Explainable and Trustworthy AI in Cybersecurity

A. Research Area / Keywords

Trustworthy AI, Explainable AI, Neuro-Symbolic AI, AI Security, Cybersecurity Analytics, AIoT Security

B. Proposed Start Date

5/1/2026

C. Abstract (≤1000 characters)

Artificial intelligence is increasingly used in cybersecurity applications such as intrusion detection, threat intelligence, and risk analysis. However, many AI models operate as black boxes, making it difficult for security analysts to understand how decisions are made or to verify whether models behave reliably in adversarial environments. This lack of transparency limits trust in AI-driven security systems, particularly in distributed and decentralized infrastructures such as AIoT environments.

This project investigates neuro-symbolic approaches to developing explainable, trustworthy AI models for cybersecurity applications. By integrating neural learning with symbolic reasoning, the research aims to improve model interpretability, robustness, and verifiability. The project will explore applications in risk analysis, AI-based intrusion detection, and security monitoring of AI modules in decentralized environments. The outcome will contribute to the development of transparent and trustworthy AI systems for next-generation cybersecurity infrastructure.

D. Research Background (≤2500 characters)

Artificial intelligence has become an important tool for addressing complex cybersecurity challenges. Machine learning models are increasingly used to analyze large-scale security data, detect network intrusions, identify anomalous behaviors, and support automated risk analysis.

These capabilities are particularly valuable in large and dynamic digital infrastructures where manual monitoring is no longer sufficient.

Despite these advantages, the adoption of AI in cybersecurity raises several important concerns. Many machine learning models, particularly deep learning architectures, function as black boxes whose internal decision-making processes are difficult to interpret. Security analysts often require explanations for alerts and predictions to verify whether the AI system is producing meaningful and reliable results. Without transparency and interpretability, AI-driven security systems may produce false positives, overlook important threats, or become vulnerable to adversarial manipulation.

These challenges become even more significant in distributed environments such as AIoT systems, where AI models operate across heterogeneous devices, networks, and infrastructures. In such environments, AI models may interact with decentralized data sources and autonomous agents, making it difficult to verify their behavior and ensure their security.

Neuro-symbolic AI has recently emerged as a promising approach for addressing these limitations. By integrating neural learning with symbolic reasoning, neuro-symbolic methods combine the pattern recognition capabilities of neural networks with the interpretability and logical structure of symbolic systems. This hybrid approach has the potential to improve the explainability, robustness, and reliability of AI models.

However, the application of neuro-symbolic approaches to cybersecurity remains largely unexplored. There is a lack of frameworks that integrate explainable reasoning mechanisms into AI-based security systems deployed in decentralized and distributed environments. Addressing this gap could significantly improve the transparency and trustworthiness of AI-driven cybersecurity solutions.

E. Objectives and Research Questions (≤2500 characters)

The objective of this research is to develop neuro-symbolic approaches that enable explainable, trustworthy, and secure AI systems for cybersecurity applications in distributed and decentralized environments.

The project seeks to design frameworks that integrate neural learning with symbolic reasoning to improve interpretability, robustness, and decision transparency in AI-based security systems.

The research will address several key questions.

First, how can neuro-symbolic approaches be used to provide interpretable explanations for AI-driven cybersecurity decisions, particularly in tasks such as intrusion detection and threat analysis?

Second, how can symbolic reasoning mechanisms be integrated with machine learning models to support structured risk analysis and security policy reasoning?

Third, how can AI modules deployed in decentralized environments be monitored and evaluated to ensure trustworthy behavior and resistance to adversarial manipulation?

Fourth, how can neuro-symbolic architectures improve the reliability and verifiability of AI models operating in distributed cybersecurity infrastructures?

By addressing these questions, the research aims to develop new methods for building explainable and trustworthy AI systems that support effective cybersecurity decision-making in complex digital ecosystems.

F. Research Methods and Approach (≤3500 characters)

The research will adopt a multi-layered methodological approach that integrates machine learning, symbolic reasoning, and cybersecurity analytics to develop neuro-symbolic frameworks for trustworthy AI systems.

The first phase of the project will focus on analyzing existing AI-based cybersecurity systems and identifying limitations in their explainability, robustness, and trustworthiness. This phase will examine common applications such as network intrusion detection, anomaly detection, and cyber risk analysis to understand how AI models are currently used and where transparency and interpretability gaps exist.

In the second phase, the research will design neuro-symbolic architectures that combine neural learning models with symbolic reasoning mechanisms. Neural models will process large-scale security data to detect complex patterns, while symbolic components will represent security rules, policies, and domain knowledge. These components will enable structured reasoning and provide interpretable explanations for AI decisions.

The third phase will focus on applying the proposed neuro-symbolic frameworks to specific cybersecurity tasks. These may include explainable intrusion detection systems, AI-assisted risk analysis models, and monitoring mechanisms for AI modules operating in decentralized environments. The research will explore how symbolic reasoning can be used to validate AI predictions and support human-understandable explanations.

The fourth phase will investigate security challenges associated with AI modules deployed in distributed environments. The research will explore methods for detecting adversarial manipulation, verifying model integrity, and monitoring AI system behavior across decentralized infrastructures such as AIoT networks.

Finally, the proposed frameworks will be evaluated through experimental studies using real-world or benchmark cybersecurity datasets. Evaluation metrics will include model accuracy, interpretability, robustness to adversarial attacks, and system scalability in distributed environments.

Through this methodology, the research aims to demonstrate how neuro-symbolic AI can enhance the transparency, reliability, and security of AI-driven cybersecurity systems.