

AGENTIC AI FOR AUTONOMOUS CYBER-DEFENSE: A HUMAN-IN-THE-LOOP FRAMEWORK FOR PAKISTAN'S CRITICAL DIGITAL INFRASTRUCTURE

Muhammad Usama Qureshi ^{*1}, Jamshed Ali², Muhammad Mujtaba³ Muhammad Saad Ali⁴

^{*1}Teacher Computer Science Govt Technical High school Bahawalpur

²Designation: Student Department: Computer System Engineering University: Sukkur IBA University

³Designation: Student Department: Computer Science University: the University of Agriculture Peshawar

⁴MS Artificial Intelligence Student Department: Faculty of Computer Science and information technology

^{*1}usamaqureshihashmi@gmail.com, ²jamshedali.becsef24@ibasuk.edu.pk, ³mujdabaibrahimi@gmail.com,

⁴msaadali605@gmail.com

DOI: <https://doi.org/10.5281/zenodo.22640884>

Keywords

Agentic Artificial Intelligence; Autonomous Cyber Defense; Human-in-the-Loop; Critical Digital Infrastructure; Cyber Resilience; Cybersecurity Risk Management

Article History

Received: 28 June 2026

Accepted: 02 August 2026

Published: 15 August 2026

Copyright @Author

Corresponding Author: *

Muhammad Usama

Qureshi

Abstract

The increasing sophistication, speed, and adaptability of cyber threats have created a critical need for intelligent and autonomous defense mechanisms for critical digital infrastructure. This study developed a Human-in-the-Loop (HITL) framework for Agentic Artificial Intelligence (AI)-enabled autonomous cyber defense within Pakistan's critical digital infrastructure. A quantitative cross-sectional research design was employed, and data were collected from 267 cybersecurity and information-technology professionals through a structured questionnaire. The findings indicated that Agentic AI Capability significantly predicted autonomous cyber-defense effectiveness ($\beta = .32, p < .001$), while Human-in-the-Loop Oversight ($\beta = .27, p < .001$), Cybersecurity Risk Management ($\beta = .29, p < .001$), and Explainability and Accountability ($\beta = .18, p = .001$) also demonstrated significant positive effects. Autonomous cyber-defense effectiveness was further strongly associated with cyber resilience ($\beta = .52, p < .001$). The results supported a risk-adaptive autonomy approach in which low-risk and reversible cybersecurity actions were automated, whereas high-impact or uncertain decisions remained subject to human authorization. The study concluded that integrating agentic AI with structured human oversight, explainability, risk management, and accountability could enhance the security and resilience of Pakistan's critical digital infrastructure. The proposed framework provides a foundation for responsible deployment and future empirical validation of autonomous cyber-defense systems.

1. INTRODUCTION

The rapid digitalization of Pakistan's financial, telecommunications, energy, healthcare, governmental, and other critical services has increased both their strategic importance and exposure to cyber threats. Contemporary attacks

are becoming more automated, adaptive, and sophisticated, challenging conventional cybersecurity systems that depend heavily on predefined rules and manual analyst intervention (European Union Agency for Cybersecurity [ENISA], 2025). Artificial intelligence (AI)

therefore offers significant potential for improving threat detection, analysis, prediction, and response.

Recent developments in agentic AI extend conventional AI by enabling systems to perceive their environment, reason over security information, plan actions, use cybersecurity tools, and execute responses with limited human intervention. Such capabilities can enhance continuous monitoring, threat hunting, incident triage, vulnerability management, and automated response (Kshetri, 2025). However, increasing autonomy also introduces risks related to erroneous decisions, adversarial manipulation, inadequate explainability, and unintended disruption of critical services (Burbano et al., 2025; Chowdhry et al., 2025).

A Human-in-the-Loop (HITL) approach can address these challenges by combining machine speed with human judgment. Under this approach, routine and low-risk actions may be automated, while high-impact or uncertain decisions require human validation. This risk-sensitive approach is consistent with contemporary principles of trustworthy AI, emphasizing accountability, transparency, monitoring, and human oversight (Autio et al., 2024).

The issue is particularly relevant to Pakistan, where the National Cyber Security Policy 2021 identifies the protection of Critical Information Infrastructure (CII) as a national priority (Ministry of Information Technology & Telecommunication, 2021). Nevertheless, limited research has developed a context-specific architecture explaining how agentic AI can be safely integrated with human oversight within Pakistan's critical digital infrastructure. This study addresses this gap by proposing a Human-in-the-Loop Agentic AI framework incorporating threat detection, contextual reasoning, risk assessment, autonomous response, human validation, and continuous feedback. The framework seeks to achieve adaptive and rapid cyber defense while maintaining operational safety, accountability, and human control.

2. Problem Statement

Pakistan's increasing dependence on interconnected digital infrastructure has intensified the potential consequences of cyberattacks. Conventional cybersecurity approaches often face limitations involving alert overload, delayed response, evolving attack techniques, heterogeneous infrastructure, and shortages of specialized cybersecurity expertise. Although agentic AI can provide continuous monitoring, autonomous reasoning, and rapid response, unrestricted autonomy may generate significant risks when incorrect or manipulated decisions affect critical systems.

The central research gap is the lack of a context-sensitive Human-in-the-Loop framework specifically designed for agentic AI-based autonomous cyber defense within Pakistan's critical digital infrastructure. Existing research largely focuses on technical AI capabilities or general AI governance, with insufficient integration of risk-based autonomy, human intervention, explainability, accountability, and Pakistan-specific infrastructure requirements. This study addresses this gap by proposing a framework that determines when AI should act autonomously, recommend actions, or require human authorization.

3. Research Questions

1. How can agentic AI be integrated into autonomous cyber defense for Pakistan's critical digital infrastructure?
2. What levels of human intervention are appropriate for different AI-driven cybersecurity decisions?
3. How can HITL architecture balance autonomous response speed with safety, accountability, and explainability?
4. What technical, organizational, and governance factors should underpin the proposed framework?
5. How can human feedback support continuous adaptation of autonomous cyber-defense systems?

4. Research Objectives

The study aims to:

1. Develop a HITL framework for agentic AI-enabled autonomous cyber defense.
2. Identify core components for threat detection, reasoning, risk assessment, and automated response.
3. Define risk-based levels of AI autonomy and human intervention.
4. Examine technical, operational, ethical, and governance risks associated with autonomous cyber defense.
5. Integrate explainability, accountability, auditability, and human feedback into the framework.
6. Contextualize the framework within Pakistan's critical infrastructure and cybersecurity requirements.

5. Significance of the Study

Theoretical: The study contributes to emerging research on agentic AI, autonomous cyber defense, and human-AI collaboration by integrating autonomy with human oversight and trustworthy AI principles.

Practical: The framework provides cybersecurity professionals with a structured approach for automating low-risk security activities while retaining human control over high-impact decisions.

Policy: The study can support policymakers in developing requirements for accountable, explainable, auditable, and risk-sensitive AI deployment within Pakistan's Critical Information Infrastructure.

National: By promoting faster and more resilient cyber-defense capabilities while preserving human control, the framework can contribute to Pakistan's broader cybersecurity resilience and continuity of critical digital services.

Literature Review

Agentic AI and Autonomous Cyber Defense

Recent cybersecurity literature increasingly positions artificial intelligence (AI) as a means of moving from reactive security monitoring toward proactive and autonomous defense. Unlike

conventional machine-learning systems that primarily classify or predict threats, agentic AI can perceive security environments, reason over contextual information, plan actions, use external tools, and execute predefined objectives. Kshetri (2025) argues that agentic AI can transform Security Operations Center activities by automating threat detection, decision-making, investigation, and incident response. However, increased autonomy also introduces risks related to erroneous decisions, adversarial manipulation, and governance.

AI-Enabled Threat Detection and Response

The growing complexity and speed of cyberattacks strengthen the need for automated defense. ENISA (2025) highlights increasing automation, supply-chain dependencies, and AI-assisted cyber threats. Agentic systems can address these challenges by continuously correlating network telemetry, endpoint events, threat intelligence, vulnerabilities, and asset information. Adabara et al. (2025) identify real-time anomaly detection, predictive threat assessment, and adaptive response as important applications of agentic AI in cybersecurity.

Nevertheless, automation does not necessarily ensure effective defense. Autonomous agents may operate on incomplete information, generate inappropriate decisions, or execute harmful actions. Consequently, autonomous cyber defense should be evaluated through detection accuracy, decision reliability, explainability, action safety, reversibility, and operational impact.

Large Language Models and Tool-Augmented Cyber Defense

Large language models (LLMs) have expanded agentic cybersecurity capabilities by supporting contextual reasoning and interaction with security tools. Recent research indicates that LLM-based agents can assist with threat analysis, incident investigation, vulnerability assessment, and defensive decision-making. However, many existing systems remain at relatively limited autonomy levels and require stronger safety, coordination, and governance mechanisms.

Operational deployment also requires secure interaction with SIEM, EDR, firewalls, identity-management systems, vulnerability scanners, and cloud platforms. Therefore, least-privilege access, authorization controls, action validation, rollback mechanisms, and comprehensive audit trails are essential for safe agentic cybersecurity.

Risks of Autonomous Cyber-Defense Agents

Despite their potential, autonomous agents introduce a new attack surface. Burbano et al. (2025) demonstrate that autonomous cyber-defense agents can potentially be influenced or steered by adversarial actors. Such manipulation could cause defensive systems to select inappropriate responses or weaken organizational security.

Other concerns include hallucination-driven tool misuse, reasoning instability, insecure agent communication, model uncertainty, and inadequate governance. These risks are particularly important for critical infrastructure because an incorrect autonomous action could disrupt essential services. Therefore, current literature supports bounded and risk-sensitive autonomy rather than unrestricted autonomous operation.

Human-in-the-Loop Cybersecurity

Human-in-the-Loop (HITL) approaches provide an important mechanism for controlling autonomous cybersecurity systems. Rather than treating human involvement as a permanent limitation, HITL can function as a risk-adaptive governance mechanism. Low-risk and reversible activities may be automated, whereas uncertain, high-impact, or irreversible actions can require human authorization.

This approach combines the speed and scalability of AI with human contextual judgment, accountability, and ethical oversight. Recent research identifies explainability, continuous learning, human oversight, and mechanisms for limiting erroneous actions as important requirements for realistic autonomous cyber-defense deployment (Vyas et al., 2025). AI risk-management principles similarly emphasize

accountability, transparency, monitoring, and human oversight throughout the AI lifecycle (Autio et al., 2024).

Critical Infrastructure and the Pakistani Context

Cybersecurity in critical infrastructure requires additional safeguards because failures may generate cascading economic, social, and national-security consequences. Pakistan's National Cyber Security Policy 2021 emphasizes the identification, assessment, prioritization, and protection of Critical Information Infrastructure, together with cybersecurity standards, monitoring, incident response, information sharing, and risk management (Ministry of Information Technology & Telecommunication, 2021).

However, Pakistan-specific research on agentic AI for critical infrastructure protection remains limited. Existing studies predominantly examine generic network environments, simulated cyber ranges, or international organizational settings. Pakistan's critical infrastructure may involve heterogeneous technologies, legacy systems, resource constraints, varying cybersecurity maturity, and sector-specific governance requirements. These contextual characteristics require a framework specifically adapted to the national environment.

Critical Synthesis and Research Gap

The literature indicates three major developments. First, cybersecurity is moving from automated threat detection toward autonomous decision-making and response. Second, agentic and LLM-based systems increasingly provide capabilities for reasoning, tool coordination, and defensive action. Third, researchers increasingly recognize that unrestricted autonomy creates significant security, operational, and governance risks.

Nevertheless, existing literature remains fragmented. Technical studies primarily emphasize autonomous agents and LLM capabilities, while governance research focuses on general AI risks. Comparatively limited research integrates risk-based autonomy, human intervention, explainability, accountability, action reversibility, and critical-infrastructure

requirements within a unified cybersecurity architecture.

This gap is particularly evident in the Pakistani context. Accordingly, this study proposes a Human-in-the-Loop, risk-adaptive agentic AI framework for Pakistan's critical digital infrastructure, in which the degree of autonomy is determined by threat severity, AI confidence, asset criticality, potential operational impact, and action reversibility. The proposed approach seeks to combine autonomous response efficiency with human control, thereby supporting secure, accountable, and resilient cyber defense.

Methodology

Research Design

The study employed a quantitative, cross-sectional research design to examine the role of agentic artificial intelligence (AI) in autonomous cyber-defense and the importance of human-in-the-loop (HITL) mechanisms within Pakistan's critical digital infrastructure. A structured survey approach was adopted to obtain standardized responses from professionals with relevant expertise in cybersecurity, information technology, artificial intelligence, and digital infrastructure. The design enabled the study to examine relationships among agentic AI capability, autonomous cyber-defense effectiveness, human oversight, risk management, explainability, and organizational cybersecurity resilience.

Population

The study population comprised cybersecurity and information-technology professionals working in organizations associated with Pakistan's critical digital infrastructure. The target population included professionals from banking and financial services, telecommunications, energy, healthcare, transportation, and government information systems. IT managers, cybersecurity analysts, security engineers, network administrators, risk-management professionals, and information-security officers were considered eligible participants because of their direct or professional exposure to cybersecurity operations and digital infrastructure.

Sampling Technique

A purposive sampling technique was employed to ensure that respondents possessed sufficient professional knowledge of cybersecurity and AI-enabled security systems. Participants were selected according to predefined criteria, including relevant professional responsibilities and a minimum level of experience in information technology, cybersecurity, information systems, or digital risk management. This approach was considered appropriate because the study required specialized respondents rather than members of the general population.

Sample Size

A total of 300 questionnaires were distributed among eligible professionals, of which 267 usable responses were obtained, representing a response rate of 89.0%. After screening, all 267 responses were retained for statistical analysis because they contained sufficiently complete information. The sample size was considered adequate for the proposed multivariate statistical analyses and provided representation from the major critical-infrastructure sectors included in the study.

Data Collection Procedures

Data were collected through a structured questionnaire administered electronically to eligible professionals. Respondents were contacted through professional networks, organizational contacts, and relevant cybersecurity and IT communities. Before participation, respondents were informed about the purpose of the study, voluntary participation, confidentiality, and the academic use of the collected information. No personally identifying information was required. The data-collection process was conducted over a defined study period, after which responses were screened for completeness, consistency, and eligibility. Incomplete and unusable responses were excluded before statistical analysis.

Instruments and Measures

A structured questionnaire was developed from relevant literature on agentic AI, autonomous cyber-defense, human-AI collaboration,

cybersecurity risk management, and trustworthy AI. The questionnaire consisted of two sections. The first section collected demographic and professional information, including age, professional position, sector, educational qualification, and cybersecurity experience. The second section measured the principal study constructs.

The major constructs included Agentic AI Capability (AAC), Autonomous Cyber-Defense Effectiveness (ACDE), Human-in-the-Loop Oversight (HITLO), Explainability and Accountability (EA), Cybersecurity Risk Management (CRM), and Cyber Resilience (CR). Each construct was measured using multiple items. Responses were recorded on a five-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree. Higher scores indicated stronger perceptions of the respective construct.

Reliability and Validity

The reliability of the measurement instrument was evaluated using Cronbach's alpha and composite reliability. The results indicated satisfactory internal consistency, with Cronbach's alpha coefficients ranging from 0.82 to 0.91, exceeding the recommended threshold of 0.70. Composite reliability values ranged from 0.85 to 0.93, further confirming the reliability of the measurement scales.

Construct validity was assessed through convergent and discriminant validity. Standardized factor loadings exceeded the recommended level of 0.60, while the Average Variance Extracted (AVE) values ranged from 0.56 to 0.72, indicating adequate convergent validity.

Discriminant validity was established because the square root of the AVE for each construct exceeded its correlations with other constructs. In addition, expert review was conducted before the main survey to establish content validity and ensure that the questionnaire items were conceptually relevant and appropriate to the Pakistani critical-infrastructure context.

A pilot test was conducted with 30 cybersecurity and IT professionals before the main survey. Feedback from the pilot study was used to improve wording, clarity, and contextual relevance of the questionnaire. The final instrument was therefore considered sufficiently reliable and valid for subsequent statistical analysis.

Data Analysis and Interpretation

The collected data were analyzed using descriptive and inferential statistical techniques. Descriptive statistics were used to summarize respondents' characteristics and the distribution of the major study constructs. Cronbach's alpha and composite reliability were used to assess internal consistency, while Pearson correlation analysis was conducted to examine associations among the variables. Multiple regression analysis was subsequently employed to evaluate the predictive relationships between agentic AI capability, human-in-the-loop oversight, cybersecurity risk management, and autonomous cyber-defense effectiveness.

Respondents' Characteristics

Table 1 presents the demographic and professional characteristics of the 267 respondents.

Table 1 Demographic Characteristics of Respondents (N = 267)

Variable	Category	Frequency	Percentage
Gender	Male	183	68.5
	Female	84	31.5
Age	25–34 years	102	38.2
	35–44 years	91	34.1
	45–54 years	52	19.5
	55 years and above	22	8.2

Variable	Category	Frequency	Percentage
Experience	1–5 years	61	22.8
	6–10 years	89	33.3
	11–15 years	71	26.6
	More than 15 years	46	17.2
Sector	Banking/Finance	55	20.6
	Telecommunications	48	18.0
	Energy	39	14.6
	Healthcare	37	13.9
	Government	54	20.2
	Transportation	34	12.7

The demographic results indicated that the sample was predominantly composed of experienced professionals. Approximately two-thirds of respondents were male, while 31.5% were female. The largest age group was 25–34 years (38.2%), followed by 35–44 years (34.1%). Regarding professional experience, 59.9% of respondents reported more than five years of experience,

suggesting that a substantial proportion of participants possessed sufficient professional exposure to cybersecurity and digital infrastructure. Representation across banking, telecommunications, energy, healthcare, government, and transportation also provided sectoral diversity.

Descriptive Statistics and Reliability

Table 2 presents the descriptive statistics and reliability coefficients for the principal constructs.

Table 2 Descriptive Statistics and Reliability Analysis

Construct	Mean	SD	Cronbach's α	Composite Reliability
Agentic AI Capability	4.02	0.68	0.88	0.91
Autonomous Cyber-Defense Effectiveness	3.91	0.71	0.86	0.89
Human-in-the-Loop Oversight	4.18	0.61	0.91	0.93
Explainability & Accountability	3.87	0.73	0.84	0.88
Cybersecurity Risk Management	4.05	0.65	0.87	0.90
Cyber Resilience	3.96	0.69	0.89	0.92

The results showed relatively high mean scores across all constructs. Human-in-the-Loop Oversight recorded the highest mean ($M = 4.18$, $SD = 0.61$), indicating strong agreement regarding the importance of human oversight in AI-enabled cyber defense. Cybersecurity Risk Management ($M = 4.05$) and Agentic AI Capability ($M = 4.02$) also received relatively high evaluations. The reliability coefficients ranged from 0.84 to 0.91, while

composite reliability values ranged from 0.88 to 0.93. These results exceeded the commonly accepted threshold of 0.70 and demonstrated satisfactory internal consistency.

Correlation Analysis

Pearson correlation analysis was performed to examine the relationships among the study variables.

Table 3 Pearson Correlation Matrix

Variables	AAC	ACDE	HITLO	EA	CRM	CR
AAC	1					
ACDE	.64**	1				
HITLO	.58**	.61**	1			
EA	.55**	.57**	.63**	1		
CRM	.59**	.65**	.62**	.60**	1	
CR	.56**	.68**	.64**	.58**	.71**	1

Note. ** $p < .01$. AAC = Agentic AI Capability; ACDE = Autonomous Cyber-Defense Effectiveness; HITLO = Human-in-the-Loop Oversight; EA = Explainability and Accountability; CRM = Cybersecurity Risk Management; CR = Cyber Resilience.

The correlation results demonstrated positive and statistically significant relationships among all major constructs. Agentic AI Capability was positively associated with Autonomous Cyber-Defense Effectiveness ($r = .64$, $p < .01$), suggesting that greater perceived AI capability was associated with stronger cyber-defense effectiveness. Human-in-the-Loop Oversight was also significantly related to autonomous cyber-defense effectiveness ($r = .61$, $p < .01$), indicating that respondents perceived human oversight as complementary to, rather than contradictory to, autonomous defense. The strongest relationship was observed between

Cybersecurity Risk Management and Cyber Resilience ($r = .71$, $p < .01$), demonstrating the importance of systematic risk management for organizational cyber resilience.

Regression Analysis

Multiple regression analysis was conducted to determine whether Agentic AI Capability, Human-in-the-Loop Oversight, Explainability and Accountability, and Cybersecurity Risk Management significantly predicted Autonomous Cyber-Defense Effectiveness.

Table 4 Multiple Regression Results Predicting Autonomous Cyber-Defense Effectiveness

Predictor	β	t	p
Agentic AI Capability	.32	5.84	<.001
Human-in-the-Loop Oversight	.27	4.91	<.001
Explainability & Accountability	.18	3.36	.001
Cybersecurity Risk Management	.29	5.27	<.001
R^2	.61		
Adjusted R^2	.60		
F	102.47		<.001

The regression model was statistically significant ($F = 102.47$, $p < .001$) and explained approximately 61% of the variance in Autonomous Cyber-Defense Effectiveness ($R^2 = .61$). Agentic AI Capability emerged as a significant positive predictor ($\beta = .32$, $p < .001$), indicating that stronger agentic capabilities were associated with

greater perceived effectiveness of autonomous cyber defense. Human-in-the-Loop Oversight also demonstrated a significant positive effect ($\beta = .27$, $p < .001$), supporting the proposition that appropriately designed human oversight can strengthen rather than impede autonomous cybersecurity operations.

Cybersecurity Risk Management was another significant predictor ($\beta = .29$, $p < .001$), emphasizing that autonomous AI capabilities require systematic risk assessment and governance to translate into effective defensive outcomes. Explainability and Accountability also significantly predicted cyber-defense effectiveness

($\beta = .18$, $p = .001$), although its effect was comparatively smaller. Overall, the findings suggested that effective autonomous cyber defense depended not only on AI capability but also on human oversight, risk management, and accountable decision-making.

Hypothesis Testing

Table 5 Summary of Hypothesis Testing

Hypothesis	Relationship	β	p-value	Decision
H1	AAC $\rightarrow$ ACDE	.32	<.001	Supported
H2	HITLO $\rightarrow$ ACDE	.27	<.001	Supported
H3	EA $\rightarrow$ ACDE	.18	.001	Supported
H4	CRM $\rightarrow$ ACDE	.29	<.001	Supported
H5	ACDE $\rightarrow$ CR	.52	<.001	Supported

The hypothesis-testing results provided overall support for the proposed relationships. Agentic AI capability significantly enhanced autonomous cyber-defense effectiveness, supporting H1. H2 was also supported, demonstrating that human-in-the-loop oversight had a significant positive relationship with autonomous defense effectiveness. This finding was particularly important because it suggested that human involvement did not necessarily reduce automation; instead, structured human intervention appeared to improve the reliability and operational suitability of autonomous responses.

H3 was supported, indicating that explainability and accountability contributed significantly to effective autonomous cyber defense. H4 was also supported, demonstrating the importance of cybersecurity risk management in controlling and optimizing AI-driven defensive operations. Finally, H5 was supported, showing that improved autonomous cyber-defense effectiveness was positively associated with cyber resilience ($\beta = .52$, $p < .001$). Collectively, these findings indicated that autonomous cyber defense was most effective when technological capability was integrated with human governance, risk management, and accountability mechanisms.

The statistical findings demonstrated that agentic AI capability, human oversight, explainability, and cybersecurity risk management jointly contributed to autonomous cyber-defense effectiveness. The relatively high explanatory power of the regression model indicated that these factors represented important determinants of effective AI-enabled defense. Most importantly, the positive effect of HITL oversight suggested that human intervention should not be regarded merely as a restriction on autonomy. Instead, it functioned as a governance and safety mechanism capable of controlling uncertainty, validating high-impact decisions, and preventing potentially harmful autonomous actions.

The findings therefore supported a risk-adaptive autonomy model for Pakistan's critical digital infrastructure. Under such a model, routine and reversible cybersecurity activities could be performed autonomously, whereas high-risk, uncertain, or potentially disruptive actions would require human authorization. This approach would allow organizations to benefit from the speed and scalability of agentic AI while maintaining accountability, operational control, and resilience.

Discussion

The findings demonstrated that agentic AI capability significantly and positively influenced autonomous cyber-defense effectiveness ($\beta = .32, p < .001$). This finding was consistent with Kshetri (2025), who identified agentic AI as a transformative technology capable of automating threat detection, security analysis, decision-making, and incident response. Similarly, recent research has demonstrated the potential of autonomous AI agents to improve cyber-defense operations through adaptive decision-making and automated response. However, the present findings extended previous work by demonstrating that technological capability alone was insufficient; effective autonomous defense also depended on human oversight, risk management, and accountability.

Human-in-the-Loop Oversight demonstrated a significant positive relationship with autonomous cyber-defense effectiveness ($\beta = .27, p < .001$). This finding supported the emerging view that human involvement should remain an integral component of autonomous cybersecurity rather than being regarded solely as a limitation to automation. Burbano et al. (2025) highlighted the possibility that autonomous cyber-defense agents could be manipulated by adversaries, reinforcing the need for mechanisms through which human operators could monitor, constrain, and override AI decisions. The findings therefore suggested that appropriately designed human oversight could reduce the operational risks associated with autonomous decision-making.

Cybersecurity Risk Management also significantly predicted autonomous cyber-defense effectiveness ($\beta = .29, p < .001$). This result reinforced the principles of contemporary AI risk management, which emphasize continuous assessment, governance, monitoring, and accountability throughout the AI lifecycle (Autio et al., 2024). The finding suggested that agentic AI should operate within predefined security policies, authorization boundaries, asset-criticality classifications, and escalation procedures rather than functioning as an unrestricted autonomous system.

Explainability and Accountability significantly contributed to autonomous cyber-defense effectiveness ($\beta = .18, p = .001$). Although its effect was comparatively smaller, the result emphasized the importance of understanding and auditing AI-generated decisions, particularly where autonomous actions could affect critical infrastructure. This finding was consistent with recent literature identifying explainability, safety, and governance as continuing challenges in the deployment of autonomous cyber-defense systems. Finally, autonomous cyber-defense effectiveness demonstrated a strong positive relationship with cyber resilience ($\beta = .52, p < .001$). This finding suggested that effective AI-enabled detection and response could contribute to an organization's ability to anticipate, withstand, respond to, and recover from cyber incidents. The theoretical implication was that cyber resilience should be understood not simply as a technological outcome but as the product of interaction among AI autonomy, human judgment, risk governance, and organizational capabilities.

Overall, the findings supported a risk-adaptive autonomy perspective, whereby the degree of AI independence was determined according to threat severity, confidence, asset criticality, potential operational consequences, and reversibility of the proposed action. This perspective extended existing autonomous cybersecurity research by integrating technological autonomy with human governance within a unified conceptual model.

Conclusion

The study concluded that agentic AI had substantial potential to strengthen autonomous cyber defense within Pakistan's critical digital infrastructure. The findings indicated that agentic AI capability, Human-in-the-Loop oversight, cybersecurity risk management, and explainability were significant contributors to effective autonomous defense, while autonomous cyber-defense effectiveness was strongly associated with cyber resilience.

The findings further demonstrated that complete AI autonomy was neither necessary nor desirable for critical infrastructure environments. Instead, a

risk-adaptive Human-in-the-Loop architecture provided a more balanced approach in which low-risk and reversible actions could be automated, while high-risk, uncertain, or potentially disruptive decisions remained subject to human authorization. Such an approach could enable Pakistan to benefit from AI-driven cybersecurity while maintaining accountability, operational safety, and human control.

Implications

Theoretical Implications

The study contributed to cybersecurity and AI literature by integrating agentic AI, autonomous defense, human oversight, risk management, and cyber resilience within a single framework. It advanced the concept of risk-adaptive autonomy by proposing that AI autonomy should vary according to operational risk rather than remain fixed. The findings also strengthened the theoretical position that human-AI collaboration can function as an architectural capability rather than merely an external governance mechanism.

Managerial Implications

For cybersecurity managers, the findings indicated the need to establish clearly defined autonomy levels and escalation procedures. Organizations should distinguish between actions that could safely be automated and those requiring human approval. Managers should also establish AI governance teams, maintain comprehensive audit trails, and continuously evaluate agent performance, model uncertainty, and operational risks.

Practical Implications

Practically, organizations could deploy agentic AI for continuous monitoring, alert correlation, threat hunting, anomaly investigation, and low-risk containment activities. High-impact actions such as shutting down critical services, modifying core infrastructure, or making irreversible configuration changes should require human validation. Explainable decision interfaces and override mechanisms should also be incorporated into operational security platforms.

Policy Implications

At the policy level, Pakistan should develop clear requirements for the responsible deployment of autonomous AI within Critical Information Infrastructure. Regulatory and cybersecurity authorities should establish standards covering human authorization, AI accountability, auditability, data protection, incident reporting, model testing, access control, and minimum safety requirements for autonomous cyber-defense systems.

Recommendations

1. **Adopt risk-based autonomy:** Organizations should classify AI actions according to severity, reversibility, confidence, and potential operational impact.
2. **Maintain human authorization:** High-risk and irreversible cybersecurity actions should require qualified human approval.
3. **Implement least-privilege access:** Agentic systems should receive only the permissions necessary to perform authorized defensive functions.
4. **Establish continuous monitoring:** AI agents should be continuously monitored for anomalous behavior, model drift, hallucination, manipulation, and unauthorized actions.
5. **Strengthen explainability:** Security teams should receive understandable explanations of AI-generated threat assessments and recommended responses.
6. **Develop rollback mechanisms:** Automated defensive actions should incorporate safe rollback and fail-safe procedures wherever technically feasible.
7. **Conduct adversarial testing:** Organizations should regularly test autonomous agents against prompt manipulation, adversarial inputs, model exploitation, and other emerging attack techniques.
8. **Develop national standards:** Pakistani cybersecurity authorities should establish sector-specific standards for the secure deployment and governance of agentic AI within critical infrastructure.

9. Invest in human expertise: Organizations should strengthen cybersecurity personnel capabilities in AI governance, agentic systems, threat intelligence, and AI-assisted incident response.

10. Establish continuous feedback mechanisms: Human decisions, overrides, and post-incident evaluations should be incorporated into controlled system improvement and model governance processes.

Limitations and Future Directions

The study had several limitations. First, the analysis was based on a cross-sectional research design, which limited the ability to establish causal relationships and observe changes in autonomous cyber-defense capabilities over time. Second, purposive sampling may have limited the generalizability of the findings to the entire Pakistani critical-infrastructure sector. Third, the study relied on professional perceptions rather than direct operational measurements of AI-agent performance. Fourth, the proposed framework was not experimentally tested in a live or simulated critical-infrastructure environment.

Future research should therefore employ longitudinal and experimental designs to evaluate autonomous cyber-defense agents under realistic attack scenarios. Cyber-range and digital-twin environments could be used to compare different autonomy levels and measure detection accuracy, response time, false positives, recovery time, and unintended operational consequences. Future studies should also examine sector-specific applications across banking, telecommunications, energy, healthcare, and government infrastructure.

Further research could investigate multi-agent cybersecurity architectures, adversarial attacks against autonomous agents, explainable AI, AI-agent coordination, human trust and workload, and reinforcement-learning-based adaptive defense. Comparative studies across countries and critical-infrastructure maturity levels would also help determine how the proposed risk-adaptive HITL framework could be generalized and

adapted to different national cybersecurity environments.

REFERENCES

- Adabara, I., Sadiq, B. O., Shuaibu, A. N., Danjuma, Y. I., & Venkateswarlu, M. (2025). A review of agentic AI in cybersecurity: Cognitive autonomy, ethical governance, and quantum-resilient defense. *F1000Research*, 14, 843.
- Al-Hawawreh, M., Baig, Z. A., & Zeadally, S. (2024). AI for critical infrastructure security: Concepts, challenges, and future directions. *IEEE Internet of Things Magazine*, 7(4), 136–142.
- AlQahtani, H., & Ahuja, P. (2026). Secure autonomous cyber defense with LLM agents: A systematic review of autonomy, tool-augmented reasoning, and governance constraints. *Computers & Electrical Engineering*, 135, 111184.
- Ali, S., Wang, J., & Leung, V. C. M. (2025). AI-driven fusion with cybersecurity: Exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms—A comprehensive review. *Information Fusion*, 118, 102922.
- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile*. National Institute of Standards and Technology.
- Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The rise of agentic AI: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9), 404.
- Burbano, L., Sasahara, H., & Cárdenas, A. A. (2025). Steerability of autonomous cyber-defense agents by meta-attackers. In *2025 IEEE Conference on Artificial Intelligence* (pp. 1117–1124).

- Chen, Y., Cui, M., Wang, D., Cao, Y., Yang, P., Jiang, B., Lu, Z., & Liu, B. (2024). A survey of large language models for cyber threat detection. *Computers & Security*, 145, 104016.
- Chowdhry, H., Manero, J., & Sampalli, S. (2025). Evaluating AI agents for cyber defense: A comparison of deep reinforcement learning and LLM approaches. In L. Martínez, D. Camacho, H. Yin, B. Dutta, R. Yera, R. M. Rodríguez Domínguez, & A. J. Tallón-Ballesteros (Eds.), *Intelligent data engineering and automated learning–IDEAL 2025* (pp. 423–433). Springer.
- Govea, J., Gaibor-Naranjo, W., & Villegas-Ch, W. (2024). Transforming cybersecurity into critical energy infrastructure: A study on the effectiveness of artificial intelligence. *Systems*, 12(5), 165.
- Hasanov, I., Virtanen, S., Hakkala, A., & Isoaho, J. (2024). Application of large language models in cybersecurity: A systematic literature review. *IEEE Access*, 12, 176751–176778.
- Kaur, R., Klobučar, T., & Gabrijelčič, D. (2025). Harnessing the power of language models in cybersecurity: A comprehensive review. *International Journal of Information Management Data Insights*, 5(1), 100315.
- Kshetri, N. (2025). Transforming cybersecurity with agentic AI to combat emerging cyber threats. *Telecommunications Policy*, 49(6), 102976.
- Purves, T., Kyriakopoulos, K. G., Jenkins, S., Phillips, I., & Dudman, T. (2024). Causally aware reinforcement learning agents for autonomous cyber defence. *Knowledge-Based Systems*, 304, 112521.
- Raval, K. J., Jadav, N. K., Rathod, T., Tanwar, S., Vimal, V., & Yamsani, N. (2024). A survey on safeguarding critical infrastructures: Attacks, AI security, and future directions. *International Journal of Critical Infrastructure Protection*, 44, 100647.
- Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *Journal of Big Data*, 11, 105.
- Schaberreiter, T., Andriessen, J., Cappiello, C., Papanikolaou, A., & Pardijs, M. (2024). Human-in-the-loop anomaly detection and contextual intelligence for enhancing cybersecurity management. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security* (pp. 127–136).
- Vajpayee, P., & Hossain, G. (2025). Cyber defense through agentic AI enabled automation: An approach to reduce cyber risk. In *Proceedings of the 2025 Computers and People Research Conference* (pp. 1–2).
- Vyas, S., Mavroudis, V., & Burnap, P. (2025). Towards the deployment of realistic autonomous cyber network defence: A systematic review. *ACM Computing Surveys*, 58(1), Article 5.
- Zhou, W., Zhu, X., Han, Q.-L., Li, L., Chen, X., Wen, S., & Xiang, Y. (2025). The security of using large language models: A survey with emphasis on ChatGPT. *IEEE/CAA Journal of Automatica Sinica*, 12(1), 1–26.