World Journal of Advanced Research
worldadvancedjournal.com
Online ISSN: 3107-8966
Received: 07-09-2026, Accepted: 07-10-2026, Published: 22-10-2026
Volume 2, Issue 4, 2026, Page No. 21-24

Explainable artificial intelligence for cybersecurity: Improving threat detection, transparency and trust

Harry Chandra Tanuwidjaja
National Institute of Information and Communications Technology (NICT), Koganei, Tokyo, Japan

Abstract
The increasing sophistication of cyber threats has encouraged organizations to adopt artificial intelligence (AI) and machine learning (ML) for threat detection, intrusion identification, malware classification, and security monitoring. Although AI-based systems can process large volumes of security data and identify complex patterns, many advanced models operate as black boxes, making their decisions difficult for security professionals to understand and validate. Explainable Artificial Intelligence (XAI) addresses this limitation by providing interpretable information about how AI systems reach particular predictions or classifications. This article examines the role of XAI in cybersecurity, focusing on threat detection, transparency, interpretability, and trust. It discusses the application of explanation techniques to intrusion detection, the importance of meaningful explanations for security analysts, the relationship between explainability and model trust, and challenges involving explanation quality, computational cost, adversarial manipulation, and standardization. The article argues that explainability should be considered an integral component of trustworthy AI-based cybersecurity systems rather than merely an additional feature.

Keywords: Explainable artificial intelligence, cybersecurity, machine learning, threat detection, intrusion detection, transparency, trust, interpretability.

Introduction

The rapid expansion of digital infrastructure has increased both the volume and complexity of cybersecurity threats. Organizations increasingly employ machine-learning models to identify malicious activities across networks, endpoints, applications, and other digital environments. These systems can analyze large quantities of security information more rapidly than manual approaches and can identify patterns that may be difficult to recognize through conventional security mechanisms. However, increased predictive capability does not necessarily guarantee that security analysts will understand or trust the decisions generated by AI systems.

The lack of transparency is particularly important in cybersecurity because analysts often need to determine why an event was classified as malicious before initiating an investigation or response. Yan *et al.* (2022) explain that cybersecurity applications require more than a simple prediction because analysts need information about the factors underlying model decisions. Their survey identifies explainability as an important component of successful ML adoption in cybersecurity.

Explainable Artificial Intelligence provides methods for making AI decisions more understandable to human users. In cybersecurity, XAI can help reveal which characteristics of network traffic, user behavior, system activity, or other security data contributed to a threat classification. Consequently, explainability can support transparency, model validation, error analysis, and human decision-making.

Explainable AI and Cybersecurity

XAI refers broadly to approaches that enable AI systems to provide understandable information about their predictions or decisions. Explanations may be generated through inherently interpretable models or through post-hoc techniques applied to more complex models. Yan *et al.* (2022) distinguish between ante-hoc approaches, in which interpretability is incorporated into the model itself, and post-hoc approaches, which attempt to explain an already trained model.

In cybersecurity, XAI has applications in intrusion detection, malware analysis, vulnerability identification, fraud detection, and security-event classification. AI models may identify suspicious activity, but an explanation can provide additional information about the characteristics that contributed to that decision. This distinction is important because security professionals must often investigate the underlying evidence before taking action.

A recent systematic review of XAI in intrusion detection found that explainability can improve transparency and help security professionals understand, evaluate, and optimize AI-based detection systems. The same review also identifies continuing challenges involving standardization, scalability, and real-time explanation generation.

Improving Threat Detection

One of the principal benefits of XAI is its potential to improve the practical usability of AI-based threat detection. A machine-learning model may classify network traffic as malicious, but a security analyst may need to know which characteristics influenced that classification. Explanation techniques can highlight influential features and provide a more detailed interpretation of the model's output.

For example, an explanation associated with an intrusion-detection alert could identify unusual traffic characteristics, abnormal connection behavior, or other relevant indicators that contributed to the prediction. Such information can assist analysts in evaluating whether an alert deserves further investigation.

XAI can also support model debugging. If a model repeatedly relies on irrelevant or misleading features,

analysts and developers can use explanations to investigate potential weaknesses. Yan *et al.* (2022) emphasize that explanations can assist developers in validating model behavior and diagnosing misclassifications.

Therefore, explainability does not necessarily replace detection models; rather, it adds an interpretive layer that can help security professionals understand and evaluate their outputs.

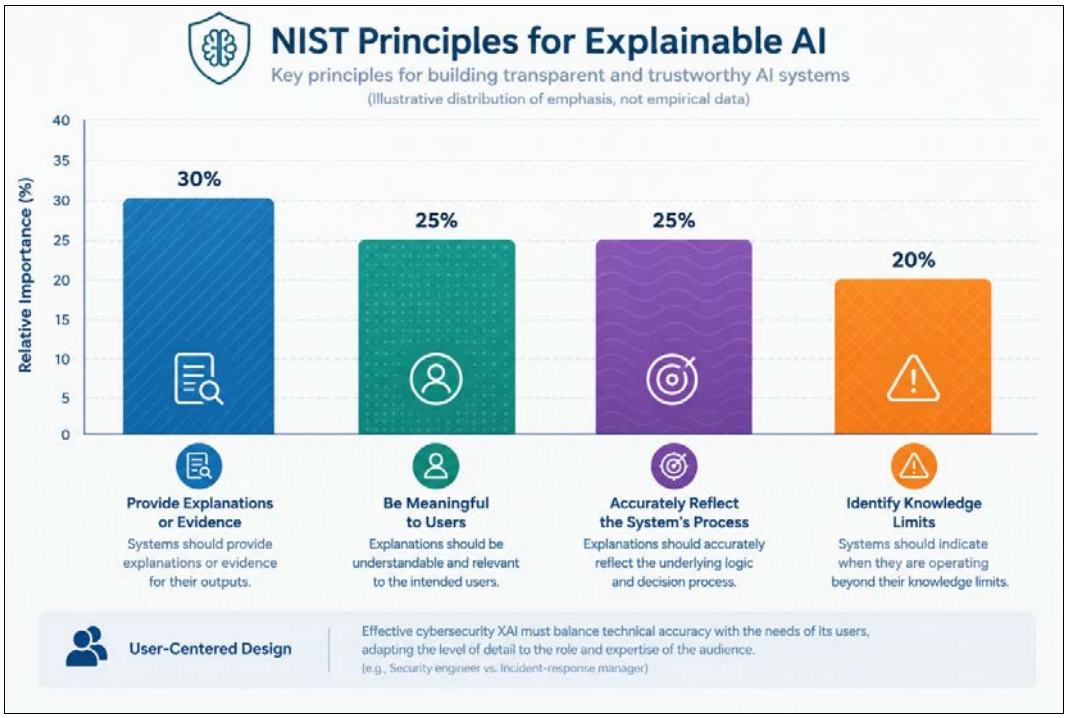
Transparency and Interpretability

Transparency is particularly important when AI systems are used in high-impact cybersecurity operations. Security analysts may be reluctant to rely heavily on predictions from systems whose decision processes cannot be examined or questioned. XAI can reduce this problem by providing explanations that connect model predictions with relevant input characteristics.

NIST identifies four principles for explainable AI: systems should provide explanations or evidence for outputs, explanations should be meaningful to users, explanations should accurately reflect the system's underlying process, and systems should identify situations in which they are operating beyond their knowledge limits.

These principles are highly relevant to cybersecurity. An explanation that is technically detailed but incomprehensible to a security analyst may have limited practical value. Similarly, an explanation should accurately represent why the model generated a particular prediction rather than providing a plausible but misleading justification.

Consequently, effective cybersecurity XAI must consider both technical accuracy and the needs of its intended users. A security engineer may require a more detailed technical explanation than an incident-response manager, meaning that explanation design should be adapted to the user's role and expertise.



Building Trust in AI-Based Security Systems

Trust is an important factor in the adoption of AI-based cybersecurity technologies. Security professionals must be able to evaluate whether a model's predictions are reliable enough to support investigative and operational decisions. XAI can contribute to this process by providing evidence that allows analysts to examine model reasoning.

However, explainability should not automatically be equated with trustworthiness. An understandable explanation does not necessarily mean that a model is accurate or secure. Security teams must also consider detection performance, data quality, robustness, generalization, and the potential for adversarial manipulation.

Research on XAI in cybersecurity emphasizes that explanations can support trust by helping users understand model behavior, but explanation quality itself must be evaluated. Zhang *et al.* (2022) describe XAI as an important mechanism for addressing the transparency and interpretability limitations of AI-based cybersecurity systems and for improving users' ability to understand and manage these systems.

Thus, trust should be viewed as a result of multiple factors rather than as a direct consequence of simply adding an explanation mechanism.

XAI Techniques for Cybersecurity

Several techniques can be used to explain machine-learning predictions. Feature-importance methods identify variables that have contributed significantly to a prediction. Local explanation techniques attempt to explain individual predictions, while global approaches seek to describe the broader behavior of a model. Popular approaches in cybersecurity research include SHAP, LIME, attention-based mechanisms, saliency techniques, surrogate models, and rule-extraction methods.

Recent research on XAI-enabled intrusion detection systems identifies SHAP, LIME, attention-based approaches, saliency methods, and rule-extraction techniques among the major explanation mechanisms used in the field.

The appropriate technique depends on the underlying model, the cybersecurity task, and the needs of the user. A highly complex explanation may provide substantial technical information but may not be useful during a time-sensitive security investigation. Therefore, explanation methods should be selected according to both model characteristics and operational requirements.

Challenges and Limitations

Despite its advantages, XAI introduces several challenges. First, explanations may be computationally expensive, particularly when applied to complex models or large volumes of real-time security events. This can create difficulties for environments that require rapid threat detection.

Second, explanation stability can be problematic. Different explanation techniques may identify different influential features for the same model prediction. Research has identified inconsistencies in explanation methods and emphasized the importance of evaluating explanation reliability rather than assuming that every explanation accurately represents model reasoning.

Third, XAI systems themselves may become targets for manipulation. Attackers may attempt to exploit information exposed by explanations or manipulate the underlying model so that explanations become misleading. Recent research on XAI-enabled intrusion detection highlights adversarial risks, including attacks targeting models and explanation mechanisms.

Another challenge involves the absence of universally accepted metrics for evaluating explanation quality. Traditional cybersecurity metrics such as accuracy, precision, recall, and F1-score evaluate predictive performance but do not necessarily determine whether an explanation is understandable, faithful, stable, or useful to an analyst.

Human-Centered Cybersecurity

The effectiveness of XAI ultimately depends on how well explanations support human cybersecurity professionals. AI should not simply generate technically sophisticated explanations; it should provide information that assists analysts in understanding and evaluating security events.

Human-centered XAI can support collaborative decision-making by allowing analysts to examine the evidence behind predictions, challenge questionable classifications, and combine automated analysis with domain expertise. Such interaction can be particularly valuable when models encounter unusual events or operate in environments different from their training data.

NIST's explainability principles emphasize meaningful explanations and recognition of knowledge limits, reinforcing the importance of designing AI systems around human interpretation and appropriate system boundaries.

Future Directions

Future research in explainable cybersecurity is likely to focus on real-time explanation, adversarially robust XAI, standardized evaluation metrics, and human-in-the-loop security operations. The development of explanations that can be generated rapidly without significantly increasing detection latency will be particularly important for operational environments.

Research should also investigate how explanations can be evaluated consistently across different cybersecurity datasets and AI architectures. Current literature identifies gaps in standardization and scalability, suggesting the need for common evaluation frameworks that assess not only predictive performance but also explanation fidelity, stability, comprehensibility, and operational usefulness.

Another important direction is the integration of XAI with emerging AI-based cybersecurity architectures. Explainability can be incorporated into intrusion detection, threat intelligence, automated incident response, and security operations centers. Such integration could allow analysts to receive both a threat assessment and an explanation of the evidence supporting that assessment.

Conclusion

Explainable Artificial Intelligence provides an important pathway toward more transparent and trustworthy cybersecurity systems. Although machine-learning models can improve the ability to identify complex cyber threats, their practical adoption can be limited when security professionals cannot understand the reasoning behind their predictions. XAI addresses this challenge by providing interpretable information about model decisions and supporting activities such as threat investigation, model validation, error analysis, and security decision-making.

However, explainability alone does not guarantee trustworthy cybersecurity. Effective XAI systems must provide accurate and meaningful explanations while addressing computational requirements, explanation instability, adversarial risks, and the diverse needs of security professionals. Future cybersecurity research should therefore treat explainability as an integral part of AI system design and evaluation. By combining strong predictive performance with transparent and human-centered explanations, XAI can contribute to cybersecurity systems that are not only capable of detecting threats but also understandable and usable by the people responsible for defending digital environments.

References

1. Mohale VZ, Obagbuwa IC. A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence,2025:8:1526221.
2. Pawlicki M, Pawlicka A, Kozik R, Choraś M. The survey on the dual nature of XAI challenges in intrusion detection and their potential for AI innovation. Artificial Intelligence Review,2024:57:330.
3. Phillips PJ, Hahn CA, Fontana AD, Broniatowski DA, Przybocki MA. Four principles of explainable artificial intelligence. National Institute of Standards and Technology, 2020.
4. Yan F, Wen S, Nepal S, Paris C, Xiang Y. Explainable machine learning in cybersecurity: A survey. International Journal of Intelligent Systems,2022:37(12):12305–12334.
5. Zhang Z, Al Hamadi H, Damiani E, Yeun CY, Taher F. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 2022.
6. Zibi Mohale V, Obagbuwa IC. A comprehensive analysis of explainable AI-driven machine learning

techniques for network intrusion detection systems. High-Confidence Computing, 2026.
7. Bhatta RP. Ensemble based machine learning model for prediction of diabetes. NPRC Journal of Multidisciplinary Research,2026:3(2):116-129.
8. Rjoub G, Bentahar J, Wahab OA, Mizouni R, Song A, Cohen R, et al. A survey on explainable artificial intelligence for cybersecurity. IEEE Transactions on Network and Service Management,2023:20(4):5115-5140.
9. Rahman MA, Haque BT. Adaptive Cybersecurity Threat Intelligence Using Explainable Artificial Intelligence for Resilient Protection of US Critical Information Systems. American Journal of Advanced Technology and Engineering Solutions,2025:1(02):216-261.
10. Alshudukhi KS, Ali S, Humayun M, Alruwaili O. Next-generation lightweight explainable AI for cybersecurity: A review on transparency and real-time threat mitigation. Computer Modeling in Engineering & Sciences,2025:145(3):3029-3085.
11. Alketbi KS, Mehmood A. A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. IEEE Access,2025:13:121772-121798.
12. Almheiri SJ, Shah AA, Abbas S, Ahmad M, Khan MA. Smart sustainable cyber security: modelling an interpretable and transparent threat detection with explainable artificial intelligence. Discover Sustainability,2025:6(1):442.
13. Salloum S, Norozpour S. XAI-IDS: A transparent and interpretable framework for robust cybersecurity using explainable artificial intelligence. Shifra, 2025, 69-80.
14. Desai B, Patil K, Mehta I, Patil A. Explainable AI in Cybersecurity: A Comprehensive Framework for enhancing transparency, trust, and Human-AI Collaboration. In 2024 international seminar on application for technology of information and communication (iSemantic). IEEE, 2024, 135-150.
15. Islam MT, Syfullah MK, Rashed MG, Das D. Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable AI. International Journal of Machine Learning and Cybernetics,2024:15(11):5337-5360.