MANTECH Publications
Journal of Next-Gen Threat Intelligence and Cyber Defense
Volume 1, Issue 3, September-December 2025

Secure and Explainable AI for Cyber Defense

Desai Mahajan[1], A. K. Verma[2], Sandeep Kulkarni[3]
Associate Professor[1], Students[2, 3]
Department of Information Technology
Ramaiah Institute of Technology
Email id: *a.k.verma650@yahoo.co.in[2]*

ABSTRACT

Artificial Intelligence (AI) has significantly transformed cyber defense by enabling automated threat detection, malware classification, anomaly detection, and real time incident response. However, the integration of AI into cybersecurity infrastructures introduces two major challenges: security of AI models themselves and lack of explainability in complex machine learning systems. Adversarial attacks, model poisoning, data leakage, and model inversion threaten AI-driven security systems. At the same time, black box models such as deep neural networks make it difficult for analysts to interpret decisions, which affects trust, compliance, and accountability. This paper presents a comprehensive review of Secure and Explainable AI (SXAI) in cyber defense. It discusses security risks in AI models, defense mechanisms, explainability techniques, regulatory implications, and future research directions. The study also highlights practical frameworks and real-world applications where secure and interpretable AI can enhance resilience against evolving cyber threats.

KEYWORDS: *Secure AI, Explainable AI, Cyber Defense, Adversarial Attacks, Model Poisoning, XAI, Intrusion Detection Systems, Trustworthy AI*

INTRODUCTION

Cyber-attacks are increasing in frequency and sophistication. Traditional signature based detection systems are not enough to handle zero day exploits, advanced persistent threats (APT), ransomware, and insider attacks. AI and Machine Learning (ML) techniques have

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

emerged as powerful tools in modern cyber defense systems. These systems can analyze massive volumes of network traffic, identify suspicious behavior patterns, and predict potential breaches before damage occurs.

However, AI systems themselves are vulnerable. Attackers are now targeting AI models directly. They manipulate training data, craft adversarial inputs, or exploit model weaknesses. Moreover, many AI based systems operate as "black boxes," where decision making logic is not transparent. In cybersecurity environment, this lack of explainability creates serious problems because security analysts need to understand why a system flagged an activity as malicious.

Secure and Explainable AI (SXAI) aims to address these two issues simultaneously. It ensures that AI models are robust against attacks and their outputs are interpretable and trustworthy.

ROLE OF AI IN CYBER DEFENSE

AI applications in cyber defense span across multiple domains and have significantly changed the way organizations monitor, detect, and respond to cyber threats. Traditional rule-based and signature-driven systems were effective against known attacks, but they struggle when facing zero-day exploits, polymorphic malware, and large-scale distributed attacks. AI systems, on the other hand, can learn patterns from historical data and generalize to unseen scenarios. This capability makes them highly valuable in modern cybersecurity infrastructure.

AI-driven systems are capable of processing massive volumes of network logs, user activity records, and system events in real time. They reduce the workload of security analysts and improve detection accuracy. However, these systems are not flawless and require careful design and monitoring.

1. Intrusion Detection Systems (IDS)

Intrusion Detection Systems are one of the earliest and most important applications of AI in cyber defense. IDS can be categorized into two types:

- **Signature-based IDS**, which detect known threats using predefined rules.
- **Anomaly-based IDS**, which detect deviations from normal behavior.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Machine learning techniques are mainly applied in anomaly-based IDS. These systems learn normal network traffic patterns and identify suspicious activities when deviations occur.

Common algorithms used include:

- **Random Forest** – It is an ensemble learning technique that builds multiple decision trees and combines their outputs. It handles large feature spaces and noisy data effectively.
- **Support Vector Machine (SVM)** – SVM is powerful for binary classification problems and works well with high-dimensional datasets.
- **K-Nearest Neighbors (KNN)** – Often used for anomaly detection in smaller datasets.
- **Deep Neural Networks (DNN)** – Deep learning models can automatically extract complex patterns from raw traffic data without manual feature engineering.

Recent IDS frameworks also use **Recurrent Neural Networks (RNN)** and **Long Short-Term Memory (LSTM)** models to analyze sequential network traffic data. These models are particularly useful for detecting slow and stealthy attacks such as Advanced Persistent Threats (APT).

AI-based IDS systems provide several benefits:

- Reduced false positive rates compared to traditional systems.
- Capability to detect unknown attack patterns.
- Real-time monitoring of high-speed networks.

However, IDS models can be vulnerable to adversarial inputs. Carefully crafted network packets can bypass detection systems, which highlights the importance of secure AI design.

2. Malware Detection

Malware detection is another critical domain where AI has shown strong performance. Traditional antivirus software relies heavily on signature matching. This approach fails against new or obfuscated malware variants.

AI models analyze both:

- **Static features** – Extracted without executing the file (e.g., opcode sequences, byte entropy, file headers, imported libraries).

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

- **Dynamic features** – Collected during runtime (e.g., API calls, system calls, registry changes, memory usage patterns).

Machine learning classifiers such as Random Forest and Gradient Boosting are widely used for malware classification. In recent years, deep learning approaches have become more popular.

Examples include:

- **Convolutional Neural Networks (CNNs)** applied to malware binaries converted into grayscale images.
- **Recurrent Neural Networks (RNNs)** used for analyzing opcode sequences.
- **Graph Neural Networks (GNNs)** used for modeling control flow graphs and call graphs.

AI-based malware detection systems can identify previously unseen malware families by recognizing behavioral similarities. This is particularly useful in combating ransomware and fileless malware.

Nevertheless, attackers have started using AI techniques themselves to generate evasive malware that can bypass ML-based detectors. This creates an ongoing arms race between attackers and defenders.

3. Phishing Detection

Phishing attacks remain one of the most common and successful cyber threats. AI has greatly improved phishing detection through Natural Language Processing (NLP) and behavioral analysis.

Traditional spam filters relied on keyword matching and blacklists. Modern AI systems go beyond this by analyzing:

- Email content semantics
- URL structure and domain reputation
- Sender behavior patterns
- Writing style and tone

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Machine learning models such as Naive Bayes, Logistic Regression, and SVM have been used for phishing classification. More recently, transformer-based models such as BERT and other contextual language models are applied to detect subtle linguistic cues.

AI systems also detect:

- Homograph attacks in URLs
- Domain spoofing
- Social engineering patterns

Behavioral biometrics can be used to detect account takeover attempts by analyzing typing speed, mouse movement, and login patterns.

Despite these advancements, phishing attacks are evolving. AI-generated phishing emails are becoming more sophisticated and personalized, making detection more challenging.

4. Threat Intelligence Automation

Cyber threat intelligence involves collecting and analyzing information about emerging threats, vulnerabilities, and attack campaigns. Manual analysis of threat feeds is time consuming and prone to errors.

AI helps automate this process in several ways:

- **Text mining** from threat reports, blogs, and dark web forums.
- **Clustering attack patterns** to identify related campaigns.
- **Predictive analytics** to forecast future attack vectors.
- **Correlation analysis** between seemingly unrelated events.

Natural Language Processing models extract indicators of compromise (IOCs) such as IP addresses, hash values, and domain names from unstructured text.

Graph-based AI models help map relationships between attackers, infrastructure, and malware families. This allows security teams to understand the broader threat landscape.

AI-powered Security Information and Event Management (SIEM) systems can automatically

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

prioritize alerts based on risk scores. This reduces alert fatigue and helps analysts focus on high-risk incidents.

However, automation must be carefully supervised. Incorrect correlation or biased training data can lead to misleading threat intelligence conclusions.

Here is the elaborated version of **Section 3: Security Threats to AI in Cyber Defense**, written in consistent academic style and slightly natural tone as in your paper.

SECURITY THREATS TO AI IN CYBER DEFENSE

AI systems deployed in cybersecurity are attractive targets for adversaries. Since these systems are responsible for detecting threats, compromising them can allow attackers to bypass defenses silently. Unlike traditional software vulnerabilities, attacks against AI systems exploit statistical weaknesses, data dependencies, and learning mechanisms.

Adversaries may target different stages of the AI lifecycle:

- Data collection
- Model training
- Model deployment
- Inference stage

Understanding these threats is important for designing robust and trustworthy cyber defense systems.

1. Adversarial Attacks

Adversarial attacks involve carefully crafted inputs that are intentionally designed to mislead AI models. These inputs often contain very small perturbations that are not easily noticeable to humans but significantly affect the model's prediction.

For example, in malware detection, an attacker can slightly modify non-functional parts of malicious code—such as adding redundant instructions or renaming variables—without changing behavior. However, these changes may cause the model to misclassify the malware as benign.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Adversarial attacks are generally divided into two categories:

Evasion Attacks (Inference Phase)

Evasion attacks occur after the model is deployed. The attacker manipulates input data to bypass detection.

Examples include:

- Modifying network packet features to avoid intrusion detection.
- Adding benign-looking API calls in malware samples.
- Slightly altering phishing URLs to evade classifiers.

In many cases, attackers use gradient-based optimization techniques to generate adversarial samples if they have knowledge about the model. Even in black-box settings, attackers can approximate model behavior by querying it repeatedly.

The main impact of evasion attacks is misclassification, which leads to false negatives and undetected threats.

- **Poisoning Attacks (Training Phase)**

Poisoning attacks target the training process. The attacker injects malicious or misleading data into the training dataset to influence the learned model.

This can happen in scenarios where:

- Data is crowdsourced.
- Models are trained on publicly available datasets.
- Federated learning is used across multiple participants.

For example, an attacker may insert mislabeled samples into a malware dataset, causing the model to learn incorrect patterns. Over time, this reduces overall accuracy or creates blind spots in detection.

Poisoning attacks are particularly dangerous because they compromise the model before deployment, and detection of such manipulation is often difficult.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

2. Data Poisoning

Data poisoning is a specific and highly impactful attack where adversaries deliberately manipulate training data to degrade model performance or achieve specific malicious objectives.

There are two main goals of data poisoning:

- **Availability Attacks** – Reduce the overall accuracy of the model, making it unreliable.
- **Integrity Attacks** – Cause the model to misclassify specific inputs while maintaining overall performance.

In cyber defense systems, attackers may:

- Inject fake benign samples that resemble malware.
- Insert malicious logs into network datasets.
- Manipulate labels in supervised learning datasets.

For example, if a phishing detection system is trained with poisoned emails labeled as legitimate, the system may start accepting malicious emails as safe.
Data poisoning can also create hidden vulnerabilities that are not immediately visible during evaluation. Since cybersecurity models often rely on continuous retraining with new data, they remain exposed to ongoing poisoning risks.

Mitigation strategies include data validation, anomaly detection during training, and robust learning algorithms. However, completely eliminating poisoning risks remains challenging.

3. Model Inversion and Extraction

Model inversion and model extraction attacks aim to compromise the confidentiality of AI systems.

Model Inversion

In model inversion attacks, adversaries attempt to reconstruct sensitive training data from the trained model. By analyzing model outputs, attackers can infer characteristics of the original dataset.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

For example:

- Reconstructing user behavior patterns from intrusion detection systems.
- Inferring confidential system logs used for training.
- Extracting personal data from models trained on sensitive information.

This is particularly concerning when AI models are trained on private enterprise security logs or user activity records.

Model Extraction

Model extraction attacks attempt to replicate a deployed model by querying it repeatedly through APIs. Attackers send carefully designed inputs and observe outputs to approximate the decision boundaries.

Consequences include:

- Theft of intellectual property.
- Discovery of model weaknesses.
- Easier generation of adversarial examples once the model is replicated.

In cybersecurity applications, if an attacker extracts a malware detection model, they can test different variations of malware offline until they find one that bypasses detection.

Cloud-based AI security services are especially vulnerable to such attacks if proper query rate limiting and monitoring mechanisms are not implemented.

4. Backdoor Attacks

Backdoor attacks are a stealthy form of poisoning where attackers embed hidden triggers in the training data. The model behaves normally under most conditions but produces malicious outputs when a specific trigger is present.

For example:

- A specific sequence of bytes in a malware file causes the model to classify it as benign.
- A particular keyword in an email bypasses phishing detection.
- A unique network packet pattern disables intrusion alerts.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

The trigger is often subtle and difficult to detect. During normal testing, the model may show high accuracy, making the backdoor remain unnoticed.

Backdoor attacks are especially dangerous in supply chain scenarios where pre-trained models are downloaded from external sources. If the model is already compromised, organizations may unknowingly deploy a vulnerable system.

Detection of backdoors requires:

- Rigorous testing with diverse inputs.
- Inspection of model behavior under rare conditions.
- Use of robust training and validation procedures.

Table 1: Common AI Security Threats in Cyber Defense

Threat Type	Target Phase	Impact on System	Example
Evasion Attack	Testing Phase	Misclassification	Adversarial malware
Data Poisoning	Training Phase	Reduced accuracy	Injected malicious samples
Model Extraction	Deployment Phase	Intellectual property theft	API probing
Backdoor Attack	Training Phase	Hidden vulnerabilities	Trigger based behavior

SECURING AI MODELS FOR CYBER DEFENSE

To ensure AI robustness, multiple defensive strategies are implemented.

1. Adversarial Training

Models are trained with adversarial samples to increase resilience.

2. Differential Privacy

Techniques ensure that sensitive training data cannot be reconstructed from the model.

3. Federated Learning

Training is distributed across devices without sharing raw data, reducing exposure risks.

4. Model Monitoring and Logging

Continuous monitoring detects unusual model behavior that may indicate attacks.

5. Secure Model Deployment

Using trusted execution environments (TEE) and hardware security modules protects models in production.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

EXPLAINABLE AI (XAI) IN CYBER DEFENSE

Explainability refers to the ability of AI systems to provide understandable justifications for their predictions.

In cybersecurity, explainability is important because:

- Analysts need to validate alerts.
- Regulatory compliance requires transparency.
- Trust in automated systems must be maintained.

1. Types of Explainability

- **Global Explainability** – Understanding overall model behavior.
- **Local Explainability** – Explaining individual predictions.

2. Explainability Techniques

- LIME (Local Interpretable Model-Agnostic Explanations)
- SHAP (SHapley Additive Explanations)
- Feature Importance Ranking
- Saliency Maps
- Decision Trees as Surrogate Models

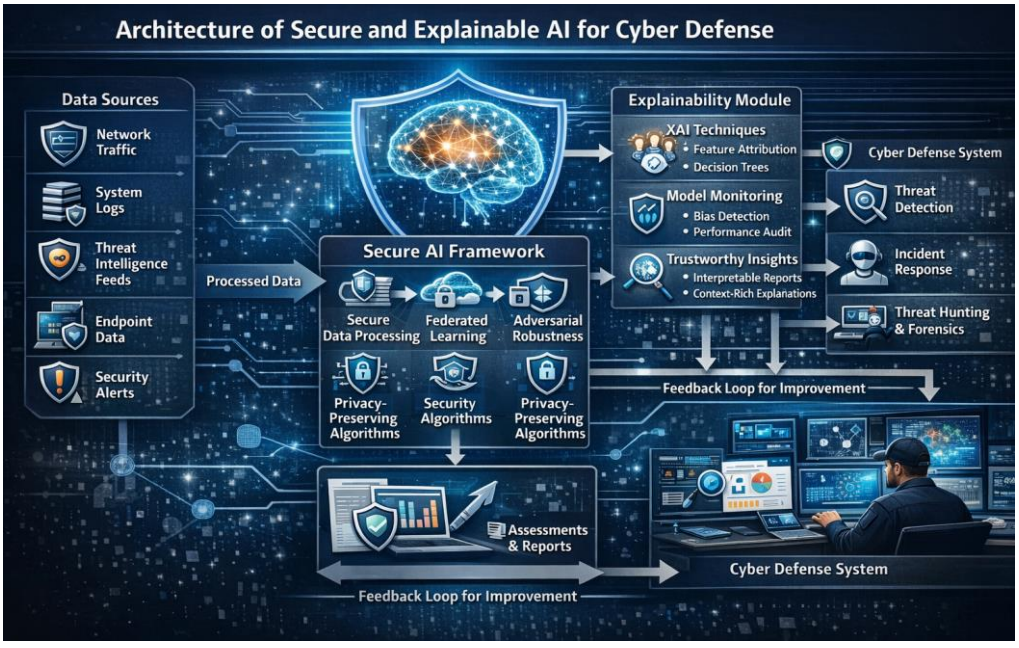


Figure 1: Architecture of Secure and Explainable AI for Cyber Defense

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Integrating Security and Explainability

Secure AI and Explainable AI should not be treated separately. Explainability can also help in identifying adversarial manipulation. For instance, if explanation patterns suddenly change, it may indicate poisoning.

1. Explainability for Threat Investigation

When an intrusion detection model flags traffic, explanation tools highlight suspicious features such as abnormal port number or unusual packet size.

2. Explainability for Compliance

Regulations such as GDPR require automated decision transparency.

3. Trade-off Between Security and Interpretability

Highly complex models (deep learning) provide better accuracy but lower interpretability. Simpler models are interpretable but may lack performance.

Table 2: Comparison of Secure and Explainable AI Approaches

Approach	Security Strength	Interpretability	Computational Cost
Deep Neural Network	High	Low	High
Random Forest	Medium	Medium	Medium
Decision Tree	Low	High	Low
Federated Learning	High	Medium	High

CASE STUDIES

1. AI Based Intrusion Detection in Cloud Systems

A cloud provider deployed deep learning based IDS. After adversarial testing, vulnerabilities were found. By integrating adversarial training and SHAP explanations, detection improved and analysts gained better insight.

2. Ransomware Detection System

An AI model detecting ransomware behavior used explainability tools to highlight abnormal file encryption patterns. This helped analysts quickly validate alerts.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

CHALLENGES

Despite progress, several challenges remain:

- Lack of standardized evaluation metrics for explainability.
- High computational overhead.
- Balancing privacy and transparency.
- Adversarial attacks targeting explanation systems.
- Limited availability of high quality labeled cybersecurity datasets.

FUTURE RESEARCH DIRECTIONS

1. Self-Healing AI Systems

AI systems that automatically detect and repair their vulnerabilities.

2. Quantum-Resistant AI Security

Preparing AI infrastructures against quantum enabled cyber attacks.

3. Explainable Deep Reinforcement Learning

Used in automated incident response systems.

4. Human-AI Collaboration Models

Designing interfaces where analysts and AI systems collaborate effectively.

CONCLUSION

Secure and Explainable AI represents a critical advancement in cyber defense. While AI enhances detection capabilities, it also introduces new vulnerabilities. Ensuring robustness against adversarial threats and providing transparent decision-making are essential for trustworthy cybersecurity systems. Integration of security mechanisms and explainability tools improves analyst confidence and regulatory compliance. However, challenges such as computational cost, trade-offs between accuracy and interpretability, and evolving attack strategies require continuous research. Future cyber defense frameworks must adopt a holistic SXAI approach where security and transparency are built into the system from the beginning rather than added later.

REFERENCES

1. Goodfellow, I., et al. (2015). Explaining and harnessing adversarial examples. ICLR.
2. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial ML. Pattern Recognition.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved

3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining classifier predictions. KDD.
4. Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. NIPS.
5. Papernot, N., et al. (2017). Practical black-box attacks against ML. AsiaCCS.
6. Shokri, R., et al. (2017). Membership inference attacks against ML models. IEEE S&P.
7. Kairouz, P., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in ML.
8. Sommer, R., & Paxson, V. (2010). Outside the closed world: ML for network intrusion detection. IEEE S&P.
9. Sculley, D., et al. (2015). Hidden technical debt in ML systems. NIPS.
10. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable ML. arXiv.

Page 191-204 © MANTECH PUBLICATIONS 2025. All Rights Reserved