Scientific Research Publishing

Journal of Information Security, 2026, 17(3), 276-291
https://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

Empowering Cybersecurity: Evaluating Agentic AI Frameworks for Threats Detection and Adaptive Incident Response

Huda Zafir Alshehri, Eiman Salamah Aljohani, Maha Badr Alyami, Abdelrahman Elsharif Karrar

College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia
Email: TU4725448@taibahu.edu.sa, ejohani@taibahu.edu.sa, TU4725291@taibahu.edu.sa, akarrar@taibahu.edu.sa

How to cite this paper: Alshehri, H.Z., Aljohani, E.S., Alyami, M.B. and Karrar, A.E. (2026) Empowering Cybersecurity: Evaluating Agentic AI Frameworks for Threats Detection and Adaptive Incident Response. *Journal of Information Security*, **17**, 276-291.
https://doi.org/10.4236/jis.2026.173014

Received: May 6, 2026
Accepted: June 27, 2026
Published: June 30, 2026

Abstract

This paper presents a comprehensive review of Agentic AI frameworks in cybersecurity, with particular emphasis on autonomous threat detection and adaptive incident response. As cyber threats continue to evolve in complexity and scale, traditional rule-based security mechanisms are becoming increasingly ineffective in responding to dynamic and sophisticated attacks. Agentic AI introduces a transformative approach by integrating real-time monitoring, continuous learning, autonomous reasoning, and adaptive decision-making into cybersecurity operations. The review examines recent advances in areas such as autonomous threat detection, SOC automation, adaptive response systems, governance, explainability, and adversarial risks. The findings indicate that Agentic AI significantly improves detection accuracy, accelerates incident response, reduces false positives, and enhances overall cyber resilience through technologies including machine learning, deep learning, reinforcement learning, and multi-agent systems. However, despite these promising developments, current research remains fragmented and faces several limitations, including limited real-world validation, insufficient explainability, governance challenges, and vulnerabilities such as adversarial attacks, data poisoning, and model manipulation. The study highlights the need for secure, transparent, and human-centered deployment strategies, as well as standardized evaluation frameworks and stronger human-AI collaboration models. Overall, Agentic AI represents a promising paradigm for building more intelligent, adaptive, and resilient cybersecurity systems capable of addressing the challenges of modern digital environments.

Keywords

Agentic AI, Cybersecurity, Threat Detection, Adaptive Incident Response,

Autonomous Cyber Defense, Multi-Agent Systems, Explainable AI, Cyber Resilience

1. Introduction

Digital technologies are evolving rapidly, making modern systems more interconnected and significantly reshaping cybersecurity [1]. While this transformation has improved efficiency and innovation, it has also intensified cyber threats such as ransomware, phishing, malware, and advanced persistent threats (APTs) [2]. Many of these attacks rely on automation and evasion techniques that traditional security tools—including signature-based detection systems, firewalls, and rule-based intrusion detection mechanisms—struggle to address effectively [3]. These conventional approaches are largely reactive, depend on known attack patterns, and often fail to detect zero-day attacks or adapt to evolving threats. In addition, Security Operations Centers (SOCs) face increasing challenges due to massive data volumes, which lead to slower response times, higher false-positive rates, and operational inefficiencies [4].

To overcome these limitations, researchers have increasingly integrated artificial intelligence into cybersecurity systems [5]. Early AI-based approaches, particularly machine learning and deep learning techniques, improved anomaly detection capabilities [6]. However, these systems still function mainly as assistive tools that require human intervention for final decision-making. As cyber threats continue to evolve, there is a growing demand for intelligent systems capable of autonomous and real-time security operations [7]. This need has driven the emergence of Agentic AI, which represents a transition from reactive support tools toward autonomous systems capable of reasoning, adapting, and acting independently [8]. Agentic AI introduces capabilities such as autonomous decision-making, continuous learning, context awareness, and adaptive response mechanisms [9]. Recent studies demonstrate that these systems can improve detection accuracy, reduce response times, lower false positives, and automate incident response processes [10]. Examples include self-healing systems, autonomous threat-hunting frameworks, and cognitive SOCs that have shown substantial improvements in cybersecurity efficiency and resilience [11].

The literature also reflects a broader shift from reactive defense models toward proactive and predictive cybersecurity strategies [12]. Agentic AI systems are increasingly capable of anticipating threats, dynamically adapting defenses, and responding autonomously in real time [13]. These advancements are supported by technologies such as reinforcement learning, deep neural networks, hybrid AI models, and multi-agent systems, which collectively improve the adaptability and robustness of cybersecurity frameworks [14]. Nevertheless, current research on Agentic AI in cybersecurity remains fragmented [15]. Many studies focus on isolated aspects such as threat detection, decision-making, or system architecture

without presenting integrated frameworks that connect these components effectively [16]. Critical areas including recovery mechanisms, explainability, governance, and real-world deployment are often insufficiently addressed [17]. Furthermore, Agentic AI introduces additional security concerns, including adversarial attacks, data poisoning, and vulnerabilities in autonomous decision-making systems [18]. These challenges highlight the need for a comprehensive and structured review of Agentic AI frameworks in cybersecurity [19]. Accordingly, this paper examines recent developments in advanced threat detection and adaptive incident response systems [20], identifies key research trends and limitations, and highlights future directions for building more resilient, autonomous, and trustworthy cybersecurity frameworks.

1.1. Research Gap

Even though more researchers are turning their attention to agentic AI for cybersecurity, the work so far is all over the map. Most papers pick one narrow slice of the problem—say, how to detect a threat or how to launch a response—without connecting those pieces into a complete picture that covers the full arc of a security event. Recovery, for example, hardly ever gets mentioned in the same breath as detection and response. Another issue is that very few studies bother to test their systems in real, live environments. Simulations and theoretical models are fine up to a point, but they do not tell you much about how something will hold up when actual attackers are knocking on the door. On top of that, nobody has agreed on a common set of benchmarks, so comparing one agentic approach against another is next to impossible. Explainability tends to get glossed over, too, even though people need to understand why an autonomous system made a particular call, especially when the stakes are high. Governance and ethics are either missing entirely or treated as an afterthought. And perhaps most worrying, hardly anyone is looking at how to protect the agentic AI systems themselves from things like adversarial attacks, poisoned data, or someone messing with the model. Taken together, these gaps mean the field is not yet ready to build the kind of integrated, dependable, and practical systems that real security operations demand.

1.2. Contributions

This paper brings several new things to the table. First, it pulls together recent work on agentic AI in cybersecurity and organizes it around six main themes: automated threat detection, agent architectures, SOC automation, governance, adversarial risks, and human-focused threats like phishing. Second, it sorts through the evidence to show where agentic AI actually delivers—faster responses, better detection rates, fewer false alarms—and where the proof is still thin or mixed. Third, the paper offers a new conceptual diagram (**Figure 1**) that ties together agentic features, AI techniques, real-world applications, outcomes, and remaining research gaps, giving readers a single place to see the whole landscape. Fourth, it includes a side-by-side comparison table (**Table 1**) that makes it easy to spot

which studies use AI techniques, which include decision mechanisms, and which have true agentic features, along with a clear note on each one's limitations. Finally, the paper lays out a targeted list of what to tackle next: building systems that handle detection, response, and recovery together; testing at scale in live environments; creating standard benchmarks; making decisions easier to understand; setting up governance and ethics guardrails; defending the AI systems themselves from attack; and adapting agentic approaches for new, spread-out environments like IoT, cloud, and edge computing. Agentic AI represents a shift from reactive tools to systems that act autonomously, set goals, and make decisions with little human input. Once a theoretical concept, it is now being deployed across healthcare, cybersecurity, software engineering, and business. Researchers are examining both its potential benefits and risks. This paper surveys recent work on agentic AI, covering technical capabilities, security concerns, real-world applications, and evaluation methods.

1.3. Scope and Operational Definitions

In this review, agentic AI refers to AI systems capable of autonomous, goal-directed, and context-aware decision-making with limited human intervention. Adaptive incident response refers to security actions that are dynamically selected and adjusted according to the evolving threat context rather than following a fixed rule set. Decision mechanisms refer to the rules, models, or learning-based processes used by a system to select a security action. Agentic features refer to system characteristics such as autonomy, planning, memory, adaptability, real-time processing, and multi-agent coordination. Studies were classified as agentic if they explicitly described autonomous or semi-autonomous threat detection, adaptive response, reasoning, learning, or coordination in cybersecurity tasks; studies that used AI only as a support tool without autonomous action were not classified as agentic AI.

1.4. Review Methods

This paper used a structured review approach to explore recent research on Agentic AI in cybersecurity. Relevant studies were gathered from well-known academic databases such as IEEE Xplore, SpringerLink, ScienceDirect, Google Scholar, arXiv, SSRN, and MDPI. The search mainly focused on publications released between 2023 and 2026. To locate relevant work, several keywords and combinations were used, including "Agentic AI," "autonomous cybersecurity," "adaptive incident response," "AI-driven threat detection," "cognitive security agents," "SOC automation," and "autonomous cyber defense". The review process was carried out in several steps. First, duplicate papers and studies unrelated to cybersecurity were removed after reviewing titles and abstracts. Then, the remaining papers were examined in full to determine whether they addressed autonomous or agent-based AI applications in cybersecurity. Finally, the selected studies were grouped and analyzed according to their contributions in areas such as threat detection,

adaptive response, governance, adversarial risks, SOC automation, and agent architectures. Studies were included when they discussed autonomous AI techniques, adaptive defense systems, or agent-based cybersecurity frameworks. On the other hand, papers were excluded if they focused only on traditional AI methods without autonomous capabilities, lacked direct relevance to cybersecurity, or did not provide enough technical or experimental detail. The final group of studies was selected based on relevance, recency, research quality, and their contribution to the main themes covered in this review

2. Background

As cyber threats grow more sophisticated by the day, traditional security tools often struggle to keep pace. That struggle has encouraged researchers to weave advanced artificial intelligence more deeply into cybersecurity work. This section walks through how AI has developed in this space, introduces the idea of agentic AI, and covers its key traits, detection methods, system designs, and role in handling security incidents. In the early days, security systems relied on fixed rules and known attack signatures, which meant they frequently missed brand new threats. Then machine learning came along, offering the ability to spot patterns across huge datasets, and deep learning later made it possible to handle messy, real world information like network logs. Still, most of these systems required a person to make the final judgment call. That missing piece—full autonomy—is what opened the door for agentic AI [2] [3] [5]. Agentic AI refers to systems that can act independently, weigh different options, and take steps toward a goal without needing constant human guidance. In cybersecurity, this means they can spot threats, assess the situation, and react instantly, unlike older tools that simply follow preset instructions [2] [3]. The main characteristics include operating without human input, learning from earlier events, running in real time, understanding context, letting multiple agents work together, and pursuing well defined objectives such as minimizing damage from an attack [2] [6] [8]. When it comes to spotting threats, these systems rely on a mix of approaches. They use machine learning to detect unusual behavior, deep learning to sift through enormous volumes of data, reinforcement learning to improve based on past outcomes, and GANs to run attack simulations. They also combine different models or set up multi agent teams to catch what a single method might miss [1] [4] [16]. Rather than sticking to rigid, prewritten response plans, Agentic AI crafts its actions based on the specific attack it sees. These actions might involve isolating infected machines, blocking malicious traffic, adjusting firewall rules, or launching recovery procedures on their own [1] [6] [10]. There are generally three levels of autonomy. In a human in the loop setup, the AI makes suggestions, but people still make the final decisions. With humans on the loop, the AI acts independently while people monitor from a distance. And in fully autonomous mode, the AI handles everything alone, from detection to response [2] [8] [14]. Cognitive autonomy takes this a step further. Here, AI agents can examine their environment,

reason through complicated situations, and make smart decisions without outside help. They blend perception, learning, memory, and decision making to understand the bigger picture and choose effective responses [3] [7]. As for system designs, common architectures include simple rule based systems which are easy to build but inflexible alongside machine learning based designs, deep learning based designs, hybrid models that combine multiple techniques, and multi agent setups where several agents coordinate their efforts [5] [14]. Finally, agentic AI speeds up incident management by automating detection, response, and recovery. It catches threats as they unfold, studies attack patterns on the fly, acts without delay, and helps bring systems back online much faster than traditional methods [6] [9] [19]. With agentic AI, systems can run on their own, think through difficult objectives, and act without much human oversight [21]. Unlike older AI that simply reacts, Agentic AI plans and adjusts as needed, working like a teammate or an independent operator [22]. Lately, studies have looked at what this technology can do and what might go wrong in medicine [23] cybersecurity [24] software development [25] and business systems [26]. This paper brings together current findings on agentic AI, covering what it promises technically [27], where security issues arise [28], how it is used across fields [29], and new ways to assess it [30]. Our goal is to give a broad picture of the remaining hurdles and how to build agentic AI that people can trust.

3. Related Works

More researchers are looking into how agentic AI can help with cybersecurity, especially for spotting threats and handling incidents. This review groups recent studies into six themes. Sheth and his team [1] built a self-healing security framework where agentic AI handles detection and response on its own. It caught 96.8% of threats, recovered 75% faster, and cut downtime by 60%. In another paper [4], they designed a threat hunting system using neural networks that hit 98.2% accuracy and a 97.1% F1-score on standard datasets. Hattali [15] showed how adaptive AI stops ransomware, phishing, and zero-day attacks in real time across finance, healthcare, and defense. Lazer *et al.* [2] said agentic AI goes beyond basic generative models by reasoning, planning, and adapting over time. They covered both defense uses and offensive risks. Adabara *et al.* [3] reviewed goal-driven agentic AI with cognitive architectures and reinforcement learning, looking at neuromorphic hardware, hybrid defense, and cross-border governance. Another survey by the same group [5] focused on reasoning, memory, and planning, noting that fixed rules do not work well against changing threats. Sugumar [5] proposed a next-generation SOC built on multiple cognitive agents that use deep learning, reinforcement learning, and knowledge graphs. It did better than standard SIEM tools. Kshetri [8] found that agentic AI can automate SOC tasks and boost efficiency, but also creates new weak spots. Adeyemi [9] systematically reviewed AI-led response systems, highlighting machine learning and NLP for finding anomalies while raising questions about openness and accountability. Indranil [6] created a Cognitive Trust Architecture that updates trust levels on the fly based on behavior,

intent, and context, and also models adversarial moves. Evani [7] put together a security framework for agentic AI with layered risk controls, dynamic trust scores, and explainable decisions. Malatji [13] matched different AI agent types to the NIST Cybersecurity Framework 2.0 to help organizations pick the right one. Chakrabarty [10] looked at how attackers target agentic AI systems, pointing to risks from system complexity and broad access rights, and proposed multi-layered defense. Balassone *et al.* [17] ran capture-the-flag tests with autonomous agents and found that defenders lost their advantage when real-world limits were added. Datta *et al.* [19] examined security risks in LLM-powered agentic AI, offering a threat classification and reviewing defense methods. Mustafa [12] used agentic AI with NLP, behavior analysis, and reinforcement learning to spot and stop phishing and vishing attacks. It outperformed traditional methods in accuracy and false alarms. Tallam [11] discussed how agentic and frontier AI are reshaping defense against advanced persistent threats, pushing for real-time monitoring and ethical oversight. Hernández-Rivas [14] looked at machine learning for catching advanced threats early, pointing to problems like models growing stale and attackers adapting. Molina *et al.* [16] reviewed 22 studies on AI-powered threat response and listed seven major challenges. Kotte [18] compared autonomous AI with traditional security, noting better speed and accuracy but also transparency headaches. Research on agentic AI shows strong potential but spotty real-world proof. Salehi *et al.* [21] found that while agentic AI works well for tasks like image segmentation in neuroradiology, clinical evidence still lags behind lab results. Leo *et al.* [22] proposed a trust-based risk framework for security, shifting focus from threats to verifiable agent behavior. Li *et al.* [23] introduced "in silico team science," where multi-agent systems run experiments and crunch omics data on their own. On the theory side, Floridi *et al.* [25] rolled out Agentic AI Optimization (AAIO), separating optimization done by agents from optimization done to them. Olujimi *et al.* [26] mapped research trends in small businesses, spotting gaps tied to limited resources. Hasan *et al.* [27] showed that current testing methods miss many emergent agent behaviors. Dholakia *et al.* [28] argued that old-school metrics don't work for goal-driven systems.

Cross-Domain Applications

Drawing on studies from neuroradiology, biomedical research, psychiatry, cybersecurity, software testing, entrepreneurship, and religious tourism, research offer a cross-sector look at current challenges and pathways toward trustworthy deployment. Sharma *et al.* [24] laid out a roadmap for psychiatric care, weighing round-the-clock monitoring against safety and bias concerns. Al-Bashrawi *et al.* [29] explored how founders and AI agents can work together to spark innovation. Lastly, Khamis [30] built and tested a working agentic system for Umrah travel planning, proving that multiple sub-agents can handle logistics, user preferences, and real-time changes in a service setting. **Table 1** summarizes recent literature on agentic AI in cybersecurity. Most studies use AI techniques and agentic fea-

tures, but only about half include decision-making mechanisms. Common gaps include poor transparency, little focus on recovery processes, and too much theory over practice. While a few works stand out, the field still needs better explainability, recovery methods, and real-world testing.

Table 1. Summary of related studies: limitations, features, and contributions.

Ref	Year	Key Contribution	Agentic-AI Techniques	Decision Mechanisms	Agentic Features	Limitation/Gaps
[1]	2025	Self-healing security framework with autonomous detection and adaptation	✓	✗	✓	Lacks transparency in decision-making
[2]	2026	Highlights key agentic AI principles for cybersecurity	✓	✓	✓	No specific cybersecurity techniques discussed
[3]	2025	Reviews ML techniques for attack prediction and identification	✓	✓	✓	Needs to address how agentic AI empowers cybersecurity
[4]	2025	AI-driven autonomous threat hunting using ML and deep RL	✓	✗	✓	Lacks decision-making transparency
[5]	2025	Reviews defense architectures and governance innovations	✓	✓	✗	Needs simpler, more accessible introduction
[6]	2024	Cognitive AI agents for faster threat identification	✓	✓	✓	Needs extension into explainable agentic AI
[7]	2025	Stakeholder perspectives on AI dependability in QA processes	✗	✗	✗	More managerial than technical
[8]	Preprint	Security framework with risk governance, trust scoring, and explainable pipelines	✓	✓	✓	Fig. 3 needs layer numbering
[9]	2025	Explores agentic AI for emerging cyber threats	✓	✓	✓	Recovery mechanisms not mentioned
[10]	2023	PRISMA-based review of AI-driven autonomous response systems	✓	✓	✓	None
[11]	2025	Analyzes adversarial attacks on agentic AI systems	✓	✓	✓	Proposed model needs more detail
[12]	2025	Agentic and frontier AI for APT defense with ethical governance	✓	✓	✓	Recovery mechanisms not mentioned
[13]	2025	Agentic AI as proactive sentinel against social engineering	✓	✓	✓	Experimental sample needs expansion
[14]	2025	Aligns AI agent architectures with NIST CSF 2.0	✗	✓	✓	Focuses on “agent” not “agentic”; agentic only theoretical
[15]	2024	ML techniques for APT detection and mitigation	✓	✓	✓	None
[16]	2024	Adaptive AI for autonomous decision-making and self-learning	✓	✓	✓	Implicitly addresses agentic AI, not explicitly
[17]	2024	Broad survey of AI-driven threat response systems (22 studies)	✗	✗	✗	Not focused on agentic AI specifically
[18]	2025	Evaluates AI agents in dynamic cybersecurity environments	✗	✓	✓	Recovery addressed via cybersecurity AI, not agentic AI
[19]	2025	Autonomous AI agents for proactive threat detection and response	✓	✓	✓	Recovery using autonomous AI agents included ✓
[20]	2025	Threat taxonomy, benchmarks, and defenses for agentic AI	✓	✓	✗	Lacks comprehensive architectural perspective

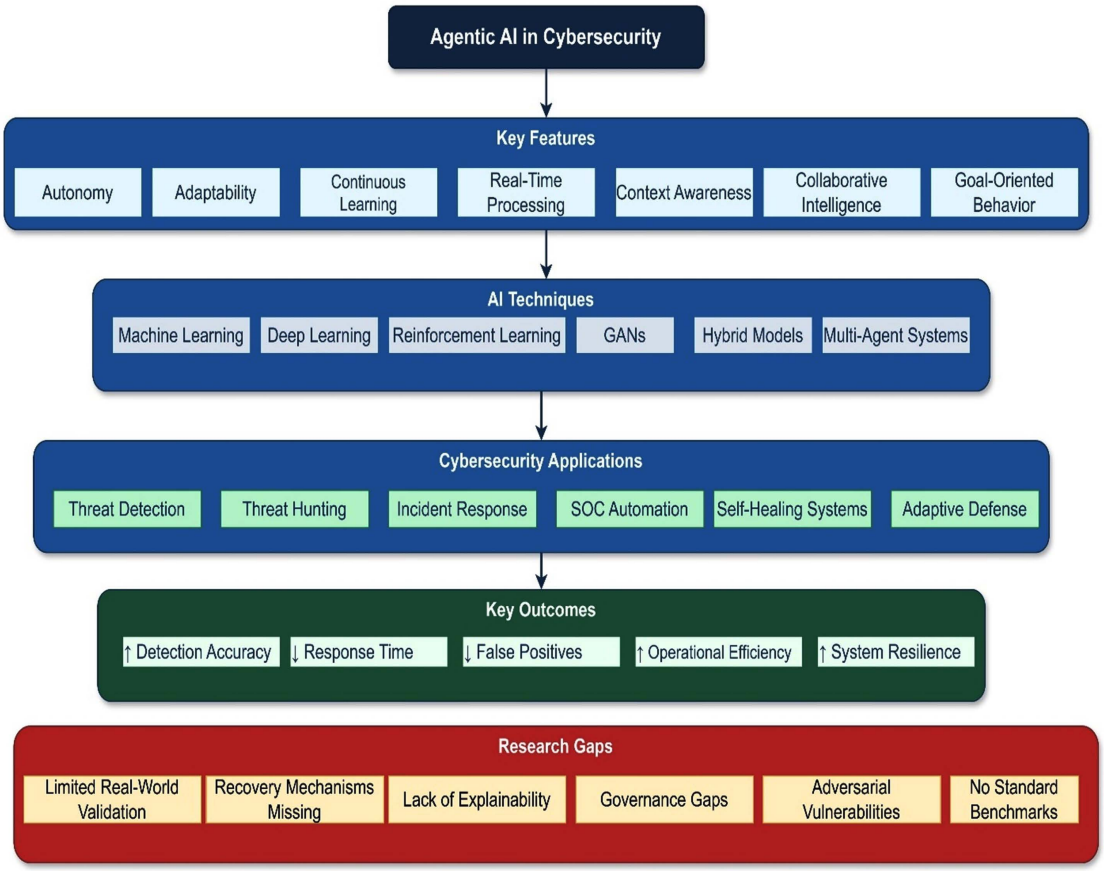

Figure 1. Agentic AI in Cybersecurity: Framework, Outcomes, and Research Gaps.

Figure 1 outlines how Agentic AI works in cybersecurity. It starts with key features like autonomy and continuous learning, which feed into AI techniques such as machine and reinforcement learning. These techniques enable applications including threat detection, incident response, and SOC automation. The outcomes are better accuracy, faster responses, fewer false alarms, and stronger resilience. The bottom section lists six research gaps, such as limited real-world testing, missing recovery mechanisms, and no standard benchmarks. The figure offers a clear snapshot of where Agentic AI stands today and what still needs work.

4. Findings

When you line up the selected studies and compare them, one thing becomes obvious: Agentic AI is shaping up to be the next big step forward in cybersecurity. Across all the research, these systems aren't just designed to detect trouble. They watch, reason, make choices, take action, and adjust as things change, all with very little need for a human to step in. You see this playing out in self-healing networks, autonomous threat hunting, automation inside security operations centers, and response frameworks that adapt on the fly [1] [4] [6] [16] [19]. A lot of the studies highlight serious improvements when it comes to autonomous detection and response. Agentic AI systems can keep an eye on things in real time, identify threats

entirely on their own, and kick off responses without waiting for a person to give the go-ahead. That slashes the workload for human teams and makes incident handling much faster, especially in high-pressure situations where every second counts [1] [6] [19]. One trait that shows up repeatedly is continuous improvement. These systems lean on machine learning, deep learning, reinforcement learning, and feedback loops to get better as time goes on. That ability lets them adapt to fresh attack techniques and catch threats no one has seen before, something static, rule-based tools just cannot do [4] [17]. Several studies also point to a clear move away from being reactive and toward being proactive. Older methods basically sit around waiting for an attack to land. Agentic AI flips that script. It helps forecast threats, go out and hunt them down ahead of time, and tweak defenses on the fly. The result is stronger overall resilience and less damage when something does slip through [9] [17]. Beyond just how well these systems perform, quite a few papers dig into ethics and control. People keep coming back to explainability, accountability, and transparency, especially around how decisions get made and how to roll these systems out in a responsible way [2] [3] [8]. Of course, Agentic AI also brings brand new problems. Researchers point to vulnerabilities like adversarial attacks, data poisoning, and outright misuse. Trust failures, misalignment between goals, and weak control mechanisms also show up as serious concerns that still need fixing [2] [11]. Looking at what gets the most attention, most studies zero in on detection techniques, adaptive learning, autonomous response, and building conceptual or architectural frameworks. Those are the parts of the field that are most mature right now. But there are also some notable gaps. Real-world testing is pretty scarce. A lot of papers rely on simulations or theoretical ideas instead of live deployments. Governance tends to get tacked on at the end rather than built into the design from the start. Human-centered security, recovery mechanisms, and system resilience get spotty coverage at best. And without standard ways to measure performance, comparing one approach to another remains really difficult. So the bottom line is this: Agentic AI is changing cybersecurity by bringing more independence, adaptability, and smarter responses. But to make it work in actual, real-world environments, the field needs implementations that are more complete, better tested, and more secure.

Evidence Strength of Reported Performance Benefits

The literature frequently reports benefits of Agentic AI in cybersecurity, particularly in terms of detection accuracy, response speed, and false-alarm reduction. However, the level of evidence supporting these outcomes varies across studies.

1) Detection Accuracy:

Claims regarding improved detection accuracy are supported primarily by a limited number of empirical case studies rather than by broad validation across the literature. For example, Sheth *et al.* [1] reported a threat detection rate of 96.8%, while autonomous threat-hunting work achieved approximately 98.2% accuracy and a 97.1% F1-score [4]. Similarly, Mustafa [12] observed improved per-

formance in phishing detection compared with conventional approaches. Although these findings suggest strong potential, evidence currently comes from a relatively small number of isolated implementations and simulation-based evaluations rather than consistent large-scale validation across multiple independent studies.

2) Response Time:

Faster incident response appears to be one of the most consistently discussed benefits in literature. Multiple studies [1] [6] [8] [18] [19] argue that autonomous decision-making and SOC automation reduce delays in incident handling and improve operational efficiency. However, only a subset of these studies provides quantitative performance measurements, while many rely on conceptual frameworks or theoretical discussions. Therefore, evidence for response-time improvement can be viewed as a repeated finding across several papers, although quantitative validation remains limited.

3) False Alarm Reduction:

The literature provides weaker evidence regarding reductions in false alarms or false positives. Only a few studies, such as autonomous threat-hunting frameworks and social engineering detection systems [4] [13], explicitly report improvements in false-positive behavior. Most studies mention this benefit conceptually without presenting measurable results. Consequently, claims regarding false-alarm reduction remain largely conceptual arguments supported by limited empirical evidence.

Overall, while Agentic AI demonstrates promising performance outcomes, current evidence remains uneven. Improvements in response speed are supported across multiple studies, whereas gains in accuracy and false-alarm reduction rely more heavily on isolated case studies and simulation-based results. More standardized benchmarks and large-scale real-world validation are required before general conclusions can be established.

5. Discussion

The findings of this review indicate that Agentic AI is fundamentally transforming cybersecurity by shifting it from traditional reactive, rule-based mechanisms toward more autonomous, adaptive, and intelligent defense systems. This transition reflects the increasing complexity, speed, and sophistication of modern cyber threats, which require security solutions capable of operating beyond static predefined rules. One of the most significant observations concerns the growing emphasis on autonomous detection and response capabilities. Agentic AI systems are increasingly designed to monitor environments, identify threats, and initiate responses with minimal human intervention, thereby reducing response delays and alleviating the operational burden on Security Operations Center (SOC) teams. However, despite these advancements, fully autonomous cybersecurity remains an evolving objective rather than a fully realized reality, as many current systems still rely on varying degrees of human supervision and validation.

Another major finding is the critical role of adaptive learning technologies, including machine learning, deep learning, and reinforcement learning, in enabling continuous system improvement over time. These approaches address one of the primary limitations of conventional rule-based security tools by allowing systems to adapt dynamically to emerging threats and changing attack patterns. Nevertheless, the effectiveness of such adaptive mechanisms heavily depends on the availability of high-quality data, reliable model training, and bias mitigation strategies. Weaknesses in data quality, model robustness, or fairness can undermine trust in AI-driven decisions and negatively affect operational reliability.

The review also highlights the important shift from reactive cybersecurity toward proactive defense strategies. Agentic AI systems have demonstrated strong potential in predicting threats, identifying vulnerabilities, and mitigating attacks before they fully materialize. While this capability represents a major advancement in cyber defense, accurate predictive performance requires robust datasets, advanced modeling techniques, and continuous validation processes—requirements that are not comprehensively addressed in many existing studies. Consequently, the gap between theoretical potential and operational deployment remains significant.

Governance, transparency, and trust emerged as equally critical considerations. As cybersecurity systems gain greater autonomy, explainability and accountability become essential requirements rather than optional enhancements. Organizations operating in high-risk environments may hesitate to adopt fully autonomous systems without clear ethical guidelines, governance frameworks, and mechanisms for transparent decision-making. Technical performance alone is insufficient to ensure acceptance; responsible deployment also requires balancing automation with ethical oversight and human accountability.

Furthermore, the review emphasizes that the same characteristics that make Agentic AI highly effective also introduce new vulnerabilities. Autonomous systems themselves may become targets of adversarial attacks, data poisoning, or model manipulation, making AI security an essential component of cybersecurity research. Protecting the AI infrastructure is therefore just as important as protecting the systems and networks it is designed to defend.

Another important observation is the persistent gap between conceptual research and real-world implementation. Although many studies propose sophisticated architectures and theoretical frameworks, relatively few provide evidence from large-scale operational deployments or real-world testing environments. This limitation highlights the need for standardized evaluation metrics, benchmarking frameworks, and practical validation studies capable of assessing the reliability, scalability, and effectiveness of Agentic AI systems under realistic attack conditions.

Overall, Agentic AI should not be viewed merely as an incremental improvement over traditional artificial intelligence approach. Rather, it represents a fundamental paradigm shift that combines autonomy, adaptability, and intelligent

decision-making within cybersecurity operations. A hybrid human-AI collaboration model appears to offer the most practical and balanced approach for future deployment. In such environments, AI systems can efficiently manage routine monitoring, detection, and response activities, while human experts maintain strategic oversight, ethical judgment, and crisis management responsibilities. Although Agentic AI can rapidly process large volumes of data and recommend appropriate responses, exclusive reliance on autonomous decision-making may introduce ethical concerns and operational risks. Human cybersecurity professionals remain essential for handling complex incidents that require contextual understanding, strategic reasoning, and ethical evaluation. Therefore, the future of cybersecurity will likely depend on achieving an effective balance between the speed and scalability of AI-driven automation and the critical oversight provided by human expertise.

Limitations

Vulnerable to adversarial attacks and data poisoning, Risks of model manipulation and system misuse, Lack of explainability and transparency, Governance, trust, and accountability issues, Limited real-world validation.

6. Conclusions

This paper examined the evolution of Agentic AI frameworks in cybersecurity, with particular emphasis on advanced threat detection and adaptive response mechanisms. The review demonstrates that Agentic AI represents a significant transition from traditional reactive security models toward more autonomous, adaptive, and intelligent cybersecurity systems. By integrating real-time monitoring, continuous learning, and autonomous decision-making, these frameworks enhance both the speed and effectiveness of cybersecurity operations. The findings indicate that Agentic AI improves detection accuracy, reduces response time, and supports proactive defense strategies against evolving cyber threats.

Despite these advantages, the current body of research remains fragmented, with limited real-world validation and considerable variation in system autonomy levels. Furthermore, important aspects such as recovery mechanisms, governance, transparency, and explainability are often insufficiently addressed. Therefore, although Agentic AI offers substantial potential, its practical implementation is still in a developmental stage. More importantly, Agentic AI should not be viewed merely as an extension of conventional AI techniques, but rather as a transformative paradigm that combines autonomy, adaptability, and intelligent response within cybersecurity environments.

Consequently, Agentic AI is expected to play a central role in shaping future cyber resilience. However, realizing this potential requires further research toward developing integrated, secure, transparent, and trustworthy systems that balance technological advancement with ethical considerations and human oversight. In conclusion, Agentic AI provides a promising pathway toward more resilient and

proactive cybersecurity frameworks, but its long-term success depends on addressing current technical, operational, and governance challenges.

7. Future Work

The limitations identified in this review highlight several important directions for future research. Current studies often focus on isolated tasks such as threat detection or response management, with limited integration of recovery mechanisms into unified security frameworks. Future work should therefore emphasize end-to-end architectures capable of managing the entire cybersecurity lifecycle. In addition, more large-scale real-world evaluations are needed to validate the reliability and adaptability of Agentic AI systems under diverse attack scenarios. Establishing standardized benchmarks for measuring detection accuracy, response speed, resilience, and autonomy is also essential for fair comparison across approaches.

As Agentic AI systems become increasingly autonomous, enhancing explainability and transparency will be critical to improving user trust and understanding of system decisions. Future research should also address ethical governance, human oversight, and accountability to ensure responsible deployment in high-risk environments. Moreover, stronger protection mechanisms are required to defend Agentic AI systems against adversarial attacks, poisoned data, and model manipulation. Hybrid human-AI collaboration models, where AI handles routine operations while humans supervise strategic decisions, represent another promising direction. Finally, emerging environments such as IoT, cloud computing, and edge infrastructures provide valuable opportunities for developing adaptive and decentralized AI-driven cybersecurity solutions.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sheth, A., Achanta, A., Matam, P., Patel, A., Sharma, P., Janapareddy, N.V.P., *et al.* (2025) AI Driven Self-Healing Cybersecurity Systems with Agentic AI for Adaptive Threat Response and Resilience. 2025 *IEEE Cloud Summit*, Washington, 26-27 June 2025, 147-153. https://doi.org/10.1109/cloud-summit64795.2025.00030

[2] Lazer, S.J., Aryal, K., Gupta, M. and Bertino, E. (2026) A Survey of Agentic AI and Cyber-Security: Challenges, Opportunities and Use-Case Prototypes. arXiv: 2601.05293.

[3] Adabara, I., Olaniyi Sadiq, B., Nuhu Shuaibu, A., Ibarahim Danjuma, Y. and Venkateswarlu, M. (2025) A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense. *F*1000*Research*, **14**, Article No. 843. https://doi.org/10.12688/f1000research.169337.1

[4] Sheth, A., Patel, A., Upadhyay, C., Ragothaman, H., Patil, B. and Udayakumar, S.K. (2025) Agentic AI for Autonomous Cyber Threat Hunting and Adaptive Defense in Dynamic Security Environments. 2025 *IEEE International Conference on Electro Information Technology* (*eIT*), Valparaiso, 29-31 May 2025, 316-321. https://doi.org/10.1109/eit64391.2025.11103697

[5] Sugumar, R. (2024) Next-Generation Security Operations Center (SOC) Resilience: Autonomous Detection and Adaptive Incident Response Using Cognitive AI Agents. *International Journal of Technology, Management and Humanities*, **10**, 62-76.

[6] Indranil, K. (2025) Cognitive Trust Architecture for Mitigating Agentic AI Threats: Adaptive Reasoning and Resilient Cyber Defense. https://philpapers.org/rec/KUMCTA

[7] Evani, P.K. (2025) Agentic AI Security: A Control Framework for Autonomous Decision-Making Systems. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5332681

[8] Kshetri, N. (2025) Transforming Cybersecurity with Agentic AI to Combat Emerging Cyber Threats. *Telecommunications Policy*, **49**, Article ID: 102976.

[9] Adeyemi, D.S. (2023) Autonomous Response Systems in Cybersecurity: A Systematic Review of AI-Driven Automation Tools. *Communication in Physical Sciences*, **9**, 878-898. https://journalcps.com/index.php/volumes/article/view/696/709

[10] Kishore Chakrabarty, P. (2025) Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies. *International Journal of Science and Research* (*IJSR*), **14**, 1367-1369. https://doi.org/10.21275/sr25417074844

[11] Tallam, K. (2025) Transforming Cyber Defense: Harnessing Agentic and Frontier AI for Proactive, Ethical Threat Intelligence. arXiv: 2503.00164.

[12] Mustafa, A. (2025) Agentic Artificial Intelligence as a Proactive Cybercrime Sentinel for Predictive Detection and Strategic Deterrence of Social Engineering Attacks. https://www.researchgate.net/publication/397997466_Agentic_Artificial_Intelligence_as_a_Proactive_Cybercrime_Sentinel_for_Predictive_Detection_and_Strategic_Deterrence_of_Social_Engineering_Attacks

[13] Malatji, M. (2025) A Cybersecurity AI Agent Selection and Decision Support Framework. arXiv: 2510.01751.

[14] Hernández-Rivas, A., Morales-Rocha, V. and Sánchez-Solís, J.P. (2024) Towards Autonomous Cybersecurity: A Comparative Analysis of Agnostic and Hybrid AI Approaches for Advanced Persistent Threat Detection. In: Rivera, G., Pedrycz, W., Moreno-Garcia, J. and Sánchez-Solís, J.P., Eds., *Innovative Applications of Artificial Neural Networks to Data Analytics and Signal Processing*, Springer, 181-219. https://doi.org/10.1007/978-3-031-69769-2_8

[15] Hattali, A. (2024) Adaptive AI for Cybersecurity: Revolutionizing Threat Detection and Incident Response through Intelligent Algorithms. https://www.researchgate.net/profile/Albert-Hattali/publication/386525333

[16] Molina, S.B., Nespoli, P. and Mármol, F.G. (2023) Tackling Cyberattacks through AI-Based Reactive Systems: A Holistic Review and Future Vision. arXiv: 2312.06229.

[17] Balassone, F., Mayoral-Vilches, V., Rass, S., Pinzger, M., Perrone, G., Romano, S.P. and Schartner, P. (2025) Cybersecurity AI: Evaluating Agentic Cybersecurity in Attack/Defense CTFs. arXiv: 2510.17521.

[18] Kotte, G. (2025) Securing the Future with Autonomous AI Agents for Proactive Threat Detection and Response. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5283830

[19] Datta, S., Nahin, S.K., Chhabra, A. and Mohapatra, P. (2025) Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges. arXiv: 2510.23883.

[20] Bandi, A., Kongari, B., Naguru, R., Pasnoor, S. and Vilipala, S.V. (2025) The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. *Future Internet*, **17**, Article 404.

https://doi.org/10.3390/fi17090404

[21] Salehi, S., Keishing, V., Singh, Y., Wei, D., Khosravi, A., Habibi, P., *et al.* (2026) Systematic Review: Agentic AI in Neuroradiology: Technical Promise with Limited Clinical Evidence. *Journal of Imaging Informatics in Medicine*. https://doi.org/10.1007/s10278-025-01839-2

[22] Leo, M., Tan, F., Miao, T. and Anand, G. (2026) From Threat to Trust: Assessing Security Risks of Agentic AI Systems. *International Journal of Information Security*, **25**, Article No. 23. https://doi.org/10.1007/s10207-025-01185-y

[23] Li, B., Saini, A.K., Hernandez, J.G. and Moore, J.H. (2026) Agentic AI and the Rise of *in Silico* Team Science in Biomedical Research. *Nature Biotechnology*, **44**, 711-725. https://doi.org/10.1038/s41587-026-03035-1

[24] Sharma, D., Meshkat, S., Perivolaris, A., Kamaleddin, M.A., Teferra, B.G., Rueda, A., *et al.* (2026) Reimagining Psychiatric Care with Agentic AI: Promise, Challenges, and a Roadmap Forward. *npj Digital Medicine*, **9**, Article No. 252. https://doi.org/10.1038/s41746-026-02453-4

[25] Floridi, L., Buttaboni, C., Gertler, N., Hine, E., Morley, J., Novelli, C., *et al.* (2026) Agentic AI Optimisation (AAIO): What It Is, How It Works, Why It Matters, and How to Deal with It. *Minds and Machines*, **36**, Article No. 25. https://doi.org/10.1007/s11023-026-09779-8

[26] Olujimi, P.A., Owolawi, P.A., Pretorius, A. and Van Wyk, E. (2025) Mapping the Research Landscape of Agentic AI in SMMEs through a Bibliometric Analysis of Patterns and Knowledge Gaps. *Discover Artificial Intelligence*, **6**, Article No. 63. https://doi.org/10.1007/s44163-025-00764-1

[27] Hasan, M.M., Li, H., Fallahzadeh, E., Rajbahadur, G.K., Adams, B. and Hassan, A.E. (2026) An Empirical Study of Testing Practices in Open Source AI Agent Frameworks and Agentic Applications. *Empirical Software Engineering*, **31**, Article No. 124. https://doi.org/10.1007/s10664-026-10857-9

[28] Dholakia, A., Wani, S.G., Ellison, D., Hodak, M., Dutta, D., Nagaraja, S., *et al.* (2026) Benchmarking Considerations for Agentic AI Systems. In: Nambiar, R. and Poess, M., Eds., *Performance Evaluation and Benchmarking*, Springer, 89-98. https://doi.org/10.1007/978-3-032-18070-4_6

[29] Al-Bashrawi, M.A., Al-Sharafi, M.A., Elgendy, I.A., Helal, M.Y.I., Anbalagan, M.K., Chae, I., *et al.* (2026) Agentic AI Systems and the Future of Entrepreneurship: A Perspective on Co-Agency, Innovation, and Ecosystem Transformation. *International Entrepreneurship and Management Journal*, **22**, Article No. 27. https://doi.org/10.1007/s11365-026-01164-2

[30] Khamis, A. (2026) Design and Evaluation of an Agentic AI Framework for Personalized Umrah Trip Planning. *Arabian Journal for Science and Engineering*, **51**, 12299-12319. https://doi.org/10.1007/s13369-025-11021-z