DOI: 10.71317/jgst.2.11.2026.993

Journal of Global Social Transformation

Journal homepage: https://rjsaonline.org/index.php/JGST

Explainable Artificial Intelligence (XAI) for Detecting Cybersecurity Threats: Enhancing Trust, Transparency, and Accuracy in AI-Driven Security Systems

Danish Hameed[1], Amna Mushtaq[2], Waleed Wahab[3]

[1] CTO at Arhamsoft

Email: Danish.hameed@arhamsoft.com

[2] Teaching Assistant, Government College University Faisalabad

Email: amna_mushtaq81@yahoo.com

[3] Masters Student in Computer Networking and IOT | Conservatoire National des Arts et Métiers (Cnam), Paris, France

Email: walees.wahab321@gmail.com

ARTICLE INFO

Submission:
September 27, 2026
Revised:
October 17, 2026
Accepted:
October 30, 2026
Available Online:
November 07, 2026

Keywords:
Explainable Artificial Intelligence (XAI); Cybersecurity Threat Detection; Trustworthy AI; Interpretability; Transparency; Human Oversight; AI-Driven Security Systems.

Corresponding author:
Danish Hameed
Email:
Danish.hameed@arhamsoft.com

ABSTRACT

Artificial intelligence (AI) is increasingly used in cybersecurity to identify unusual patterns, detect complex attacks, and support responses to emerging threats. However, the limited transparency of many AI-powered security systems raises concerns about trust, accountability, and reliability in security operations. Explainable artificial intelligence (XAI) seeks to make AI-generated predictions and decisions more understandable to human users. Its usefulness depends on the fidelity, clarity, and security of the explanations it provides. This paper examines how XAI may support trust and transparency in AI-driven cybersecurity threat detection and considers its relationship with predictive accuracy and human decision-making. The study adopts a qualitative, interpretivist, exploratory, and documentary approach based on secondary evidence, including previous research, cybersecurity standards, technical publications, and relevant institutional reports. Thematic analysis is used to identify recurring patterns, conceptual relationships, implementation challenges, and governance implications. Five interconnected dimensions are examined: (1) the interpretability of threat-detection decisions; (2) informed trust and human oversight; (3) false-positive and false-negative detection; (4) complex and evolving cyber threats; and (5) governance concerns involving explanation fidelity, privacy, adversarial manipulation, and model complexity. A central finding is that explainability does not necessarily improve predictive accuracy or cybersecurity performance. Rather, it can give security analysts greater insight into system outputs, helping them assess automated recommendations and make informed decisions within appropriate operational and governance limits. The study offers a conceptual framework for understanding how humans and AI can work together in cybersecurity threat detection and how AI can be governed and used responsibly. It concludes that explainability should be combined with comprehensive security assessments, privacy safeguards, and meaningful human oversight to support more transparent, trustworthy, and reliable cybersecurity systems.

Introduction

As digital infrastructures become more complex and interconnected, and cyber threats become increasingly sophisticated, reliable and adaptable cybersecurity mechanisms are becoming more critical. Artificial intelligence (AI) and machine learning (ML) are increasingly used in cybersecurity to identify suspicious activities, detect anomalous network behaviour, classify malicious software, and support intrusion detection. These technologies can help security professionals identify patterns that may not be readily apparent through rule-

based approaches. However, their performance depends on data quality, model design, the nature of emerging threats, and the conditions in which they are used. AI can enhance cybersecurity, but it should not replace traditional security measures or professional expertise. In this context, explainable artificial intelligence (XAI) has attracted considerable academic attention because it addresses a key challenge associated with sophisticated AI models: understanding the processes underlying automated decision-making (Charmet et al., 2022; Ferrag et al., 2023).

AI-powered cybersecurity extends beyond threat detection and network classification to the use of intelligent systems in broader security operations. Security analysts need to understand why an activity has been flagged, whether it represents a genuine threat, and whether a response is necessary. These responsibilities become particularly challenging when systems generate large numbers of alerts or operate in environments where attack techniques change rapidly. False-positive alerts can consume investigative resources, disrupt normal workflows, and contribute to alert fatigue; false negatives can allow malicious activity to go undetected. Although AI algorithms may identify potentially harmful behaviour, an alert stating only that a threat has been detected does not explain the evidence behind that conclusion. Effective cybersecurity therefore requires more than predictive performance. It also depends on the comprehensibility of supporting information, human review, and confidence in the reliability of system outputs. In cybersecurity, explainability concerns the interpretability of AI-based systems' automated decisions and the security of the mechanisms used to generate explanations (Charmet et al., 2022; Ferrag et al., 2023).

Many machine-learning models, particularly deep neural networks, operate as “black boxes”: they may identify relationships among input features without making those relationships readily understandable to users. In cybersecurity, analysts must determine whether an alert represents a meaningful threat signal, an unusual but legitimate activity, a data-quality problem, or an unintended combination of features. Such uncertainty can complicate incident analysis and information sharing, while making accountability for AI-assisted decisions more difficult to establish. Pawlicki et al. (2024) discuss technical and operational challenges associated with explainability in intrusion detection. Charmet et al. (2022) similarly examine the applications and limitations of XAI in cybersecurity. Together, these contributions highlight the need to consider the interpretability of security decisions alongside the reliability and security of the mechanisms that generate explanations.

Explainable artificial intelligence (XAI) refers to a family of techniques and approaches intended to make AI-generated predictions, classifications, and decisions more understandable to people. Its purpose is not simply to provide more information about a model, but to offer insight into the factors or reasoning associated with a particular output. The National Institute of Standards and Technology (NIST) report by Phillips et al. (2021) outlines principles concerning the evidence supporting explanations, users' understanding of those explanations, their faithfulness to the system's decision-making process, and the conditions under which the system operates. These principles are particularly important in cybersecurity: explanations should be understandable without misrepresenting the underlying model or creating unjustified confidence in its results. XAI therefore offers conceptual and technical approaches that security professionals can use to interpret automated threat assessments and judge their suitability for operational use.

XAI methods can broadly be divided into two categories: inherently interpretable models and post-hoc explanation methods. Decision trees and rule-based systems are examples of models whose decision processes may be relatively understandable to users. Post-hoc methods, by contrast, seek to explain the outputs of models that are not inherently easy to interpret. Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) can highlight features associated with a particular prediction or patterns in a model's outputs (Pawlicki et al., 2024). In cybersecurity, these techniques may be used to examine network traffic, investigate unusual user activity, and analyse events considered suspicious. However, an explanation generated after a prediction should not be treated as a complete description of the model's internal decision process. Its usefulness depends on the method, the model, the quality of the underlying data, and the question the explanation is intended to address (Charmet et al., 2022).

The relationship between explainability and trust is another important aspect of AI-powered cybersecurity. Security professionals must be able to assess automated recommendations, decide whether to escalate alerts, and determine when human intervention is necessary. Clear explanations may support communication among technical analysts, security managers, and other organizational stakeholders. However, trust should not mean simply accepting an AI system's outputs. An explanation may appear convincing while remaining incomplete, unstable, or inaccurate in representing the model's behaviour. Phillips et al. (2021) emphasize that an explanation should accurately reflect the process by which an AI system produces its output. This distinction separates genuine transparency from explanations that merely sound persuasive. The NIST AI Risk Management Framework provides broader guidance for managing AI trustworthiness throughout design, deployment, use, and evaluation (NIST, 2023). In this context, the purpose of explainability is to support appropriately calibrated trust, human oversight, and critical evaluation of AI-generated outputs.

The relationship between XAI and detection accuracy must also be considered carefully. Predictive accuracy concerns the correctness of a model's predictions, whereas explainability concerns the extent to which the reasons or factors behind those predictions can be understood. These dimensions are related but not interchangeable. An interpretable model may still misclassify events, while a model with strong predictive performance may produce explanations that are difficult to understand or evaluate. XAI may be useful when analysts need to verify an automated classification, investigate conflicting features, or examine a false-positive or false-negative result. However, an explanation alone cannot establish whether a model is accurate or whether its use improves an organization's security.

Existing research raises questions about the evaluation and trustworthiness of explanation methods and their applications (Charmet et al., 2022; Pawlicki et al., 2024). XAI should therefore be assessed as a potential aid to security decision-making, not as proof that detection performance has improved.

Privacy, adversarial security, and the complexity and evolution of cyber threats create further challenges for explainability. Cybersecurity data may reveal network patterns, user behaviour, system configurations, or organizational practices. Explanations that disclose too much information may therefore create privacy or security risks, particularly if they are accessible to unauthorized parties. In addition, attackers may exploit knowledge of a detection system or manipulate its inputs to influence its predictions. Explanation mechanisms can also increase computational demands or produce inconsistent explanations across different conditions. Charmet et al. (2022) examine security issues associated with XAI techniques, while Pawlicki et al. (2024) discuss broader challenges in explainable intrusion detection. These concerns suggest that explanations should be evaluated in terms of clarity, fidelity, consistency, privacy protection, adversarial robustness, and operational suitability. The central challenge is to provide meaningful information about AI-generated security decisions without creating vulnerabilities or weakening detection capabilities.

Although substantial research has examined XAI techniques and their potential cybersecurity applications, important questions remain about their practical use. Charmet et al. (2022) review explainability in cybersecurity tasks and the security of XAI methods, while Ferrag et al. (2023) survey explainable AI approaches from a cybersecurity perspective. Pawlicki et al. (2024) also discuss challenges associated with XAI. These contributions provide a basis for examining technical interpretability alongside human trust, transparency, reliability, and accountability. The present study does not assume that previous research has ignored these relationships. Instead, it brings relevant issues together in a qualitative discussion, focusing on their theoretical connections, practical implications, and unresolved tensions. It treats explainability not merely as an additional technical feature, but as one element of responsible cybersecurity decision-making.

This study critically examines the role of explainable artificial intelligence in supporting trust, transparency, and responsible interpretation of AI-powered cybersecurity threat-detection systems. It adopts a qualitative, interpretivist, exploratory, and documentary research design based on secondary evidence. The analysis draws on existing scholarly literature, relevant cybersecurity standards, technical publications, and institutional reports to interpret recurring concepts, challenges, opportunities, and governance implications. The study is guided by the following questions: What challenges arise when AI is used for cybersecurity threat detection? How can XAI support the interpretation of cyber-threat detection decisions? Which aspects of XAI remain insufficiently explored? How are explainability, security, and human oversight related in cybersecurity? What implications does XAI have for detection accuracy, reliability, false positives, and false negatives? Finally, what technical, ethical, operational, and governance issues constrain the implementation of explainable cybersecurity systems?

The study addresses four objectives: (1) to conceptualize the use of XAI in interpreting cybersecurity threats; (2) to critically analyse the relationship between XAI, trust, transparency, and human oversight; (3) to examine explanations as tools for interpreting threats in relation to detection accuracy and reliability, without assuming that detection performance improves; and (4) to investigate the technical, ethical, operational, and governance challenges associated with explainable security systems. Through documentary analysis and thematic interpretation, the study seeks to clarify these dimensions and their implications for responsible cybersecurity practice.

The significance of this study lies in its effort to connect the technical, human, and organizational dimensions of XAI with the needs of cybersecurity decision-making. It provides a conceptual basis for understanding how AI-assisted threat identification can be interpreted, evaluated, and used responsibly. The discussion is relevant to cybersecurity analysts, system architects, organizational decision-makers, and policymakers concerned with the responsible use of AI in security operations. The study does not present experimental results or validate a specific model; instead, it synthesizes documentary evidence about the potential capabilities and limitations of XAI. It recommends integrating explainability with security evaluation, privacy protection, adversarial-risk management, and human intervention. This measured approach recognizes the potential of XAI to support transparency and trust without overstating its contribution to threat-detection performance.

Literature Review

The Conceptual Foundation and Evolution of Explainable Artificial Intelligence (XAI)

The increasing complexity of machine-learning systems and the difficulty of understanding their outputs have contributed to the emergence of explainable artificial intelligence (XAI) as a research area. As AI is used more widely in domains such as cybersecurity, understanding how a model arrives at a prediction becomes important to assessing its usefulness, reliability, and suitability for operational decisions. In cybersecurity, explainability may inform incident prioritization, threat investigation, and defensive action. The conceptual framework of XAI must, however, distinguish among several related but different concepts. Interpretability concerns how readily a model's structure or behaviour can be understood, while explainability concerns the methods used to communicate or justify an AI output. Transparency involves visibility into a system's design, operation, limitations, and decision-making. Accountability concerns responsibility for decisions and outcomes, while trustworthiness encompasses reliability, safety, security, and appropriate human oversight.

The principles described by the National Institute of Standards and Technology (NIST) provide a useful introduction. Phillips et al. (2021) identify four principles of explainable AI: (1) an explanation should provide evidence or reasons for an output; (2) it should be understandable to its intended audience; (3) it should accurately reflect the system's decision-making process; and (4) it should operate within the system's knowledge limits and stated conditions. These principles show that explainability involves more than making an algorithm's output visually accessible. An explanation should be understandable and provide a reasonably faithful account of the process it describes. This distinction is particularly important in cybersecurity: a misleading threat alert may send an incident-response team in the wrong direction or create complacency when an alert is incorrectly dismissed.

XAI literature commonly distinguishes between inherently interpretable models and post-hoc explanation methods. Inherently interpretable models, such as decision trees, can offer relatively direct insight into how decisions are made. When a model is difficult to interpret, post-hoc techniques may be used to explain its predictions. LIME (Ribeiro et al., 2016) uses a locally approximating model to explain an individual prediction. Lundberg and Lee (2017) introduced SHAP, a framework that attributes a prediction to input features using Shapley values. Both methods have been applied to explanation tasks involving complex machine-learning models, including cybersecurity applications.

These methods also have limitations. A local explanation may not represent model behaviour in other cases, and feature attribution cannot establish causal relationships or reveal the entire decision-making process. The importance assigned to a feature should therefore not be treated as proof that the feature caused a malicious classification. Charmet et al. (2022) explain that XAI methods in cybersecurity should be evaluated in relation to the models and security tasks for which they are used. Explainability is therefore not a single property; it involves several qualities, including fidelity to the model, coverage, understandability, and suitability for the intended task.

AI and machine-learning tools are widely studied in cybersecurity because they can help analyse large volumes of network traffic, system events, and other complex security data. Their applications include intrusion detection, malware identification, phishing detection, anomaly recognition, and network-traffic classification. Traditional rule-based security methods can recognize known patterns and enforce predefined policies. Machine-learning methods can also identify statistical relationships and patterns that may not have been specified in advance through manually written rules. Deep-learning techniques extend these capabilities by learning complex representations from data, although their suitability depends on the task, available information, and deployment environment.

Despite advances in AI-based detection, longstanding cybersecurity challenges remain unresolved. Threat behaviour changes, legitimate activity may resemble malicious activity, and threat information may be incomplete, noisy, or unrepresentative. Differences in network configurations, user behaviour, software, and attack strategies can also limit the transferability of models between environments. Class imbalance and the absence of examples of emerging attacks may further complicate performance evaluation. These limitations demonstrate why published detection results should be evaluated independently rather than treated as evidence of general superiority.

Ferrag et al. (2023) survey explainable AI in cybersecurity, categorizing threats, security concerns, explanation approaches, and outstanding challenges. Zhang et al. (2022) provide an overview of research on XAI applications in cybersecurity and demonstrate the range of tasks in which interpretability has become relevant. Explainability is therefore not limited to one detection method or security application. However, an explanation technique suitable for one task may not be appropriate for another. For example, explaining an anomalous network connection may require different contextual information from explaining a suspicious user login or a malware classification.

An AI-driven security system may generate alerts without providing enough information for analysts to determine why an event is suspicious or how urgently it should be investigated. False positives can increase investigators' workloads and contribute to alert fatigue, while false negatives can allow malicious activity to remain undetected. Although accuracy remains important when evaluating detection systems, operational reliability also depends on the quality of the evidence, the relevance of alerts to the security context, and analysts' ability to assess system outputs. The literature therefore supports integrating explainability with predictive performance and established security requirements rather than treating it as a replacement for them.

Using Explainable AI in Cybersecurity Threat Detection

Research on XAI in cybersecurity examines how the outputs of security systems can be explained to analysts. In intrusion detection, explanations may identify features associated with a suspicious classification and help analysts investigate why a network event was flagged. In malware analysis, explanation techniques may highlight model features associated with a classification. In phishing detection and anomaly identification, explanations may indicate which attributes contributed to a prediction. These applications illustrate how XAI can connect automated classification with human investigation. Their usefulness depends, however, on whether the identified features are meaningful in the security domain under consideration.

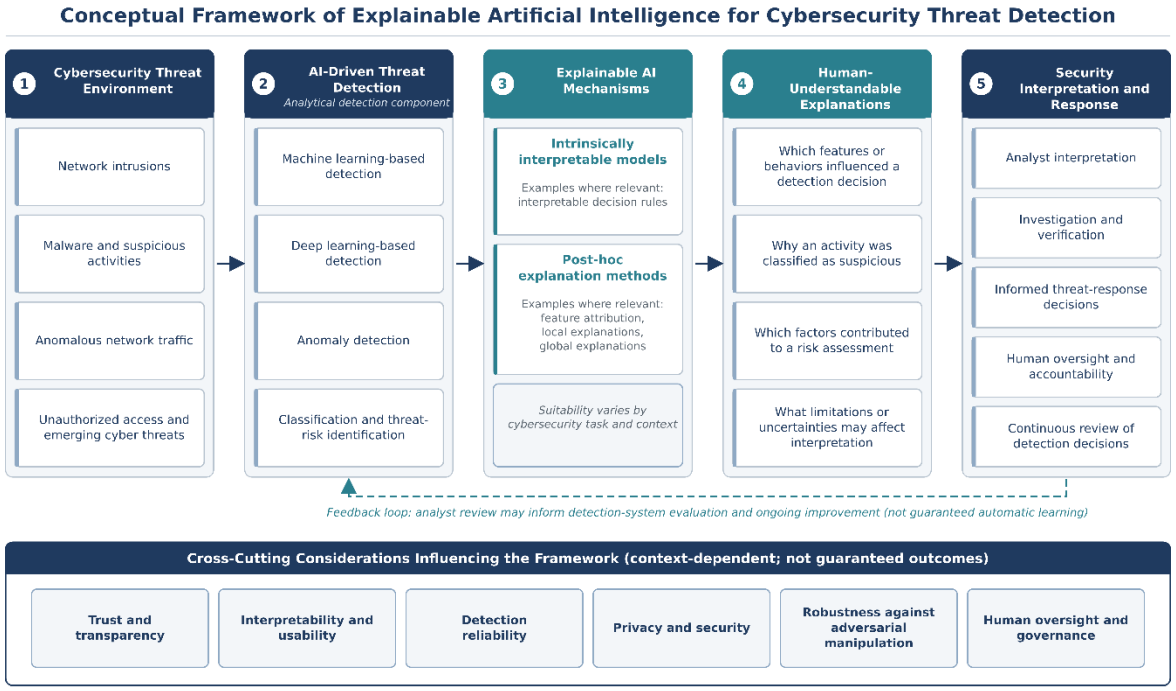
Charmet et al. (2022) provide an overview of XAI in cybersecurity and highlight challenges involving the interpretability and security of AI-powered solutions. Their discussion forms part of a broader effort to understand how people interpret security-related predictions and which explanations may support that process. Zhang et al. (2022) examine the development of explainability in cybersecurity

applications, while Ferrag et al. (2023) classify the literature according to cybersecurity threats, explanation methods, and limitations of existing approaches. Together, these contributions provide a foundation for studying XAI in threat detection.

An important distinction concerns local and global explanations. Local explanations address an individual prediction, such as why a particular network event was classified as suspicious. Global explanations describe model behaviour at a broader level, for example by identifying features that commonly influence predictions. Both forms can support security analysis, but neither independently establishes model reliability. A local explanation may help with an immediate investigation, whereas a global explanation may provide broader insight into model behaviour without explaining every individual incident.

The literature also identifies challenges in making explanations useful to different audiences. A machine-learning specialist may understand a technical feature name or attribution score, while a security analyst may find it difficult to translate that information into an operational response. Conversely, simplifying an explanation may conceal technical details or uncertainty. Pawlicki et al. (2024) discuss these challenges in the development and implementation of XAI. Their analysis reinforces the need to consider the intended user, security task, and operational context when designing explanations.

Another issue concerns the relationship between an explanation and the model it describes. Post-hoc explanations can be useful without providing an exhaustive or fully accurate account of the model's decision-making process. Explanations may also be unstable if small changes in the input produce substantially different results. In addition, changes in attack patterns or operating conditions may make explanations produced during an earlier evaluation less informative. XAI methods should therefore be evaluated for fidelity, stability, relevance, comprehensibility, and usefulness in relation to the specific cybersecurity decisions they are intended to support.



Explainability, Trust, Transparency, and Human Oversight

Explainability and trust are important to the responsible use of AI in cybersecurity. Security analysts need to assess whether AI-generated alerts are genuine, whether the supporting information is adequate, and whether a suggested action is appropriate. Explanations may help identify factors associated with a machine-generated classification and provide a basis for further investigation. Trust, however, should reflect the demonstrated capabilities and limitations of a system rather than follow automatically from the presence of an explanation.

Phillips et al. (2021) emphasize the importance of clear and understandable explanations. Tabassi (2023), by contrast, places AI trustworthiness within a broader framework for identifying, assessing, and managing risks throughout an AI system's lifecycle, as described in the NIST AI Risk Management Framework. These perspectives are related but distinct: the former focuses on explainability, while the latter addresses AI risk management more broadly. In cybersecurity, they suggest that explanations should be considered alongside system reliability, security, uncertainty, and the organizational processes through which AI-assisted decisions are reviewed.

The literature also highlights the risk of over-reliance on automated recommendations and the importance of justified trust. An analyst may accept a model's output because its explanation appears plausible, even when that explanation is incomplete or incorrect. This can result in automation bias, in which automated recommendations are treated as more reliable than the available evidence warrants. Conversely, highly technical or ambiguous explanations may lead users to disregard potentially useful outputs. Trust should therefore be calibrated to what the system can demonstrate, its known limitations, and the quality of the available information.

Human oversight is essential. AI-generated explanations may inform an investigation, but analysts remain responsible for interpreting the information, questioning doubtful classifications, and deciding what action to take. Oversight requires more than permission to approve an automated decision. Reviewers need sufficient knowledge, authority, time, and access to relevant information to challenge the system when necessary. The relationship between XAI and human judgement is therefore sociotechnical, involving algorithms, users, organizational processes, and security objectives.

Transparency has several dimensions. Model-level transparency concerns how a system functions; explanation-level transparency concerns the reasons provided for specific outputs; operational transparency concerns how outputs are incorporated into security workflows; and institutional transparency concerns responsibility for decisions and actions based on those outputs. Structured management is also important for establishing and continually improving an organization's use of AI, as reflected in responsible AI management standards (International Organization for Standardization [ISO], 2023). However, governance standards alone do not demonstrate that an explanation is faithful or that a threat-detection model is effective. Technical and institutional safeguards should be considered together.

It is important to distinguish explainability from predictive accuracy. Accuracy concerns how often predictions are correct under specified evaluation conditions, whereas explainability concerns how well the factors or processes behind a prediction can be understood. An accurate model need not be interpretable, and an interpretable model may still make incorrect predictions. An explanation should therefore not be treated as proof that a threat classification is correct or that one security system is superior to another.

XAI may support investigations of false positives and false negatives. When legitimate activity is incorrectly classified as malicious, an explanation may help analysts examine the alert's features and determine whether the classification reflects a misleading pattern. When a threat is missed, information about model outputs may help technical teams identify gaps in the model's features or decision process. However, explanations generated for predictions do not automatically explain events that the system failed to flag. Missed threats may require alternative evidence, analytical methods, and investigation of system blind spots.

The distinction between explanation quality and system performance is important when interpreting published studies. Charmet et al. (2022), Ferrag et al. (2023), and Pawlicki et al. (2024) discuss different aspects of XAI in cybersecurity, including applicability and implementation challenges. Their work can inform future assessments of explanation quality as a distinct dimension of security-system evaluation. However, the use of explanation methods does not by itself demonstrate that analysts make fewer errors, investigate alerts more quickly, or interpret results more accurately. Such claims require appropriate operational or experimental evidence.

Reliability concerns the consistency and dependability of a system's performance under the operating conditions for which it is intended. A security system should perform appropriately in those conditions, provide explanations that are sufficiently supported, and communicate uncertainty where necessary. The fidelity, stability, and consistency of explanations, their comprehensibility, and their computational feasibility are therefore important evaluation criteria. An explanation method that poorly represents the model's behaviour may be unsuitable for consequential security decisions. Similarly, explanations that arrive too late or require excessive analyst time may have limited operational value.

The relationship between complex and interpretable models adds further complexity. Models capable of representing intricate patterns may be difficult to understand, whereas simpler models may be easier to interpret but may not capture all relevant patterns in a particular security context. The choice is not simply between accuracy and transparency; it depends on the intended use, the consequences of errors, the available contextual evidence, and organizational needs. XAI should therefore form part of a broader security-assurance approach that considers predictive performance, operational suitability, uncertainty, and human review.

One of the major challenges in developing and implementing XAI in cybersecurity is explanation fidelity. Useful explanations should accurately represent the model behaviour they are intended to describe. Feature-attribution techniques can identify attributes associated with a prediction, but these associations should not be interpreted as evidence of causation. If an analyst receives an approximate explanation rather than a faithful account of the model's reasoning, the analyst may draw incorrect conclusions about a threat or overlook weaknesses in the detection process. Phillips et al. (2021) identify fidelity as a key component of explainability, not an optional feature.

Adversarial security presents a related challenge. Malicious actors may deliberately alter inputs, exploit weaknesses in an environment, or change their behaviour in response to security measures. Explanation methods may themselves be manipulated or may disclose information that helps an adversary understand a detection system. Charmet et al. (2022) and Pawlicki et al. (2024) discuss security issues associated with XAI and intrusion detection, emphasizing the importance of considering explanations within the system's threat

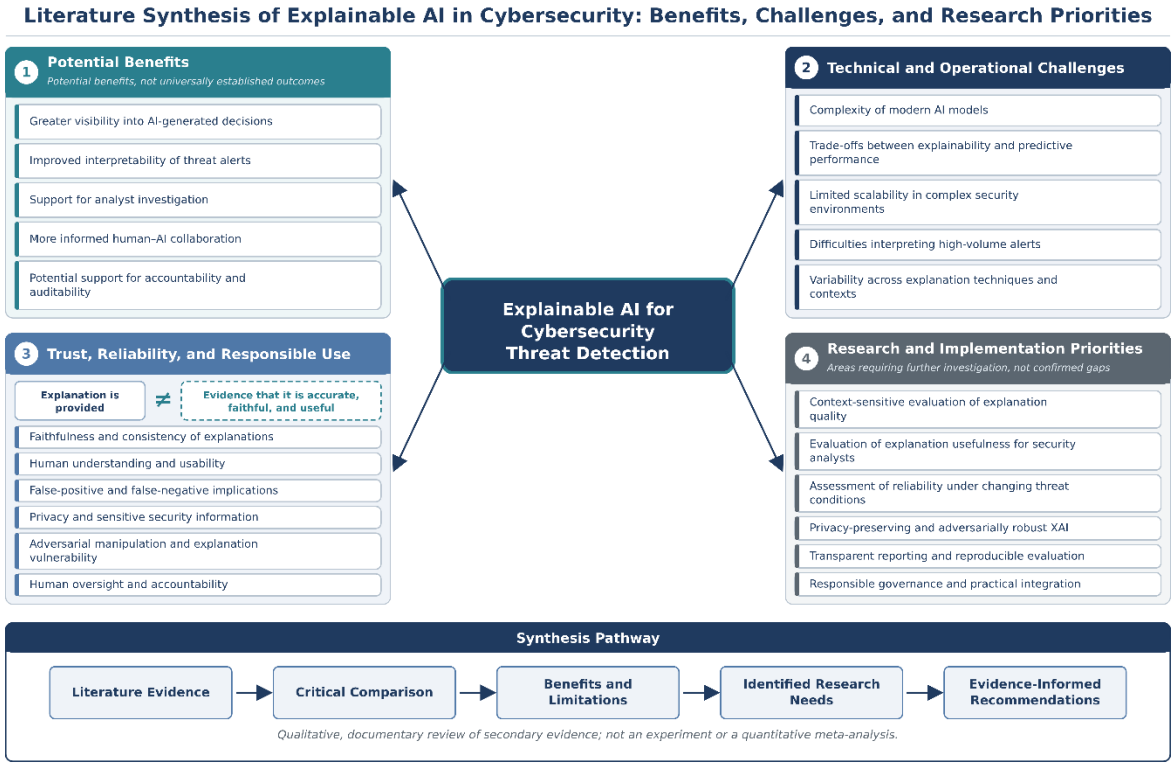
model. XAI should therefore be assessed both for its usefulness to authorized analysts and for the risks arising from disclosure, manipulation, or misinterpretation.

Confidentiality and privacy are also essential. Security information may include sensitive network details, user activity, system configurations, and defensive infrastructure. Some explanation methods may reveal information that should remain confidential. Responsible implementation therefore requires decisions about access control, data minimization, disclosure, and the intended recipients of explanations. Not every user or situation requires the same level of technical detail.

Concept drift and changing threats create additional difficulties. Patterns learned by a model may lose their relevance when network infrastructure, software, user practices, or attack methods change. Under new conditions, explanations based on earlier behaviour may become less useful. Systems and their explanations should therefore be tested continuously to determine whether they remain appropriate for their intended uses. Explainability is not a one-time task completed during model development.

Operational constraints also affect adoption. More detailed explanations may require greater computational resources, increase response times, and create information overload unless the detail contributes to an investigation. A security specialist may need more technical information than a manager responsible for escalating incidents or overseeing organizational risk. Explanation design should therefore account for users' expertise, need for timely action, available resources, and decision-making authority. The quality of an explanation should be assessed in relation to the task it supports, not solely by its technical detail.

Governance and accountability remain unresolved challenges. Organizations need protocols specifying who reviews AI-generated alerts, how disagreements with AI recommendations are handled, and how decisions are documented. The NIST AI Risk Management Framework emphasizes identifying and managing AI risks within an organization and throughout the system lifecycle (Tabassi, 2023). ISO/IEC 42001:2023 also provides a framework for establishing and continually improving AI management systems. These resources can support governance, but their effectiveness depends on implementation, institutionalization, and context-specific review and oversight.



Theoretical Concepts and Perspectives

Three complementary perspectives help explain how XAI may support cybersecurity threat detection: (1) trustworthy AI governance, (2) calibrated human–AI trust, and (3) sociotechnical systems thinking. These perspectives are useful because explainability is not determined solely by the properties of an algorithm. It also depends on the available data, the role of the user, and the organizational context in which AI-supported decisions are made.

The responsible-AI perspective emphasizes reliability, transparency, accountability, safety, and risk management. Tabassi (2023) explains that the NIST AI Risk Management Framework provides guidance for assessing and mitigating risks throughout the AI system lifecycle. Similarly, ISO/IEC 42001:2023 provides a framework for managing AI systems within organizations. These sources provide a foundation for considering explainability within broader governance arrangements. XAI is not a complete solution to cybersecurity threats; it is one component of a wider strategy for responsible AI use. Its value depends on whether explanations appropriately support evaluation, risk management, and accountability.

The human–AI interaction perspective emphasizes users' judgements and experiences of AI recommendations. Cybersecurity analysts must interpret explanations in light of their expertise, the available evidence, and the possible consequences of error. An explanation that is technically correct but too complex, or that does not address the analyst's immediate question, may have little practical value. The opposite risk also exists: an overly simple explanation may conceal important limitations. It is therefore useful to consider whether explanations support informed judgement or encourage uncritical acceptance, a distinction captured by the concept of calibrated trust.

Sociotechnical systems thinking considers cybersecurity outcomes to arise from interactions among technical systems, people, procedures, and operating environments. From this perspective, it is not enough to examine the explanation generated by an XAI system. Researchers must also consider how explanations are used in incident response, how analysts can challenge automated classifications, and who is responsible for the consequences of decisions. These perspectives inform the present study's interpretation of explanation quality, human understanding, operational reliability, and institutional accountability.

Research Gap and Critical Synthesis

The reviewed literature establishes XAI as an active and developing area of cybersecurity research. Charmet et al. (2022), Zhang et al. (2022), and Ferrag et al. (2023) examine explanation methods and cybersecurity applications, while Pawlicki et al. (2024) discuss challenges in explainable intrusion detection. This work shows that the debate has moved beyond whether AI outputs should be explainable to questions about the quality, security, utility, and limitations of explanations. The present study does not claim that XAI in cybersecurity is unexplored. Instead, it examines how established technical discussions relate to human, organizational, and governance considerations in responsible security decision-making.

The first issue concerns the relationship between explanation quality and predictive accuracy. Techniques are available for explaining complex model outputs, but there is no guarantee that an explanation is correct or that the prediction itself is correct. Likewise, a correct prediction does not mean that security professionals understand how the model reached its decision. This distinction supports an integrated approach in which interpretability, accuracy, reliability, and operational usefulness are treated as analytically distinct but related dimensions. Without this distinction, claims about XAI may incorrectly imply that better understanding of a threat necessarily leads to better threat detection.

The second issue concerns explainability and justified trust. Technical explanations may help analysts examine alerts, but their usefulness depends on fidelity, comprehensibility, uncertainty, and context of use. A persuasive explanation may still be inadequate, while a highly technical explanation may be difficult to translate into a security response. Principles of explainability and responsible AI can help address these problems, but they must be considered alongside the analyst's role, organizational processes, and the consequences of security failures. The research question therefore extends beyond whether an AI system can provide explanations to whether those explanations support appropriate human judgement and accountability.

The third issue concerns the relationship between explanation mechanisms and operational environments. The usefulness of XAI may be affected by privacy risks, adversarial manipulation, evolving threat patterns, computational limitations, and the retrospective nature of some explanations. These concerns are interconnected. For example, a more detailed explanation may help an analyst in the short term but could disclose confidential information about defensive processes. Similarly, an explanation method suitable for one threat scenario may need to be reassessed for another. These issues support a holistic view of XAI within cybersecurity and governance.

The study seeks to contribute an integrated conceptual understanding of these issues. Previous reviews have focused on XAI techniques, cybersecurity use cases, or technical challenges. The present study brings together explanation fidelity, calibrated trust, detection reliability, human oversight, operational constraints, and institutional accountability in a qualitative documentary synthesis. It does not claim that these issues have never been studied. Rather, it seeks to provide a structured understanding of how they relate to one another and what those relationships mean for the responsible use of AI in threat detection.

A qualitative, interpretivist, exploratory, and documentary approach is appropriate because the study examines concepts, arguments, limitations, and governance implications rather than experimentally measuring the performance of a particular model. It seeks to clarify when explainability may support the interpretation of cybersecurity decisions, when it may be less useful, and why it must be combined with technical assessment and human oversight. This provides the conceptual foundation for the methodology and analysis and supports a balanced assessment of XAI's potential contribution to trustworthy, transparent, and operationally responsible cybersecurity threat detection.

Research Methodology

Research Philosophy

This study adopts an interpretivist philosophy to explore the meaning and use of explainable artificial intelligence (XAI) in cybersecurity threat detection. Interpretivism emphasizes the context, assumptions, and meanings underlying a phenomenon rather than statistical measurement or experimental testing of causal relationships. This orientation is appropriate because the study examines academic and institutional discussions concerning explainability, transparency, trust, predictive reliability, human oversight, and responsible AI governance. These concepts are not treated as isolated technical features; instead, they are considered within the organizational and operational contexts of cybersecurity. The importance of an explanation may vary according to its technical accuracy, the user's knowledge, and the consequences of the decision it informs.

Research Design and Approach

The study uses a qualitative, interpretivist, exploratory, and documentary research design based on secondary evidence. The qualitative approach supports an in-depth examination of existing knowledge, while the exploratory approach helps identify relationships among concepts discussed across technical, organizational, and governance literature. Documentary research is used to examine published scholarly studies, cybersecurity standards, technical reports, and relevant institutional documents.

The research design focuses on five areas: explainability and interpretability in cybersecurity; trust, transparency, and human oversight; detection accuracy and operational reliability; technical, ethical, and adversarial challenges; and responsible AI governance and implementation. These areas provide an analytical structure rather than predetermined findings. The study includes no primary data collection, interviews, surveys, experiments, model training, benchmark testing, or statistical analysis. It does not claim that particular XAI techniques have a causal effect. Instead, it critically discusses the available evidence and the potential benefits and limitations of explainability in AI-based cybersecurity threat detection.

The study draws on secondary documentary evidence concerning XAI, machine-learning interpretability, cybersecurity threat detection, trustworthy AI, and responsible security governance. The intended source base includes peer-reviewed journal articles, scholarly conference papers, established review studies, cybersecurity standards, technical publications, and authoritative institutional or policy reports. Recent publications are included to reflect developments in XAI, particularly research published between 2020 and 2026. Earlier foundational publications are included where necessary to explain the concepts and methods discussed.

Document selection is guided by relevance to the research questions, scholarly credibility, conceptual contribution, methodological transparency, and discussion of the technical or practical implications of explainable cybersecurity systems. Relevant topics include intrusion detection, malware classification, anomaly detection, explanation fidelity, interpretation of false positives and false negatives, privacy, adversarial risks, human–AI interaction, and AI governance.

The methodology does not claim that a particular academic database was used unless this can be demonstrated in the research records. Similarly, it does not report a search date, search string, number of records retrieved, screening total, or final corpus size without supporting documentation. This avoids presenting undocumented search activities as completed methodological procedures.

Documentary selection and organization

The document-selection process is guided by relevance and credibility. Publications may be included if they address XAI methods, explainability in cybersecurity, trust and transparency in AI-supported decisions, threat-detection reliability, human oversight, or the ethics and governance of XAI security systems. Relevant institutional publications and technical standards are also considered where they provide authoritative guidance on AI risk management, security, or accountability.

Sources are excluded from the analytical discussion when they have no meaningful connection to the research topic, provide insufficient information for critical interpretation, contain bibliographic details that cannot be verified from available sources, or make claims that cannot be adequately evaluated. Duplicate records should be identified and treated consistently. These criteria are intended to keep the evidence relevant and credible while preserving differences in scholarly perspectives and acknowledging source limitations.

These selection criteria describe the methodological principles used to identify evidence; they do not indicate that a formal screening procedure or systematic-review protocol was followed. The analysis remains dependent on the sources that can be substantiated and on the scope and quality of the available documentary material.

Data is retrieved and recorded

A structured documentary-extraction approach is used to organize relevant information. For each source, the analytical record includes bibliographic details, context, research focus, XAI techniques or interpretability approaches discussed, the cybersecurity application, and claims concerning transparency, trust, predictive performance, and operational reliability. Where relevant, the record also captures

technical considerations, sensitivity, privacy concerns, vulnerabilities in adversarial environments, human factors, governance issues, and recommendations for future research.

These categories are not a measurement instrument; they serve as qualitative organizing principles. They can be used to compare documents and identify shared arguments, disagreements, and unanswered questions. For example, a paper focused on feature attribution in intrusion detection may be considered alongside research on analyst oversight or explanation fidelity. Such comparisons help connect technical explanation methods with broader questions about operational usefulness and responsible implementation.

Documentary analysis examines the meaning and implications of published evidence within its original context. It supports the systematic examination and interpretation of documents, including the identification of meaning, the development of understanding, and the generation of insights, as described by Bowen (2009). In this study, documentary analysis supports synthesis; it does not assume that findings from different cybersecurity environments are directly comparable. The study does not claim the use of specialized qualitative software, an independent coding team, or a formal inter-coder reliability procedure.

Qualitative Thematic Analysis

Thematic analysis is used to interpret the documentary evidence, drawing on the approach described by Braun and Clarke (2006). It provides a flexible method for identifying and interpreting patterns of meaning in qualitative material. This approach is appropriate because the study explores conceptual relationships and recurring issues rather than statistical associations across scholarly and institutional documents.

The analytical process follows six phases. First, the documents are read to become familiar with their issues, contexts, and arguments. Second, initial codes are developed by identifying passages and concepts relevant to the research questions, including explanation fidelity, calibrated trust, predictive reliability, human oversight, and security governance. Third, the codes are organized into provisional themes that capture broader patterns across the evidence. Fourth, the preliminary themes are reviewed against the documents for coherence, contradictions, and underdeveloped interpretations. Fifth, the themes are refined and labelled according to their analytical focus. Finally, an interpretive synthesis connects the themes to the study's objectives and the wider academic literature.

The analysis is organized around five thematic areas:

1. Explainability and Interpretability in Cybersecurity: Examines what explanations are and how they may help, or sometimes hinder, the understanding of threat-detection decisions.
2. Trust, Transparency, and Human Oversight: Explores the relationships among explanations, calibrated trust, analyst judgement, and accountability.
3. Detection Accuracy and Operational Reliability: Examines how XAI may help explain the limitations of predictions, including false positives and false negatives.
4. Technical, Ethical, and Adversarial Challenges: Considers explanation fidelity, privacy, changing threats, and potential manipulation.
5. Responsible AI Governance and Implementation: Examines human oversight, organizational accountability, and governance arrangements for explainable security systems.

These categories provide a preliminary analytical framework focused on the research questions. They are not presented as final findings; their scope and boundaries may be refined as the documentary analysis develops. Additional ideas and differing interpretations that emerge from engagement with the texts may be incorporated. The analysis aims to present documentary evidence transparently and coherently.

The standards for analytical rigor and trustworthiness include the following:

- **Methodological rigour** is supported by checking source details, documenting analytical decisions, comparing scholarly arguments, and making explicit connections between interpretations and documentary evidence. Credibility is strengthened by drawing on authoritative publications and peer-reviewed literature while considering contrasting evidence rather than selecting only sources that support a particular argument.
- **Dependability** is addressed through consistent document organization, clear definitions of relevant concepts, and review of provisional themes. Confirmability is supported by distinguishing claims made by sources from the researcher's interpretations and reflecting on how prior assumptions about AI explainability may influence analytical judgements. Transferability is addressed by describing the conceptual domains and cybersecurity contexts to which the interpretations may apply, rather than claiming that they apply universally.

- Scholarly research, standards, and institutional reports can provide complementary insights into technical performance, governance, and organizational responsibilities. However, their different purposes and evidential strengths must be distinguished. Agreement among sources does not eliminate interpretive bias, and institutional guidance does not constitute empirical validation. These measures support a more transparent analytical process but do not represent formal external validation, participant validation, or independent auditing.

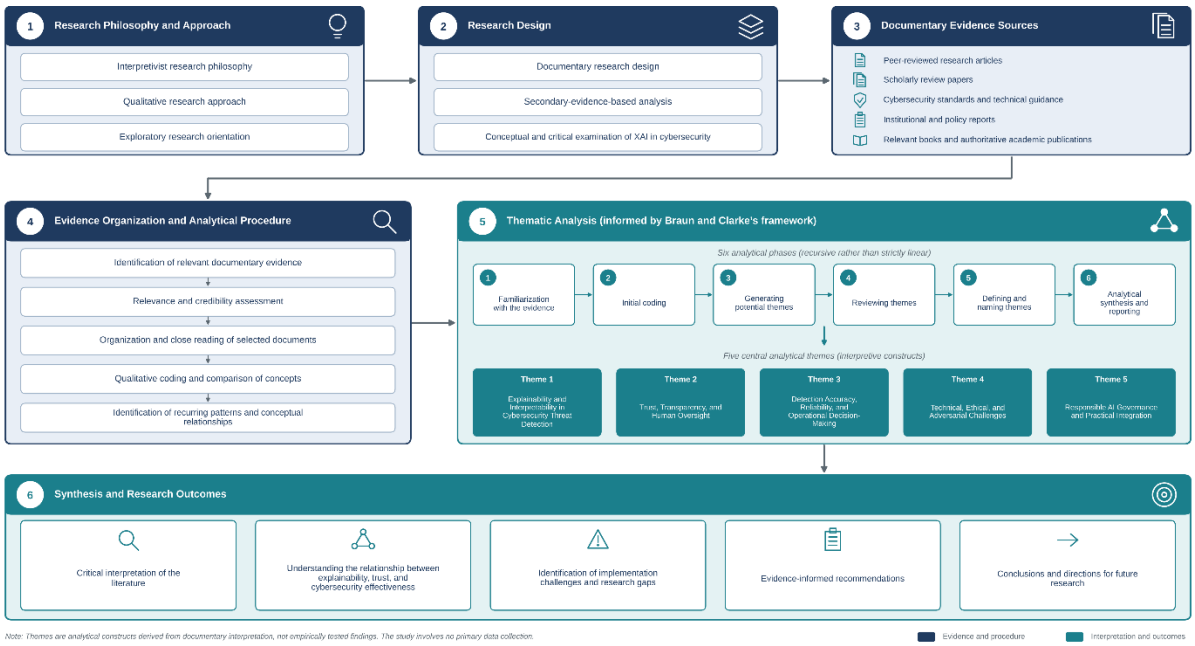
Ethical Considerations and Research Limitations

This study analyses existing documentary sources and does not collect data from human participants. Its ethical considerations therefore concern accurate attribution, responsible interpretation, respect for intellectual property, and fair representation of published arguments and findings. Sources should be cited correctly, and conclusions should not be overstated or presented as facts when they are not supported by the evidence. No formal ethics approval is claimed for this study.

The study has several limitations. Published sources may vary in availability, quality, and relevance. Differences in evaluation conditions and operational environments can make cybersecurity applications difficult to compare directly. Findings may also be affected by rapidly changing attack techniques and AI technologies. Documentary research may not capture every relevant publication or the experiences of security practitioners who are not represented in the literature. In addition, the interpretation of documents necessarily depends on the researcher's analysis and synthesis of the material.

The study does not independently verify the predictive accuracy of an AI model, the causal effectiveness of an XAI technique, or the claim that explainability improves cybersecurity outcomes. Its findings are based on the available documentary evidence and the scope of the analysis. Within these limits, the methodology provides a structured basis for critically examining the relationships among explainability, trust, transparency, detection reliability, human oversight, and responsible governance in AI-based cybersecurity threat detection.

Figure 1. Research Methodological Framework for Analyzing Explainable Artificial Intelligence in Cybersecurity Threat Detection



Findings and Discussion

This section critically analyses the main themes identified through the conceptual framework and documentary material in this qualitative, interpretivist, and exploratory study. It examines explainable artificial intelligence (XAI) and its application to cybersecurity threat detection, focusing on the relationships among explainability, interpretation of AI-generated security alerts, trust, predictive accuracy, operational reliability, and responsible governance. The study does not assume that XAI always improves cybersecurity outcomes. Instead, it assesses the potential value of explanations and the technical and organizational conditions required for their effective use. The discussion is organized around five themes: (1) explainability and interpretability; (2) trust, transparency, and human oversight; (3) detection accuracy and operational reliability; (4) technical, ethical, and adversarial challenges; and (5) responsible AI governance. These themes provide an analytical lens for examining XAI's role in security decision-making and the limitations that must be addressed before it can be relied upon in operational environments.

Theme 1: Explainability and Interpretability in AI-Enabled Cybersecurity Threat Detection

A central point in the literature is that explainability should be treated as a means of examining AI-generated security decisions, not as evidence that those decisions are correct. Machine-learning systems can identify patterns in network traffic, system events, files, user behaviour, and other security-related data. Deep-learning models may be useful when relationships among these features are complex, but their decision processes can be difficult to interpret. This creates a practical challenge: a model may generate a potentially useful alert without providing enough information for a security professional to understand it. Explainable intrusion-detection systems are therefore an important research area, although transparency must be considered alongside model performance and other security requirements (A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems, 2025; Pawlicki et al., 2024).

The distinction between interpretability and explainability helps clarify this issue. Interpretability generally refers to how readily a model's structure or behaviour can be understood, whereas explainability concerns the explanations provided for a model's outputs. These concepts are related but are not synonymous in every research context. For an inherently interpretable model, such as a decision tree with understandable rules, the contribution of features to a classification may be examined directly. A post-hoc method, by contrast, attempts to explain the behaviour of an already trained model without necessarily changing its structure. Such methods can make complex models more accessible, but they raise an important question: does the explanation faithfully represent the model's decision process?

Feature-attribution methods such as SHAP and local surrogate methods such as LIME illustrate both the potential and limitations of post-hoc explanations. They can provide explanations for individual predictions or estimate model behaviour for a particular instance (Lundberg & Lee, 2017; Ribeiro et al., 2016). In cybersecurity, this may help analysts investigate an intrusion alert by examining which traffic characteristics, event patterns, or other model inputs contributed to the classification. A local explanation may be useful when investigating one suspicious event, while a global explanation may help a technical team examine broader patterns in model behaviour. These approaches are complementary: an explanation of one alert does not describe the entire model, and a general description of model behaviour may not explain an unusual individual event.

These methods may be relevant to several security applications, but their usefulness depends on the application and the information supplied to the model. An explanation may help identify the traffic characteristics associated with an intrusion alert. In malware analysis, it may highlight file properties associated with a malicious classification. In phishing detection, it may identify message characteristics or other inputs that contributed to a prediction. In anomaly detection, it may identify the features that caused an event to depart from an expected pattern. These are potential applications discussed in the literature, not empirical results produced by this study. An explanation is useful only when the features it identifies are relevant, accurately represented, and understandable in the operational context. A feature associated with malicious traffic in one development environment, for example, may also be common in legitimate traffic in another environment. The explanation may faithfully reveal the model's reliance on that feature while also exposing a weakness in the model.

This discussion is consistent with the principles of explainable AI described by Phillips et al. (2021): explanations should be understandable to their intended users, accurately reflect the system's processes, and acknowledge the limits of the system's knowledge. The relevant question is not simply whether a human reader finds an explanation persuasive. The explanation must also be faithful to the underlying model and suitable for its intended user.

A detailed explanation may be difficult for a frontline analyst to understand, while a simplified explanation may omit information needed for specialist investigation. Explanation quality is therefore context-dependent: it depends on the model, the cybersecurity task, the user's knowledge, and the decision the explanation is intended to support.

An explanation of a model output is not necessarily an explanation of a correct security decision. A model may classify an event as malicious even when the activity is legitimate. A plausible explanation for a false positive does not make the alert correct. Classification accuracy depends on data quality, model quality, the relevance of learned patterns, and the context in which the model is deployed. Explainability can help analysts investigate these issues, but it does not resolve them automatically. Research on explainable intrusion-detection systems continues to identify challenges involving explanation validation, scalability, and the relationship between interpretability and detection performance (Pawlicki et al., 2024; A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems, 2025).

The interpretation of this theme is that explainability should not be treated as an inherent virtue of any AI system that produces an explanation. It should instead be evaluated as one component of cybersecurity decision-making. Explanations may help analysts examine model behaviour, question unusual events, and identify cases requiring further investigation. Their effect on operational decisions may be positive or negative, depending on fidelity, relevance, comprehensibility, and the availability of independent evidence. This distinction helps explain how XAI may support justified trust while also reducing the risk of over-reliance on AI outputs.

Theme 2: Trust, Transparency, and Human Oversight

Providing more explanations does not necessarily create greater or more appropriate trust. In cybersecurity, analysts must interpret alerts in time-sensitive environments, decide which incidents require escalation, and assess whether automated recommendations justify a response. Explanations can make system outputs more accessible for questioning and investigation. However, users can reasonably rely on an AI system only when its explanations accurately represent its behaviour and communicate its limitations. It is therefore important to distinguish users' perceptions of trust from the trustworthiness of the system and the arrangements surrounding its use.

This distinction is consistent with the NIST AI Risk Management Framework, which treats validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy, and fairness as related characteristics of trustworthy AI (Tabassi, 2023). These characteristics are not interchangeable. They should be considered in relation to the system's context and intended use.

In cybersecurity, an explanation may clarify a model's decision without resolving questions about trust, manipulation, or suitability for the operating environment. A detector does not become trustworthy merely because it provides detailed explanations; its reliability and security must also be assessed.

Human judgement is central to this relationship. Explanations can provide analysts with information to consider alongside other evidence, compare predictions with known security conditions, and question results that appear inconsistent. This is particularly important when a false positive disrupts legitimate activity or a false negative allows malicious activity to continue. Human review creates an opportunity to challenge model outputs rather than accept them as final decisions. However, oversight is meaningful only when reviewers have the expertise, time, access to information, and authority needed to assess the output. A nominal human-approval step may simply confirm automated recommendations without meaningfully evaluating them.

The concept of calibrated trust helps explain this issue. Users should understand an AI system's capabilities and limitations so that they can rely on it appropriately. Analysts may place too much confidence in a system and accept its outputs uncritically, or place too little confidence in it and disregard useful warnings. The objective is not to maximize confidence, but to align reliance with demonstrated capability and the consequences of error. Phillips et al. (2021) emphasize users' understanding and the limits of a system's knowledge. Operationally, explanations should communicate relevant uncertainty and limitations rather than present predictions as unquestionable conclusions.

Transparency cannot be achieved simply by adding an explanation panel or feature-attribution display. It also involves making the system's operation and limitations visible, assessing whether its explanations are consistent in similar situations, ensuring that they are understandable to the intended audience, and providing information relevant to the decision. An explanation may have limited value if it uses unfamiliar terminology or fails to connect model features with security concepts. A concise explanation may be easy to understand but misleading if it omits important qualifications. Explanation design must therefore balance technical fidelity with the needs of system users.

Organizational arrangements also influence trust. Analysts need clear procedures for evaluating explanations, escalating cases when explanations appear inconsistent, and recording explanations that inform consequential decisions. Managers need to understand how AI-assisted decisions fit within ongoing security responsibilities, while technical teams need sufficient information to investigate model behaviour. These groups may require different levels of detail. A security analyst may need feature-level information, whereas an organizational decision-maker may need a concise summary of the reasons, limitations, and implications. One explanation format may therefore be insufficient for all stakeholders.

From a sociotechnical perspective, trust emerges from the interaction of the AI model, explanation method, user expertise, organizational processes, and consequences of relying on AI. It cannot be established through technical design alone. The presence of human oversight does not guarantee effective review: responsibility may be unclear, or organizational pressure may encourage rapid acceptance of automated recommendations. A defensible approach is to rely on XAI only when explanations are sufficiently faithful, comprehensible, appropriately limited, and embedded in procedures that allow meaningful human challenge. Conversely, persuasive explanations may create false confidence if they are mistaken for proof of correctness or if human review is merely nominal.

Theme 3: Detection Accuracy, Reliability, and Operational Decision-Making

Explanation quality and detection accuracy concern different aspects of an AI-based security system. Accuracy concerns the correctness of predictions under specified evaluation conditions, whereas explainability concerns whether model behaviour or individual predictions can be understood. Correct predictions are not the only measure of reliability; a system must also perform consistently under its intended operating conditions. These dimensions should not be treated as equivalent, and improvement in one does not necessarily improve the others. XAI may help professionals investigate the reasoning behind a prediction without necessarily changing detection effectiveness.

The distinction between false positives and false negatives is important. A false positive occurs when legitimate activity is incorrectly classified as malicious, potentially increasing alert volume and the time analysts spend investigating. A false negative occurs when

malicious activity is classified as legitimate or is not detected, allowing a threat to continue. In either case, explanations may provide information that supports investigation. Analysts can examine the properties of an alert, consider which features influenced the classification, and compare the output with available contextual evidence. In a false-negative investigation, understanding model behaviour may help technical teams identify why expected features did not influence a classification. However, identifying a possible source of error is not the same as demonstrating that the error rate has decreased.

Post-hoc explanations do not necessarily change the prediction that has already been made. A feature-attribution method may show that a model relies heavily on a particular feature, but it does not automatically correct an inappropriate reliance on that feature. Similarly, a local explanation may make a classification easier to understand without establishing whether it is correct. Demonstrating improved detection accuracy requires appropriate empirical evaluation. Evidence about review processes and decisions is also needed to establish whether explanation-assisted review improves human decision-making. The present documentary study can synthesize these requirements and existing arguments, but it cannot determine whether XAI has improved the performance of a particular detector.

Reviews of XAI-supported intrusion detection illustrate the importance of these distinctions. Research in this area considers not only interpretability but also trade-offs involving detection performance, explanation methods, scalability, and deployment. For example, the 2025 review titled “A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems” examines XAI-based intrusion-detection research and identifies continuing issues involving standardization and scalability. Pawlicki et al. (2024) also discuss challenges associated with explainable intrusion detection.

These discussions show why the use of an explanation method should not be treated as proof that a security system is more accurate or effective.

A related issue concerns the choice between inherently interpretable models and complex models supported by post-hoc explanations. Decision rules in simpler models may be relatively easy to examine, while complex architectures may contain relationships that are difficult to express through simple rules. Neither type of model is guaranteed to produce the best outcome in every cybersecurity setting. An interpretable model may still be inaccurate, poorly calibrated, or unsuitable when attack patterns change. A complex model may perform well in a particular context but provide insufficiently useful explanations. The relevant question is therefore how explanation quality, model performance, operational requirements, and the consequences of error should be evaluated separately.

Uncertainty and instability in explanations add to this challenge. Analysts may find it difficult to determine whether an apparent explanation pattern is meaningful or merely reflects the sensitivity of the explanation method. A system may also produce a confident-looking explanation when used outside the environment for which it was designed. NIST's explainability principles emphasize the importance of operating within a system's design limits and acknowledging the boundaries of its knowledge (Phillips et al., 2021). This is relevant in cybersecurity when traffic patterns change, new attack methods emerge, or a detector is deployed in an environment different from its development setting.

Explanations alone cannot address every factor affecting operational reliability. Model predictions may be affected by poor-quality input data, incomplete network visibility, changes in legitimate user behaviour, and evolving attack strategies. A model may provide an understandable explanation even when its input data are incomplete or inaccurate. The explanation may faithfully describe the model's response to those inputs without revealing the underlying data problem. Explainability must therefore be considered alongside data-quality controls, input monitoring, model validation, security controls, and procedures for responding to uncertain predictions.

Operational usefulness also depends on timing and resource requirements. An explanation designed for retrospective investigation may be too slow, detailed, or computationally expensive for a high-volume environment where alerts must be triaged quickly. A concise explanation may help during an incident but may not provide enough detail for a full investigation. No single approach is likely to be optimal for intrusion detection, malware classification, phishing detection, and every other security application. The value of a method depends on what it explains, how quickly it produces an explanation, the resources it requires, and the decision it is intended to support.

The key implication is that XAI should not be evaluated using a single criterion. Detection performance, explanation fidelity, explanation stability, analyst comprehensibility, computational overhead, and operational usefulness address different questions. A system may perform well on one dimension and poorly on another. Documentary evidence therefore supports treating XAI as a means of investigating and interpreting security decisions, not assuming that it increases predictive accuracy or overall cybersecurity effectiveness. Such improvements must be tested separately through task-specific evaluations.

Theme 4: Technical, Ethical, and Adversarial Challenges

The use of XAI in cybersecurity involves more than making complex models understandable. Explanations may be inaccurate, inconsistent, vulnerable to manipulation, or unsuitable for their intended audience. They may also introduce security vulnerabilities or operational burdens. These concerns are interconnected: an explanation may be clear but inaccurate, accurate but too narrowly focused, or informative in one environment but unsuitable in a changed threat environment. The central question is not simply whether an explanation can be generated, but whether it is reliable and safe under the conditions in which it will be used.

Explanation fidelity is a major concern. Post-hoc techniques often approximate or summarize model behaviour, and their outputs should not automatically be treated as complete accounts of a model's internal processing. If an explanation does not faithfully represent the model, analysts may form an incorrect understanding of why an alert was generated. This can undermine model debugging, incident investigation, and accountability, particularly when explanations are used to support decisions with operational or legal consequences. Phillips et al. (2021) distinguish between explanations that are understandable and explanations that accurately represent what is happening in a system. Both qualities matter, but they are distinct. Explanations should therefore be assessed for accessibility and fidelity.

Adversarial manipulation is another security concern. Explanation methods may be used during normal operations to help users understand or investigate a model, but attackers may attempt to manipulate the model, the explanation, or the relationship between them. Baniecki and Biecek (2024) discuss attacks and defences involving explainable AI, including adversarial examples, data poisoning, and backdoor attacks. Their discussion highlights that explanations should not automatically be assumed to be secure simply because their purpose is to make model behaviour more understandable.

This creates a double-edged issue in cybersecurity: information that helps defenders understand a detection system may also provide information that an adversary could misuse.

The implications depend on the information disclosed and the capabilities of the potential attacker. Details about the input features that influence a detector may help analysts identify inappropriate model dependencies. However, the same information could help an adversary modify malicious activity to avoid detection. This does not mean that explanations should be abandoned or that every explanation creates an exploitable vulnerability. Rather, disclosure should be assessed in relation to the threat model, intended audience, and sensitivity of the information. Detailed explanations may be appropriate for internal investigations but unsuitable for unrestricted disclosure.

Privacy and confidentiality present further challenges. Cybersecurity systems may process network records, user activity, device information, incident logs, and other sensitive operational data. Explanations may reveal information about individuals, organizational systems, or defensive capabilities. Even when raw records are not disclosed, an explanation may become sensitive when combined with other available information. Removing obvious identifiers may therefore be insufficient to protect privacy. Organizations should consider what an explanation reveals, who can access it, how long it is retained, and whether disclosure is necessary for the decision it supports. Excessive restrictions, however, may prevent analysts from obtaining information needed to investigate threats. The challenge is to balance confidentiality with meaningful, authorized access.

Cyber threats and operating environments also continue to change. Attacker behaviour, software configurations, user activity, and network settings may differ over time, while cybersecurity models are developed using data from particular environments. A feature that was informative in the past may become less relevant, and new features or behaviours may not be adequately represented in the model. Explanations may therefore remain understandable while becoming less useful for current threats. NIST's adversarial machine-learning taxonomy emphasizes the need to consider the stages at which attacks can affect machine-learning systems, as well as attackers' objectives and capabilities (Vassilev et al., 2025).

Human interpretation and operational usability must also be addressed. An explanation may be technically correct but fail to support a decision if it relies on specialist jargon, lists too many features, or does not distinguish relevant evidence from incidental correlations. Under time pressure, analysts may be overwhelmed by lengthy explanations and struggle to compare them with other alerts. Different roles also require different levels of detail. A technical specialist investigating a complex intrusion may need feature-level information, while an incident manager may need a concise summary of uncertainty, alert significance, and possible next steps. Explanation systems should account for these differences to avoid both information overload and excessive simplification.

These constraints are closely related to accountability. If an AI output is incomplete and a human reviewer accepts it, the organization cannot assume that meaningful human oversight has occurred. Organizations should define responsibilities for system design, explanation evaluation, incident review, approval of consequential actions, and remediation of identified weaknesses. This is particularly important when AI-supported decisions may block traffic, restrict network access, escalate an incident, or trigger another significant defensive action. Explanations should inform organizational decisions, not shift responsibility from people or institutions to the model.

Taken together, these issues suggest that explanation mechanisms should be considered part of the cybersecurity system's potential attack surface. Fidelity, privacy, adversarial robustness, usability, and governance are interconnected and cannot be addressed in isolation. Improving one dimension may create trade-offs elsewhere: greater detail may help an investigation but reveal sensitive information, while greater simplicity may improve accessibility but omit important qualifications. Responsible design therefore requires context-sensitive evaluation, explicit threat modelling, controlled disclosure, and ongoing monitoring. The value of XAI depends not only on making AI behaviour visible, but also on avoiding confusion, vulnerabilities, and unjustified confidence.

Theme 5: Responsible AI Governance and the Gap Between XAI Deployment and Practice

The final theme concerns the organizational arrangements required to implement explainability responsibly in cybersecurity. The preceding themes indicate that XAI may help interpret AI-generated alerts, but its usefulness depends on explanation quality, system trustworthiness, human expertise, information security, and effective alert-management processes. Governance coordinates these requirements. Explainability should therefore be incorporated into the design, deployment, monitoring, and review of AI-powered security systems rather than added only after a model has been developed.

The NIST AI Risk Management Framework is one relevant resource. It organizes AI risk management around the functions of Govern, Map, Measure, and Manage, and emphasizes consideration of risks throughout the AI system lifecycle. Its relevance to explainable cybersecurity lies in connecting technical performance with transparency, accountability, safety, security, privacy, and appropriate use.

The ability of a system to generate an explanation does not, by itself, demonstrate that the explanation is adequate. Organizations should assess whether the explanation is suitable for the task, whether its limitations are understood, and whether the information it provides supports the decision for which the system is used.

A lifecycle approach begins by defining the purpose of an explanation. During system design, developers and security practitioners should determine whether explanations will support alert triage, model debugging, incident investigation, auditability, or another specific task. Each purpose may require different information. An explanation useful for model development may not meet the needs of a frontline analyst making a time-sensitive decision. Likewise, information sufficient for routine alert triage may not be adequate for a post-incident investigation. Defining the intended purpose makes it possible to evaluate transparency against specific requirements rather than relying on vague claims that a system is or is not transparent.

Governance should also establish procedures for evaluating explanations. Relevant criteria may include fidelity to model behaviour, comprehensibility for intended users, consistency under similar conditions, sensitivity to input changes, and the risk of disclosing protected information. These criteria should be considered alongside detection performance and operational requirements. A persuasive explanation should not be accepted without evidence of fidelity. Similarly, a technically accurate explanation may need to be redesigned if analysts cannot understand or use it. Governance should clarify what constitutes an acceptable explanation and ensure that evaluation is proportionate to the risk of error. Human oversight should be supported by clear responsibilities and escalation procedures. Security practitioners should know when an explanation is sufficient to guide an investigation, when additional evidence is needed, and when a decision should be referred to a more experienced reviewer. Organizations should also define how to handle conflicts between AI recommendations and other evidence. Such procedures can reduce the risk that explanations displace professional judgement and can support documentation of the reasons for decisions, the evidence considered, and the treatment of uncertainty.

Documentation is essential to accountability. Merely having access to an explanation when an alert is generated is not sufficient. Useful records may include intended uses, known limitations, explanation methods, evaluation criteria, known failure conditions, access controls, and procedures for reviewing consequential decisions. Documentation can help distinguish the model's output from the human reviewer's interpretation and subsequent action. It may also support later investigation into whether a model relied on inappropriate features or whether operational procedures contributed to an error. Documentation does not prove that a decision was correct; it makes the decision process more transparent and open to review.

Ongoing monitoring is also essential. An explanation method that is useful when a system is first deployed may become less suitable as configurations, user behaviour, or attack techniques change. Organizations should therefore monitor the continued suitability of explanations as well as model behaviour. This monitoring should be integrated with existing security-monitoring and model-risk processes rather than treated as a separate XAI activity. If significant limitations are identified, an organization may need to revise the explanation interface, reassess the model's suitability, adjust operating parameters, or restrict automated use until the concerns are addressed. These are governance priorities derived from the issues discussed in the literature, not results of testing a specific implementation.

Governance decisions must also reflect security and privacy requirements. Authorized analysts may need detailed explanations, but unrestricted disclosure of defensive mechanisms or system characteristics may be inappropriate. Disclosure policies and access controls should therefore reflect the sensitivity of the information and the organization's operational context. These controls should allow authorized users to investigate incidents without exposing information to inappropriate audiences. Decisions about retaining explanations should balance their investigative and accountability value against confidentiality requirements and the risks of storing sensitive security information. Formal standards can inform this process. NIST has published a voluntary framework for identifying and managing AI risks (NIST, 2023), as well as a taxonomy for analysing adversarial machine-learning threats (Vassilev et al., 2025).

These documents can inform organizational procedures, risk assessment, and technical evaluation. However, citing a framework or standard does not guarantee that an explainable intrusion-detection system will achieve a particular level of accuracy, robustness, or operational effectiveness. Governance guidance should inform local assessment rather than replace it.

These issues can be understood through a sociotechnical perspective. The reliability of explainable cybersecurity depends on modelling techniques, explanation methods, analysts, organizational policies, technical infrastructure, and operating conditions. A technically sophisticated explanation may have little value if an analyst lacks the knowledge or authority to act on it. Conversely, even an experienced specialist may be misled by a persuasive but unreliable explanation. Organizational policies cannot compensate fully for a model that is unsuitable for its operational context. Responsible implementation therefore requires coordination among technical development, security operations, risk management, and organizational leadership.

The interpretation of this theme is that explainability should be understood as an ongoing organizational capability, not merely a technical feature. Responsible use requires explanations to have a clear purpose, to be assessed for fidelity and usability, and to be reviewed through meaningful human oversight. Disclosure must be controlled, monitoring must continue, and decision-making responsibilities must be clear. These arrangements do not guarantee that every threat will be detected or that a cybersecurity system will be error-free. They do, however, provide a more structured basis for assessing whether AI-generated explanations are suitable for informing security decisions and for addressing the limitations of AI systems.

Integrative Discussion

Overall, the five themes present XAI as one component of cybersecurity threat detection. Explanation quality, calibrated trust, predictive reliability, and organizational governance provide related but distinct perspectives for understanding its role. Explainability concerns the information available about model behaviour; trust concerns whether reliance on the system is justified; accuracy concerns the correctness of predictions; and governance concerns how the system is evaluated, used, and held accountable. Improvement in one dimension may support the others, but does not guarantee improvement across all dimensions.

This interpretation reflects the multiple dimensions of AI trustworthiness, in which explainability, reliability, security, transparency, privacy, and accountability are interconnected but distinct (Tabassi, 2023). It also reflects challenges identified in research on explainable intrusion detection, including explanation quality, scalability, performance considerations, and adversarial vulnerabilities (Pawlicki et al., 2024; Baniecki & Biecek, 2024).

The central argument is not that XAI necessarily makes cybersecurity systems more accurate or trustworthy. Rather, XAI may help analysts investigate and question AI-generated decisions when appropriate technical and organizational controls are in place.

The practical implication is that explanations should be tailored to the investigation and operational context. Their value should be assessed through fidelity, comprehensibility, stability, security, and usefulness, while detection performance is evaluated independently. Human oversight and governance arrangements should address situations in which explanations may mislead users, disclose sensitive information, or be exploited by adversaries. Future research should examine these dimensions together while distinguishing explanation quality from predictive effectiveness and perceived trust from justified reliance.

The qualitative, interpretivist, and documentary approach used in this study is suitable for critically synthesizing these conceptual relationships and the limitations identified in previous research. The study does not validate a specific XAI method, demonstrate improved detection results in a real-world cybersecurity environment, or establish causal effects. Its contribution lies in integrating the conditions under which explainability may support more transparent, accountable, and informed cybersecurity decisions, while showing why explainability alone is insufficient to ensure trustworthy security outcomes.

Table 1: Critical Synthesis of Findings on Explainable Artificial Intelligence in Cybersecurity Threat Detection

Analytical theme	Key findings from documentary evidence	Critical interpretation and discussion	Implications for cybersecurity practice	Challenges and limitations
1. Explainability and interpretability	XAI helps interpret AI-generated alerts. Intrinsically interpretable models differ from post-hoc methods such as SHAP and LIME (Lundberg & Lee, 2017; Ribeiro et al., 2016). Local explanations address individual alerts; global explanations summarize model behaviour.	Explanation quality depends on the model, task, user, and decision. A plausible explanation is not necessarily faithful or useful (Phillips et al., 2021). Explainability should be evaluated rather than assumed.	Match explanation type and detail to user needs. Treat explanations as investigative aids and corroborate them with independent evidence.	Post-hoc fidelity may be uncertain. Detailed explanations can overwhelm analysts, whereas simplified ones may omit essential information. Validation and interpretability–performance trade-offs remain concerns (Pawlicki et al., 2024; A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems, 2025).

2. Trust, transparency, and human oversight	Explanations may help analysts examine AI decisions, but more explanation does not necessarily produce appropriate trust. Trustworthiness also involves reliability, security, accountability, transparency, and privacy (Tabassi, 2023).	Trust is sociotechnical, depending on explanation quality, user expertise, organizational processes, and consequences. Oversight is meaningful only when reviewers have sufficient knowledge, time, access, and authority.	Communicate uncertainty and system limitations. Define escalation procedures and accountability. Tailor explanations to analysts and decision-makers.	Over-reliance, under-reliance, misleading explanations, time pressure, and unclear responsibility can weaken oversight. No single explanation format suits every stakeholder.
3. Detection accuracy, reliability, and operational decisions	Explanation quality and predictive accuracy are distinct. XAI may support investigation of false positives and false negatives, but explanations do not themselves correct predictions. Standardization, scalability, and performance trade-offs remain concerns (Pawlicki et al., 2024; A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems, 2025).	Documentary evidence supports examining XAI as an aid to interpretation, not assuming that it improves accuracy. Effectiveness requires task-specific evaluation of both detection and explanations.	Evaluate predictive performance and explanation quality separately. Combine explanations with data-quality checks, model validation, and monitoring. Use concise explanations for triage and detailed ones for investigation.	Explanation instability, changing traffic, evolving attacks, computational costs, and method-specific limitations may undermine reliability. An available explanation does not prove that a detection system performs well.
4. Technical, ethical, and adversarial challenges	Explanation fidelity, privacy, usability, and adversarial robustness are important concerns. Relevant threats include adversarial examples, data poisoning, and backdoor attacks (Baniecki & Biecek, 2024). Evolving threats may also reduce the relevance of existing approaches (Vassilev et al., 2025).	Explanation mechanisms form part of the system's potential attack surface. More detail may support investigation but expose sensitive information; simplification may omit important qualifications. These are literature-based risks, not incidents established by this study.	Apply role-based access controls, assess information disclosure, protect explanation records, and avoid relying solely on explanations for consequential security actions.	Misleading explanations, privacy–utility trade-offs, scalability constraints, changing threat patterns, information overload, and diffused accountability remain challenges.
5. Responsible AI governance and practical integration	XAI should be considered throughout design, deployment, monitoring, and review. The NIST AI Risk Management Framework supports lifecycle-based AI risk management (Tabassi, 2023). Evaluation may consider fidelity, comprehensibility, consistency, and disclosure risks.	Explainability is an organizational as well as technical capability. Standards inform risk management but do not establish the effectiveness of a specific implementation. Priorities must be assessed in context.	Define explanation purposes and evaluation criteria. Document intended use and limitations, establish escalation procedures, monitor performance, and coordinate security teams with developers and decision-makers.	Frameworks cannot guarantee accuracy or robustness. Disclosure must be balanced against confidentiality, and changing threats or configurations may reduce explanation suitability.

Note. XAI = explainable artificial intelligence; SHAP = SHapley Additive exPlanations; LIME = Local Interpretable Model-Agnostic Explanations. The table synthesizes documentary evidence and critical interpretation; it does not report original experimental results.

Conclusion

This study examines the potential role of explainable artificial intelligence (XAI) in cybersecurity threat detection, with particular attention to transparency, interpretability, justified trust, detection reliability, and responsible decision-making. It uses a qualitative, interpretivist, exploratory, and documentary approach based on secondary evidence to synthesize scholarly discussions, technical perspectives, and institutional guidance. The analysis considers how explainability relates to trustworthy AI in security systems. The central issue is not simply whether AI decisions can be explained, but whether explanations can support sound human decisions without encouraging over-reliance on automated threat classifications.

The thematic synthesis suggests that explainability can help analysts investigate why AI systems generate security alerts. Explanation methods may help analysts examine suspicious behaviour, question potentially misleading results, and identify features or aspects of model behaviour that influenced a classification. The usefulness of these explanations depends on their fidelity, comprehensibility, consistency, and relevance to the cybersecurity task. A convincing explanation is not necessarily a faithful representation of the model's decision process, nor does it prove that a classification is correct. Explainability should therefore be assessed from the perspective of intended users and the decisions they need to make.

The main argument is that explainability, trustworthiness, and effectiveness are distinct but related aspects of AI in cybersecurity. Explainability concerns how well the behaviour and outputs of a model can be understood. Trustworthiness encompasses broader requirements, including reliability, security, accountability, privacy, and responsible human oversight. Detection effectiveness concerns a system's ability to identify relevant threats in its operating environment. These dimensions may support one another, but they are not interchangeable, and improvement in one does not guarantee improvement in the others. Explanation-assisted investigation should not be confused with improved predictive accuracy. Any claim of improved detection must be supported by empirical evaluation, while claims about the value of explanations for human decision-making require evidence of their use and impact.

The study also identifies technical and organizational challenges that may constrain the responsible use of XAI. These include explanation fidelity, adversarial manipulation, confidentiality, model complexity, computational demands, and differences in users' interpretive abilities. These challenges demonstrate the need for security and privacy safeguards. Detailed explanations may assist specialist investigations but may also reveal sensitive information about defensive systems. Conversely, simplified explanations may misrepresent model behaviour or omit important limitations. Transparency therefore requires a balance among accessibility, technical accuracy, contextual relevance, and information protection.

The study contributes to cybersecurity research by bringing together four dimensions in a qualitative interpretation: technical interpretability, human–AI interaction, detection reliability, and responsible AI governance. Moving beyond the view of explainability as a feature of an individual machine-learning model, this approach places it within the broader sociotechnical context of security decision-making. For cybersecurity analysts and developers, this highlights the need to select explanation methods appropriate to specific tasks and to evaluate explanation quality separately from predictive performance. Security architects and organizational decision-makers should ensure that AI-supported processes are subject to meaningful human review, that responsibilities are documented, that monitoring is ongoing, and that procedures exist for ambiguous or potentially incorrect outputs. For policymakers and governance practitioners, the findings underline the difference between commitments to transparency and effective implementation.

This study has several limitations. It is based on documentary evidence and therefore depends on the availability, quality, and relevance of existing publications and institutional materials. Its findings cannot be generalized automatically to all cybersecurity applications because research contexts, explanation methods, evaluation measures, and threat environments differ. The presentation of evidence and the researcher's interpretive process may also influence the thematic synthesis. Cyber threats and AI technologies continue to evolve, so explanations and governance approaches that are useful in one context may need to be reassessed in another. The study does not independently evaluate AI models, assess trust through primary data, test original cybersecurity datasets, or establish a causal relationship between explainability and detection accuracy. It therefore provides an interpretive synthesis rather than experimental evidence of XAI effectiveness.

Future research should examine explanation fidelity and stability across different threat scenarios, the resistance of explanation mechanisms to adversarial manipulation, and privacy-preserving methods that still support meaningful investigation. Empirical studies could investigate how analysts interpret different explanation formats, how those formats affect calibrated trust and decision-making, and whether explanation-assisted investigations improve operational outcomes under clearly specified conditions. Further work is also needed on integrating explanation-quality assessment into broader governance and monitoring frameworks without imposing excessive computational or operational costs. Such research would help distinguish potential benefits from benefits that have been demonstrated empirically.

Overall, XAI may improve the transparency and auditability of AI-assisted decision-making in cybersecurity, but it does not guarantee accurate, trustworthy, or effective security decisions. Responsible use requires relevant explanations, reliable detection, human judgement, appropriate security and privacy protections, and effective governance.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

2. Arreche, O., Guntur, T., & Abdallah, M. (2024). XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. Applied Sciences, 14(10), Article 4170. https://doi.org/10.3390/app14104170

3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

4. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

5. Braun, V., & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE Publications.

6. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

7. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

8. Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

9. Hermosilla, P., Berríos, S., & Allende-Cid, H. (2025). Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models. Applied Sciences, 15(13), Article 7329. https://doi.org/10.3390/app15137329

10. Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. ACM Computing Surveys, 55(2), Article 39. https://doi.org/10.1145/3491209

11. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

12. Liu, H., & Lang, B. (2019). machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences, 9(20), Article 4396. https://doi.org/10.3390/app9204396

13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

14. Mittelstadt, B. D., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (pp. 279–288). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287574

15. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

16. National Institute of Standards and Technology. (2020). Security and privacy controls for information systems and organizations (NIST Special Publication 800-53, Revision 5). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-53r5

17. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

18. National Institute of Standards and Technology. (2024). The NIST cybersecurity framework (CSF) 2.0 (NIST Cybersecurity White Paper 29). U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.29

19. Nugraha, B., Jnanashree, A. V., & Bauschert, T. (2025). A versatile XAI-based framework for efficient and explainable intrusion detection systems. Annals of Telecommunications, 80, 1095–1120. https://doi.org/10.1007/s12243-025-01118-9

20. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In 2016 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 372–387). IEEE. https://doi.org/10.1109/EuroSP.2016.36

21. Pascoe, C., Quinn, S., & Scarfone, K. (2024). The NIST cybersecurity framework (CSF) 2.0 (NIST Cybersecurity White Paper 29). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29

22. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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 (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

23. Rjoub, G., Bentahar, J., Abdel Wahab, O., Mizouni, R., Song, A., Cohen, R. S., Otrok, H., & Mourad, A. (2023). A survey on explainable artificial intelligence for cybersecurity. IEEE Transactions on Network and Service Management, 20(4), 5115–5140. https://doi.org/10.1109/TNSM.2023.3282740

24. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

25. Sharma, A., Rani, S., & Shabaz, M. (2025). A comprehensive review of explainable AI in cybersecurity: Decoding the black box. ICT Express, 11(6), 1200–1219. https://doi.org/10.1016/j.icte.2025.10.004

26. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

27. Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 56–67). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372870

28. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

29. Wan, X., Xue, G., Zhong, Y., et al. (2025). Separating prediction and explanation: An approach based on explainable artificial intelligence for analyzing network intrusion. Journal of Network and Systems Management, 33, Article 16. https://doi.org/10.1007/s10922-024-09891-z

30. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

31. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Paper 601, pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300831

32. Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2021). "Why should you trust my explanation?" Understanding uncertainty in LIME explanations. In Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency (pp. 575–585). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445921



2026 by the authors; Journal of Global Social Transformation (JGST). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).