Agentic AI, SHAP and LIME for Cyber Threat Pre-emption: Emerging Paradigms in Intelligent Cybersecurity

Mohammed Sadath P[1], Dr R. Kaviyarasi[2]
[1]*Research Scholar, Department of Computer Science and Information Technology, Yenepoya (Deemed to Be University), Bangalore, Karnataka, India.*
[2]*Research Supervisor, Department of Computer Science and Information Technology, Yenepoya (Deemed to Be University), Bangalore, Karnataka, India.*

***Abstract*—The increasing complexity and sophistication of cyber threats have challenged conventional cybersecurity approaches that primarily rely on reactive detection and response mechanisms. The emergence of Artificial Intelligence (AI)-driven cybersecurity has introduced new possibilities for predictive and preventive defence strategies. Among recent technological developments, Agentic Artificial Intelligence (Agentic AI), combined with explainable artificial intelligence (XAI) techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), represents a transformative approach for cyber threat preemption. Unlike traditional AI systems that perform isolated analytical tasks, Agentic AI possesses autonomous decision-making capabilities, enabling continuous monitoring, adaptive learning, threat anticipation, and automated defensive actions. However, the increasing adoption of autonomous AI systems introduces challenges related to transparency, accountability, reliability, and ethical governance. This article examines the role of Agentic AI, SHAP, and LIME in advancing cyber threat preemption by enabling early identification, interpretation, and mitigation of emerging cyber risks. It explores how autonomous AI agents enhance threat intelligence, anomaly detection, vulnerability assessment, and incident response. Furthermore, the study analyses the importance of explainability frameworks such as SHAP and LIME in improving trust, interpretability, and accountability within AI-driven cybersecurity systems. The article argues that integrating autonomous AI agents with explainable machine learning techniques can establish a proactive cybersecurity ecosystem capable of anticipating threats before significant damage occurs. Future cybersecurity frameworks must prioritise human-AI collaboration, ethical AI governance, and transparent decision-making to ensure secure and responsible deployment of intelligent cyber defence systems.**

***Index Terms*—Agentic AI, Explainable Artificial Intelligence, SHAP, LIME, Cyber Threat Preemption, Cybersecurity, Machine Learning, Threat Intelligence, AI Governance.**

I. INTRODUCTION

The rapid digital transformation of contemporary societies has resulted in unprecedented dependence on interconnected technological infrastructures. Organisations, governments, financial institutions, and individuals increasingly rely on digital platforms for communication, data management, and critical operations. However, this expansion of digital ecosystems has simultaneously created complex vulnerabilities, enabling cybercriminals to develop increasingly sophisticated attack strategies. Traditional cybersecurity approaches based primarily on signature-based detection and post-incident response are becoming insufficient against modern threats characterised by speed, adaptability, and complexity.

Cyber threats such as advanced persistent threats (APTs), ransomware, zero-day exploits, supply-chain attacks, and AI-generated cyber offences demonstrate the limitations of conventional security mechanisms. These threats often evade traditional detection systems because they exploit unknown vulnerabilities and operate through subtle behavioural changes rather than recognisable patterns. Consequently, cybersecurity has shifted from reactive defence towards proactive threat

anticipation, commonly referred to as cyber threat preemption.

Artificial Intelligence has emerged as a critical technology in this transition. Machine learning algorithms, deep neural networks, and intelligent analytics systems can process massive volumes of cybersecurity data, identify hidden patterns, and predict potential attacks. However, conventional AI systems often function as passive analytical tools, requiring human intervention for decision-making and response execution. The emergence of Agentic AI introduces a new paradigm by enabling autonomous AI systems capable of perceiving environments, reasoning about threats, planning actions, and executing defensive strategies.

Agentic AI represents a significant advancement because it transforms artificial intelligence from a predictive technology into an active cybersecurity participant. AI agents can continuously monitor networks, analyse threat intelligence, investigate anomalies, coordinate defensive responses, and adapt to evolving attack techniques. Such capabilities have the potential to significantly reduce response times and improve cyber resilience.

Nevertheless, the increasing autonomy of AI systems creates a fundamental challenge: cybersecurity professionals must understand why AI models make particular decisions. Many advanced machine learning models operate as “black boxes,” producing accurate predictions without providing understandable explanations. In high-risk domains such as cybersecurity, unexplained AI decisions may reduce human trust and create difficulties regarding accountability.

Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME address this challenge by providing interpretability mechanisms for complex AI models. SHAP explains model predictions by calculating the contribution of individual features, while LIME creates simplified local explanations of specific predictions. These methods enable cybersecurity analysts to understand AI-generated alerts, validate decisions, and improve human-machine collaboration.

This article examines the integration of Agentic AI with SHAP and LIME for cyber threat preemption. It explores their technological foundations, applications in cybersecurity, advantages, limitations, and future possibilities in establishing transparent, autonomous, and proactive cyber defence systems

II. EVOLUTION FROM REACTIVE CYBERSECURITY TO CYBER THREAT PREEMPTION

Traditional cybersecurity models have historically focused on identifying and responding to threats after malicious activities have occurred. Antivirus systems, intrusion detection systems, and firewalls primarily relied on predefined signatures and known attack patterns. Although effective against conventional threats, these approaches struggle against rapidly evolving attacks.

The emergence of machine learning introduced a significant transformation by enabling systems to detect behavioural anomalies rather than relying exclusively on known signatures. AI-based cybersecurity platforms analyse network traffic, user behaviour, system activities, and threat intelligence data to identify suspicious patterns.

Cyber threat preemption represents the next stage of cybersecurity evolution. Instead of waiting for attacks to occur, preemptive cybersecurity aims to predict potential threats and neutralise them before they cause significant damage. This approach requires continuous intelligence gathering, autonomous reasoning, and adaptive decision-making, which are central characteristics of Agentic AI.

III. AGENTIC AI: AUTONOMOUS INTELLIGENCE FOR CYBER DEFENCE

Agentic AI refers to artificial intelligence systems capable of independently pursuing objectives by observing environments, analysing information, making decisions, and executing actions. Unlike conventional AI models designed for specific tasks, Agentic AI systems demonstrate greater autonomy and adaptability.

In cybersecurity environments, Agentic AI agents can perform several critical functions:

3.1 Autonomous Threat Detection

AI agents continuously analyse network activities, authentication patterns, endpoint behaviour, and communication flows. By identifying deviations from

normal behaviour, they can detect potential cyber threats at early stages.

3.2 Intelligent Threat Hunting

Traditional threat hunting requires significant human effort and expertise. Agentic AI can automate threat discovery by examining large-scale datasets, correlating indicators of compromise, and identifying hidden attack pathways.

3.3 Automated Incident Response

Agentic AI can initiate defensive actions such as isolating compromised systems, blocking malicious traffic, updating security policies, and generating incident reports. This reduces response time and minimises potential damage.

3.4 Adaptive Learning

Cyber threats constantly evolve. Agentic AI systems can continuously learn from new attack patterns and modify defensive strategies accordingly, creating dynamic cybersecurity environments.

Despite these advantages, autonomous cybersecurity systems require transparency and human oversight. This necessity has increased the importance of explainable AI approaches such as SHAP and LIME.

IV. SHAP: ENHANCING INTERPRETABILITY IN AI-BASED CYBERSECURITY

SHAP (Shapley Additive Explanations) is an explainable AI technique based on cooperative game theory. It determines how much each feature contributes to a machine learning model's prediction. In cybersecurity, SHAP helps analysts understand why an AI system identifies specific activities as malicious. For example, if an AI model classifies network behaviour as a ransomware attack, SHAP can identify contributing factors such as unusual file encryption patterns, abnormal login behaviour, or suspicious network communication.

The applications of SHAP in cybersecurity include:

- Malware classification explanation
- Intrusion detection analysis
- Risk assessment
- Fraud detection
- User behaviour analysis
- Vulnerability prioritisation

By improving interpretability, SHAP enables cybersecurity professionals to verify AI decisions and develop more effective defence strategies.

V. LIME: LOCAL EXPLANATIONS FOR CYBER THREAT ANALYSIS

LIME (Local Interpretable Model-Agnostic Explanations) provides explanations for individual AI predictions by creating simplified models around specific decisions. Unlike SHAP, which provides mathematically consistent feature contributions, LIME focuses on explaining particular cases.

In cybersecurity, LIME can assist analysts in understanding why a specific email was classified as phishing, why a user account was identified as suspicious, or why a particular network connection was considered dangerous.

Applications of LIME include:

- Explainable intrusion detection
- Phishing classification
- Malware detection
- Security alert investigation
- User anomaly detection

LIME enhances cybersecurity operations by transforming complex AI outputs into understandable explanations, allowing analysts to make informed decisions.

VI. INTEGRATING AGENTIC AI, SHAP AND LIME FOR CYBER THREAT PREEMPTION

The integration of Agentic AI with explainable AI techniques creates a powerful cybersecurity framework combining autonomy, intelligence, and transparency. Agentic AI provides autonomous reasoning and action capabilities, while SHAP and LIME ensure that these decisions remain interpretable.

A cyber threat preemption framework based on this integration can operate through four stages:

Threat Perception: AI agents continuously collect data from networks, endpoints, cloud systems, and threat intelligence sources.

Threat Prediction: Machine learning models analyse patterns and estimate potential security risks.

Decision Explanation: SHAP and LIME provide explanations regarding the factors influencing AI predictions.

Autonomous Response: AI agents execute appropriate defensive actions while maintaining human supervision.

This integrated approach addresses one of the greatest challenges in cybersecurity: achieving automation without sacrificing transparency.

VII. CHALLENGES AND ETHICAL CONSIDERATIONS

Despite their potential, Agentic AI and explainable AI systems face several limitations. Autonomous decision-making raises concerns regarding accountability, especially when AI actions produce unintended consequences. Determining responsibility between developers, organisations, and AI systems remains a significant legal challenge.

Explainability techniques also have limitations. SHAP and LIME provide useful interpretations but may not always fully represent the internal reasoning processes of complex AI models. Additionally, attackers may attempt to manipulate AI systems through adversarial attacks designed to deceive machine learning algorithms.

Privacy concerns represent another major issue. AI-driven cybersecurity systems require access to large volumes of sensitive information, creating potential risks related to data misuse and surveillance.

Therefore, responsible deployment requires strong governance frameworks, human oversight, continuous validation, and ethical guidelines.

VIII. FUTURE DIRECTIONS

The future of cybersecurity is likely to witness deeper integration between autonomous AI agents, explainable machine learning, and human expertise. Future Agentic AI systems may operate as intelligent cybersecurity assistants capable of predicting threats, coordinating defence mechanisms, and supporting security analysts.

Advancements in quantum computing, federated learning, and adaptive AI models may further enhance cyber threat preemption capabilities. However, technological advancement must be accompanied by ethical standards ensuring transparency, accountability, and responsible AI usage.

The development of global standards for explainable and autonomous cybersecurity systems will be essential for establishing trust and reliability in AI-driven defence environments.

IX. CONCLUSION

The emergence of Agentic AI represents a fundamental transformation in cybersecurity by shifting the focus from reactive defence mechanisms towards proactive cyber threat preemption. Unlike conventional AI systems that primarily analyse information, Agentic AI possesses autonomous capabilities that enable continuous monitoring, intelligent reasoning, and adaptive responses. This evolution provides significant opportunities for addressing increasingly complex cyber threats, including advanced persistent attacks, ransomware, zero-day exploits, and AI-enhanced cybercrime.

However, the growing autonomy of AI-driven cybersecurity systems introduces critical challenges concerning transparency, accountability, and human trust. Explainable AI techniques such as SHAP and LIME provide essential mechanisms for interpreting AI decisions and ensuring that cybersecurity professionals can understand, validate, and effectively utilise automated recommendations. By revealing the factors influencing AI predictions, these approaches strengthen collaboration between human experts and intelligent systems.

The integration of Agentic AI with SHAP and LIME offers a promising framework for developing cybersecurity infrastructures that are not only intelligent but also transparent and accountable. Such systems can identify emerging threats, explain their reasoning, and initiate preventive measures before significant damage occurs. This represents a transition from traditional incident response towards anticipatory cyber defence.

Nevertheless, successful implementation requires careful attention to ethical governance, privacy protection, adversarial resilience, and regulatory compliance. Autonomous cybersecurity systems must operate under appropriate human supervision to prevent unintended consequences and maintain accountability.

In conclusion, Agentic AI combined with explainable AI technologies such as SHAP and LIME represents a significant advancement in cyber threat prevention. By merging autonomous intelligence with interpretability, these technologies have the potential to create a new

generation of cybersecurity systems capable of protecting digital environments against increasingly sophisticated threats.

REFERENCES

[1] Adadi, A., and Berrada, M., "Peeking inside the black-box: A survey on explainable artificial intelligence," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[2] Arrieta, A. B., *et al*., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[3] Lundberg, S. M., and Lee, S.-I., "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[4] Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022.

[5] Ribeiro, M. T., Singh, S., and Guestrin, C., "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[6] Russell, S., and Norvig, P., *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ, USA: Pearson, 2021.

[7] Sarker, I. H., "AI-based cybersecurity: A comprehensive review," *Journal of Information Security and Applications*, vol. 66, pp. 1–14, 2022.

[8] Sommer, R., and Paxson, V., "Outside the closed world: On using machine learning for network intrusion detection," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2010, pp. 305–316.