Agentic AI for Autonomous Cyber Defense: Architecture, Evaluation, and Risk Management

Pritika Mehra
Assistant Professor, Khalsa College for Women, Amritsar
doi.org/10.64643/IJIRTV13I1-205726-459

***Abstract*—Agentic AI is shifting artificial intelligence from passive prediction to autonomous systems that can perceive, reason, plan, act, and use tools across digital environments. In cybersecurity, this shift is timely because defenders need faster triage, adaptive response, and safer automation against increasingly complex attacks, while recent reports note that cybersecurity leaders are investing more in AI and agentic defenses. This paper proposes a research framework for agentic AI in cyber defense, with emphasis on architecture, evaluation, and governance. It also addresses agent-specific risks such as tool misuse, cascading failures, prompt injection, and goal drift.**

***Index Terms*—agentic AI, cybersecurity, intrusion detection, autonomous systems, AI governance**

I. INTRODUCTION

Artificial intelligence has evolved rapidly from narrow classifiers to foundation-model systems capable of reasoning over long tasks and interacting with external tools. The latest research describes agentic AI as a unified system where perception, planning, action, collaboration, and tool use are integrated into a single operational loop. In cybersecurity, this matters because attack surfaces are expanding and defenders are under pressure to respond at machine speed. At the same time, autonomous operation introduces new risks that are not fully covered by traditional AI safety methods or conventional software security practices.

A major gap remains between promising demonstrations and secure real-world deployment. Recent work suggests that existing governance frameworks need agent-specific extensions because autonomy changes how identity, authorization, monitoring, and accountability work in practice. This paper argues that agentic AI can improve cyber defense, but only if autonomy is constrained by policy, auditability, and human oversight. The central research question is whether an agentic framework can outperform standard ML/DL intrusion detection in responsiveness and operational value without reducing safety.

II. RELATED WORK

Recent literature frames agentic AI as a distinct research direction rather than a simple extension of large language models. One 2026 arXiv paper proposes a taxonomy spanning perception, brain, planning, action, tool use, and collaboration, and it highlights open problems such as hallucination in action and prompt injection. Another 2026 framework extends NIST-style risk management to autonomous agents and emphasizes agent-specific failures such as cascading errors, accountability diffusion, and goal drift.

Cybersecurity-specific work is also moving quickly. Industry and research sources in 2026 describe rising investment in AI and agentic defenses, while threat reports describe attackers using agentic workflows to automate reconnaissance, phishing, credential testing, and infrastructure rotation. Academic and survey work on agentic AI security further notes that these systems create amplified risks because they can execute actions across web, software, and physical environments. For intrusion detection, the research opportunity is to combine strong detection models with agentic orchestration for triage, investigation, and response.

III. PROPOSED FRAMEWORK

The proposed framework contains five layers: perception, reasoning, planning, action, and governance. Perception ingests network flows, endpoint telemetry, alerts, logs, and threat intelligence. Reasoning interprets context, correlates events, and generates hypotheses about malicious activity.

Planning selects a sequence of actions such as enrichment, containment recommendation, or escalation. Action executes limited, policy-guarded operations through approved tools. Governance enforces identity, scope, audit logs, and human approval for high-impact steps.

A practical architecture can be implemented as a multi-agent SOC assistant. One agent can profile the environment, another can map assets, another can analyze threats, and another can recommend mitigations; a recent 2026 cybersecurity paper describes a six-agent system built along similar lines for risk assessment. In the context of intrusion detection, the agent does not replace the classifier. Instead, it sits on top of the detector and converts alerts into decisions, summaries, and response workflows. That design keeps the detection model focused on prediction while the agent handles workflow automation.

IV. METHODOLOGY

The study can be designed as a comparative experiment with two systems: a baseline intrusion detection pipeline and an agentic intrusion response pipeline. The baseline may use a standard ML or deep learning model trained on datasets such as CICIDS2017 or UNSW-NB15, while the agentic system uses the same detector but adds autonomous triage, enrichment, and response planning. Evaluation should include detection accuracy, false positive rate, response latency, analyst workload reduction, and explanation quality. Safety metrics should include unsafe tool invocation, incorrect containment recommendation, and resistance to prompt injection.

To make the paper publishable, the methodology should also include a governance design. Agent permissions should be limited by least privilege, high-risk actions should require approval, and every action should be logged for audit. This is important because current guidance on agentic AI emphasizes human oversight, containment, and continuous monitoring as essential controls. If you have access to real security logs, the paper can also include a case study; otherwise, a benchmark-based simulation is acceptable for a first submission.

V. EXPERIMENTAL DESIGN

A strong experiment would follow this structure. First, train or use a pre-trained intrusion detector on the chosen dataset. Second, generate alerts and feed them into an agentic orchestration layer that performs enrichment, summarization, and decision support. Third, compare the baseline and agentic workflows over multiple runs. Fourth, measure performance with statistical testing, ideally paired tests across folds or repeated runs.

Key metrics can include detection accuracy, precision, recall, F1-score, and ROC-AUC; mean time to detect and mean time to respond; alert triage workload, measured by number of manual steps removed; safety violations, including wrong tool use or overconfident actions; and explanation usefulness, assessed by analyst review or rubric scoring. For a cybersecurity-oriented paper, it is useful to include both offensive and defensive considerations. Recent reports warn that adversaries are already using agentic frameworks to speed up attack chains, so the defensive system should be evaluated under adversarial conditions too. This makes the contribution more realistic and more relevant to current threat conditions.

VI. RISK ANALYSIS

The largest risk in agentic AI is not raw model error; it is autonomous error propagation. Once an agent can plan and act, a bad inference can become a sequence of bad actions. This is why prompt injection, tool misuse, infinite loops, and goal drift are central concerns in the current literature. A cyber-defense agent that misreads evidence or overreacts to a benign pattern can trigger unnecessary disruption.

Governance is therefore part of the technical design. Recent 2026 frameworks argue that agent identities, scoped permissions, lifecycle governance, and auditability are required for safe deployment. In practice, this means that containment actions should be limited, reversible, and logged, while destructive operations should remain human-approved. A good paper should present these controls as necessary safeguards rather than optional best practices.

VII. DISCUSSION

The main advantage of agentic AI in cybersecurity is operational speed. A mature agent can summarize alerts, correlate logs, look up context, and draft response steps much faster than a human analyst working manually. It also helps reduce analyst fatigue by removing repetitive triage tasks. However, autonomy can create hidden complexity, especially if the model becomes overly confident or interacts with untrusted inputs.

The broader implication is that intrusion detection is moving from static classification toward autonomous security operations. That shift is why agentic AI is considered one of the most important AI research trends of 2026. For your background in machine learning and cybersecurity, this topic is especially strong because it connects novel AI with a concrete domain problem. It also supports a paper structure that is both current and academically defensible.

VIII. CONCLUSION

This paper positions agentic AI as a new and promising direction for autonomous cyber defense. The proposed framework combines intrusion detection with policy-guarded planning and response, while explicitly addressing safety, governance, and auditability. The evidence from 2026 research shows both the promise of agentic systems and the urgency of secure deployment. A future-ready contribution will not only improve detection workflows but also show how autonomy can be made trustworthy.

REFERENCES

[1] "Agentic Artificial Intelligence (AI)," *arXiv*, 2026.
[2] "Five Trends in AI and Data Science for 2026," *MIT Sloan Management Review*, 2026.
[3] Agentic AI Risk-Management Standards Profile Discussion," *UC Berkeley Center for Long-Term Cybersecurity (CLTC)-Related Coverage*, 2026.
[4] *NIST AI 100-4: Guidance on Reducing Risks of AI in Agentic Systems*. National Institute of Standards and Technology (NIST), 2026.
[5] "Cybersecurity Leaders Investing in AI and Agentic Defenses," *Ernst & Young (EY)*, 2026.
[6] "Agentic AI Security Survey," *arXiv*, 2026.
[7] "A Safety and Security Framework for Real-World Agentic Systems," *arXiv*, 2026.
[8] An Agentic Multi-Agent Architecture for Cybersecurity Risk Assessment," *arXiv*, 2026.