Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00388-1

Review Article

Explainable Artificial Intelligence as a National Security Imperative in Modern Cybersecurity Systems

Bright Duffour [1,]

1. *Computer Science, University of the Potomac, Washington, USA*

Received: June 29, 2026 | Review began: August 03, 2026 | Review ended: October 22, 2026 | Published: November 24, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

These concerns have accelerated global interest in explainable artificial intelligence (XAI) as a strategic requirement for modern cybersecurity systems. XAI aims to improve transparency, interpretability, accountability, and human understanding within AI-driven decision-making processes. Explainable cybersecurity frameworks enable analysts and organizations to understand why specific threats are detected, how risks are prioritized, and what factors influence automated defensive responses. Such transparency strengthens institutional trust, improves regulatory compliance, enhances forensic investigations, and supports collaborative human-AI cybersecurity operations.

Although substantial research has explored AI-based cyber threat detection, malware analysis, intrusion detection, and explainability techniques individually, comparatively little attention has been devoted to positioning explainable AI as a strategic national security capability rather than merely a machine learning (ML) enhancement. As governments and organizations increasingly deploy autonomous AI systems across critical digital infrastructures, ensuring transparency, accountability, and operational trust has become essential for maintaining cyber resilience and protecting sensitive information ecosystems.

This paper investigates the strategic role of XAI in modern cybersecurity systems and examines how interpretable AI architectures can strengthen digital trust, cyber defense accountability, and operational resilience. The study synthesizes current research on explainable cybersecurity, evaluates the security implications of opaque AI decision-making, and analyzes emerging applications across phishing detection, intrusion detection, cyber threat intelligence, misinformation defense, fraud detection, Zero Trust security, and autonomous incident response systems.

This paper makes four primary contributions. First, it establishes the relationship between explainable cybersecurity and national security protection. Second, it analyzes the transparency and trust challenges associated with autonomous AI-driven cyber defense systems. Third, it proposes a conceptual explainable cybersecurity framework integrating interpretable ML, human oversight, governance, and adaptive threat intelligence. Finally, it identifies emerging research directions involving generative AI security, adversarially robust explainable systems, federated explainable learning, trustworthy autonomous cyber defense architectures, and human-centered cybersecurity.

Categories: Network Security, Ethical AI and Responsible Technology, AI applications

Keywords: explainable artificial intelligence (xai), cybersecurity, national security, threat intelligence, phishing detection, trustworthy ai, adversarial machine learning, generative ai security, zero trust security, xai

Introduction And Background

Artificial intelligence (AI) has become an important component of modern cybersecurity systems, supporting threat detection, security operations automation, and the analysis of large-scale network activity [1,2]. Machine learning (ML) and deep learning techniques have expanded the capacity of cybersecurity systems to identify complex attack patterns, process substantial volumes of security data, and support timely responses to emerging threats [1,3]. However, the increasing adoption of complex AI architectures has introduced challenges concerning transparency, interpretability, accountability, and operational trust [4,5]. Many advanced models operate as black-box systems, producing classifications or recommendations without providing explanations that security analysts can readily interpret or independently assess [4,6]. These limitations are particularly consequential in security-sensitive environments where incorrect or poorly understood decisions may affect critical services, institutional operations, or national security interests [7,8].

The evolving cyber threat landscape further complicates these challenges. Generative AI, automated attack techniques, and adversarial ML can increase the sophistication and adaptability of malicious activities [9,10]. Although AI-based cybersecurity tools offer opportunities to improve detection and response capabilities, their outputs may be affected by data quality limitations, adversarial manipulation, model uncertainty, and changing operational conditions [3,11]. In high-consequence environments, effective cybersecurity therefore requires attention not only to predictive performance but also to the ability of human operators to evaluate automated recommendations, understand relevant limitations, and exercise appropriate oversight [7,12].

Explainable artificial intelligence (XAI) encompasses methods and approaches intended to make AI-generated predictions, classifications, and recommendations more understandable to human stakeholders [4,5]. In cybersecurity, explainability may help analysts investigate suspicious activity, interpret model outputs, identify potentially unreliable predictions, and document the reasoning supporting security decisions [1,2]. Nevertheless, explanation methods differ in their technical foundations, fidelity, computational requirements, and usefulness to practitioners [5,13]. The availability of an explanation does not necessarily establish that a model is reliable, that its explanation accurately represents its internal reasoning, or that its use improves operational outcomes [12,13]. These distinctions are especially important when evaluating the potential contribution of XAI to national security and critical infrastructure protection [7,8].

This narrative review examines the relationship between XAI, cybersecurity operations, and national security governance. It critically synthesizes literature on interpretable ML, black-box security models, adversarial threats, human oversight, and AI accountability. Particular attention is given to applications involving phishing detection, intrusion detection, malware analysis, threat intelligence, misinformation defense, Zero Trust security architectures, and autonomous incident response [2,14,15]. The review also considers how explainability may support analyst decision-making and organizational accountability while recognizing the limited availability of large-scale operational evidence from government, military, and critical infrastructure environments.

The central argument is that explainability represents a potentially important component of trustworthy AI-enabled cyber defense, particularly in environments requiring auditable decisions and meaningful human oversight [7,8]. However, the review does not assume that explainability is universally necessary or sufficient for effective cybersecurity. Instead, it evaluates the conditions under which XAI may provide operational value, the limitations of existing methods, and the research needed to establish their effectiveness in national security applications.

Review objectives

This review is guided by four principal objectives:

To examine the evolution of AI techniques in modern cybersecurity and assess their implications for security-sensitive and national security environments.

To critically evaluate major XAI methods used in cybersecurity, including their interpretability, technical limitations, potential vulnerabilities, and suitability for high-consequence applications.

To analyze how XAI may support transparency, accountability, human oversight, and operational decision-making across cybersecurity applications, including Zero Trust architectures and autonomous incident response.

To identify research gaps, implementation challenges, evaluation requirements, and future directions for integrating explainability into national cyber defense and critical infrastructure protection.

These objectives establish the scope of the narrative synthesis and provide the basis for evaluating the technical, operational, and governance dimensions of explainable cybersecurity.

Evolution of AI-driven cybersecurity

The evolution of cybersecurity technologies provides important context for understanding the growing interest in explainability. Traditional signature-based and rule-based detection mechanisms rely primarily on predefined patterns and known indicators of malicious activity. Subsequent developments in ML and deep learning expanded the use of statistical classification, behavioral analysis, and automated pattern recognition [3,16]. More recently, generative AI and increasingly autonomous security technologies have introduced additional capabilities alongside new concerns regarding reliability, adversarial exploitation, and human control [9,15].

Figure *1* presents a conceptual overview of this technological progression and highlights the relationship between increasingly sophisticated cybersecurity capabilities and the challenges of interpreting automated security decisions.

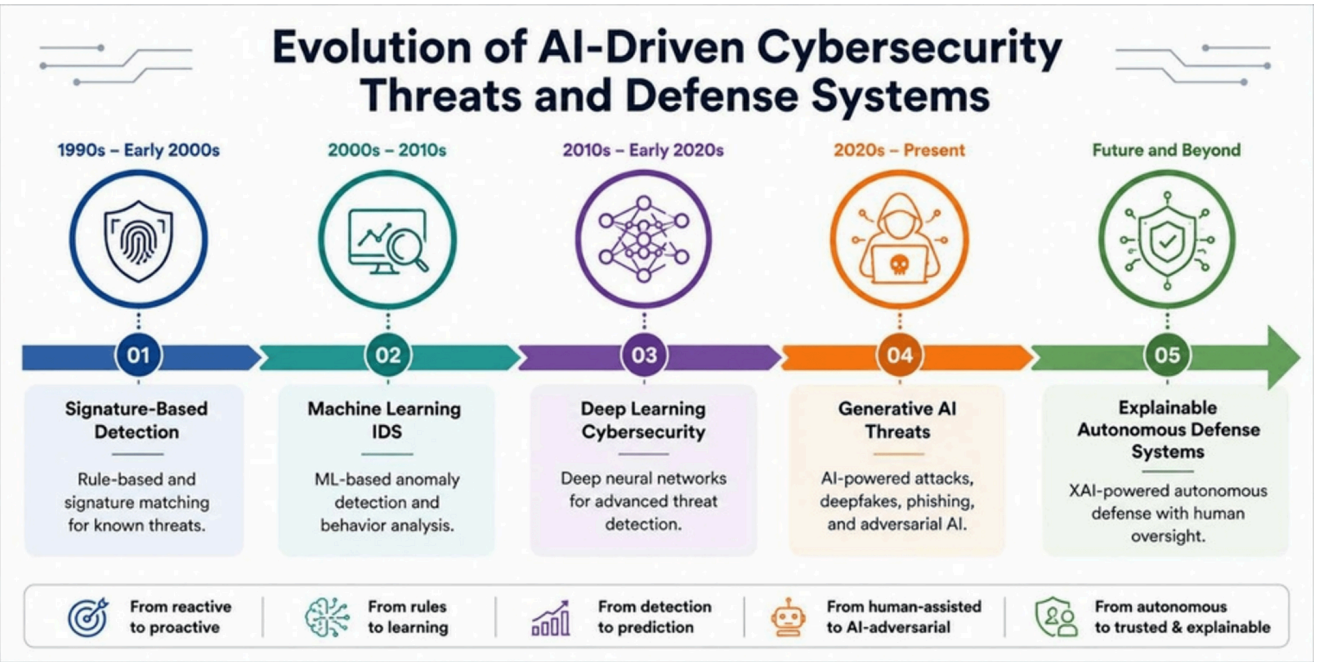

FIGURE 1: Evolution of AI-Driven Cybersecurity Threats and Defense Systems

AI, Artificial Intelligence; ML, Machine Learning

Figure *1* conceptually illustrates the progression from traditional signature-based detection to ML, deep learning, and emerging AI-enabled cybersecurity architectures. It highlights how increasing automation and model complexity have created additional requirements for transparency, verification, and human oversight. XAI is presented as a complementary approach intended to improve the interpretability of AI-assisted security decisions, rather than as a replacement for established cybersecurity controls or a guaranteed solution to emerging threats.

The progression illustrated in Figure *1* provides a foundation for examining the relationship between predictive capability, interpretability, operational oversight, and cybersecurity governance. It also highlights the importance of evaluating explainability as part of broader security architectures rather than treating it as an isolated technical feature.

Research gap and contribution of this review

Previous research has examined XAI methods and their applications to cybersecurity, including the technical characteristics, implementation challenges, and limitations of explanation techniques. Reviews by Capuano et al. [1], Charmet et al. [17], and Rjoub et al. [2] provide important foundations for understanding XAI within cybersecurity. Reynaud and Roxin [18] also contribute to this developing literature, while Yapar [8] explicitly addresses explainability in national security through the perspectives of trust and accountability.

These contributions establish that explainability in cybersecurity and national security is not an unexplored research area. Accordingly, the present review does not claim originality through the identification of XAI as a national security concern or through the introduction of previously unknown explanation techniques. Instead, it focuses on the analytical relationship among technical explanation methods, human interpretation, operational decisions, governance responsibilities, and adaptive security responses.

The specific research gap addressed is the need for a cross-disciplinary synthesis that connects these dimensions within a common conceptual structure and identifies how their interactions might be evaluated in security-sensitive environments. Technical explanations alone do not establish operational trustworthiness. Their potential value depends on whether explanations are sufficiently faithful and understandable, whether analysts can use them appropriately, whether decisions remain subject to meaningful oversight, and whether governance mechanisms can support accountability when automated systems behave unexpectedly [7,12,13].

To examine these relationships, the review proposes a conceptual framework organized around five interconnected dimensions:

Explainable security analytics: The generation of interpretable information about AI-assisted threat detection and classification.

Human interpretation and oversight: The ability of security analysts to assess explanations, challenge automated outputs, and exercise appropriate decision-making authority.

Operational decision support: The integration of explanations into threat investigation, incident prioritization, and security response workflows.

Governance and accountability: The documentation, evaluation, and oversight of AI-assisted decisions in accordance with organizational responsibilities and applicable security requirements.

Adaptive feedback and threat intelligence: The use of analyst feedback, operational observations, and evolving threat information to support continued evaluation and refinement of AI-enabled security processes.

The contribution of this framework is its organization of technical, human, operational, and governance considerations into an integrated structure for examining explainability in cybersecurity. This synthesis is intended to support the identification of evaluation criteria, implementation dependencies, and research priorities that may not be fully captured when explanation methods are considered independently of their operational context.

The proposed framework is conceptual and has not been empirically validated or demonstrated to outperform existing cybersecurity architectures. It should therefore be understood as an analytical and research-oriented contribution rather than an experimentally verified security solution. Future investigations are needed to assess its applicability, practical limitations, and potential benefits through comparative studies, analyst-centered evaluations, and operational testing in appropriate security environments.

By distinguishing established findings from conceptual proposals and unresolved research questions, this review aims to contribute a balanced assessment of XAI as a potentially valuable component of trustworthy and accountable cybersecurity systems.

Review

Review methodology

This study employed a structured narrative review methodology to examine the development, applications, limitations, and national security implications of XAI in cybersecurity. Unlike a systematic review or meta-analysis, the purpose was to synthesize relevant technical and interdisciplinary literature, compare established approaches, identify recurring challenges, and develop an integrated conceptual perspective on explainability, operational decision-making, human oversight, and cybersecurity governance.

Literature search strategy

Relevant literature was identified through searches of major scientific databases and digital libraries, including IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Wiley Online Library, Google Scholar, arXiv, NIST publications, and USENIX conference proceedings.

The literature search focused primarily on publications from 2015 through 2025, reflecting the rapid evolution of explainable AI, modern ML, and AI-driven cybersecurity. Searches were performed using combinations of keywords including:

"Explainable Artificial Intelligence"

"Explainable AI"

"XAI"

"Cybersecurity"

"National Security"

"Trustworthy AI"

"Machine Learning Security"

"Intrusion Detection"

"Adversarial Machine Learning"

"Critical Infrastructure"

"Cyber Threat Intelligence"

"Large Language Model Security"

Boolean operators (AND, OR) were used to refine search queries. Representative searches included:

"Explainable AI" AND Cybersecurity

"XAI" AND National Security

"Trustworthy AI" AND Cyber Defense

"Adversarial Machine Learning" AND Explainability

Eligibility criteria: Publications were considered eligible when they examined XAI or interpretable XAI, relevant cybersecurity applications, cyber defense operations, or national security implications. Eligible source types included peer-reviewed journal articles, conference proceedings, government publications, recognized technical reports, and relevant research preprints. Publications were restricted to English-language materials.

Sources were excluded when they lacked a meaningful connection to explainability or cybersecurity, duplicated previously identified publications, or contained insufficient methodological or technical information for the intended synthesis.

Preprints were considered where relevant to emerging technologies, particularly generative AI and large language model (LLM) security. Their preliminary publication status was considered when interpreting the strength of the evidence.

Study selection and evidence synthesis

Titles and abstracts were screened for relevance, followed by full-text examination of potentially eligible publications. Selected sources were evaluated qualitatively according to their methodological focus, cybersecurity application, explanation approach, reported benefits, limitations, and relevance to operational accountability.

The synthesis was organized around technological evolution, XAI methods, cybersecurity applications, emerging adversarial threats, human oversight, governance, and national security implications.

Because this was a narrative review, no quantitative meta-analysis was conducted. The synthesis emphasizes relationships among technical capabilities, operational requirements, and unresolved research questions rather than pooled estimates of effectiveness.

The available manuscript does not document the total number of records retrieved, duplicates removed, publications screened, full texts assessed, or studies finally included. These counts should be reported only if they can be reconstructed from the original search and screening records. Their absence limits the reproducibility of the literature selection process.

Evolution of AI-driven cybersecurity threats and defense systems

The development of cybersecurity technologies has been shaped by changes in attack techniques, computing capabilities, and the availability of security-relevant data. Traditional defenses relied primarily on manually configured rules, known threat signatures, and network access policies. ML subsequently introduced statistical pattern recognition and anomaly detection, while deep learning expanded the capacity to analyze complex and high-dimensional security data [1-3].

More recent developments in generative AI, reinforcement learning, and automated security orchestration have introduced additional capabilities for threat analysis and response. These advances also create challenges concerning model reliability, adversarial robustness, interpretability, and human oversight [9,11,15].

The relationship between predictive performance and interpretability is not necessarily a fixed trade-off. Some interpretable models achieve competitive performance for particular tasks, while explanation methods can make complex models more accessible without fundamentally changing their predictive architecture [4,5]. Nevertheless, highly complex systems may be difficult to audit or interpret, especially when explanations are unstable, incomplete, or poorly aligned with analysts' operational needs [12,13].

Traditional cybersecurity approaches

Rule-Based Systems

Rule-based cybersecurity systems use predefined logical conditions to identify suspicious events or enforce security policies. They remain important in intrusion detection, email filtering, access control, and automated security workflows.

Their principal advantages include transparency, predictable behavior, and ease of auditing. Analysts can generally identify which rule triggered an alert and determine whether the associated condition remains appropriate.

However, manually maintained rule sets may struggle to identify unfamiliar attack patterns and require regular updates as adversaries change their techniques. Rule-based systems therefore provide valuable control and interpretability but are not sufficient as standalone defenses against every emerging threat [1,3].

Signature-Based Detection

Signature-based detection identifies known malicious activity through established patterns, fingerprints, or indicators of compromise. Antivirus products and intrusion detection systems have historically relied on signatures to recognize previously characterized malware and attacks.

Signature matching remains useful when reliable indicators are available. However, obfuscation, polymorphism, and rapidly changing malicious artifacts can reduce the effectiveness of static signatures. Research on adversarial evasion demonstrates how attackers can deliberately manipulate inputs to circumvent ML-based security classifiers [10,19].

Consequently, signature-based methods are commonly complemented by behavioral analysis, heuristic detection, and ML.

Static Firewalls

Traditional firewalls enforce network traffic policies using predefined conditions involving addresses, ports, protocols, and connection states. They remain foundational security controls for restricting unauthorized communication.

However, static filtering alone cannot reliably identify every threat involving compromised credentials, malicious activity within authorized sessions, or complex application behavior. Modern cybersecurity architectures therefore combine network controls with monitoring, identity management, endpoint protection, and behavioral detection [14].

These traditional approaches illustrate an important principle: interpretability and operational usefulness do not depend exclusively on advanced AI. XAI should complement established security controls rather than replace them without demonstrated benefit.

Machine learning in cybersecurity

ML supports cybersecurity by identifying statistical patterns associated with malicious or anomalous activity. Depending on the application, models may use supervised classification, unsupervised anomaly detection, clustering, or other learning approaches.

ML can improve detection flexibility where appropriate training data and evaluation procedures are available. However, performance depends on factors such as dataset representativeness, class imbalance, evolving threats, adversarial manipulation, and operational deployment conditions [3,11].

Sommer and Paxson [3] highlighted the difficulties of transferring promising experimental results in ML-based network intrusion detection into complex operational environments. This distinction remains relevant when evaluating contemporary AI-enabled security systems.

Intrusion detection systems

ML-based intrusion detection systems analyze network flows, system logs, and behavioral indicators to identify suspicious activity. Supervised approaches classify known attack patterns, while anomaly detection techniques may identify deviations from expected behavior [2,3].

These methods can support detection of denial-of-service attacks, unauthorized access, lateral movement, and other suspicious activities. Nevertheless, false positives, changing network conditions, and limited generalizability across datasets remain important concerns.

Wei et al. [20] investigated explainable deep learning-based network intrusion detection in the context of active intrusion responses, demonstrating the relevance of connecting explanation mechanisms with security operations.

Explainability can assist analysts in examining which features contributed to a classification. However, an explanation should be evaluated for fidelity and usefulness rather than assumed to establish that the underlying prediction is correct [13,20].

Behavioral analytics

Behavioral analytics examines patterns in user activity, authentication events, network access, and system interactions. Models may identify deviations from established activity baselines that warrant further investigation [1,2].

Applications include account compromise detection, insider threat monitoring, and anomalous access analysis. These systems may support security investigations, but unusual behavior does not necessarily indicate malicious intent.

Human interpretation remains important because legitimate changes in user activity can produce anomalous patterns. Explanations that identify relevant contextual features may help analysts distinguish potentially meaningful security indicators from benign deviations.

Malware classification

ML-based malware classification analyzes characteristics such as executable structure, system behavior, API calls, and communication patterns to distinguish malicious software from benign applications.

Unlike static signature matching, learning-based approaches may identify relationships among features that are not captured by known signatures. However, malware classifiers remain vulnerable to distribution shifts, obfuscation, and adversarial manipulation [10,11].

Severi et al. [21] demonstrated that explanations themselves can be exploited in attacks against malware classifiers. Their work highlights the importance of considering explanation mechanisms as potential components of the attack surface.

Explanation methods may help analysts examine why a model classified a file or behavior as suspicious. Their value depends on whether the identified features are meaningful, stable, and useful for investigation [13,21].

Fraud detection

ML is also used in fraud detection systems to analyze transactions, authentication behavior, device characteristics, and historical activity patterns.

Such systems can assist in identifying suspicious transactions or potential account compromise. However, fraud detection involves significant challenges relating to false positives, privacy, fairness, and the consequences of incorrectly restricting legitimate users [5,7].

Explainable outputs may support case review and accountability by identifying relevant decision factors. Nevertheless, explanations must be evaluated carefully because revealing detailed detection logic can also create opportunities for adversarial exploitation.

Deep learning and autonomous cyber defense

Deep learning enables the analysis of complex security data using multilayer neural architectures. Potential applications include malware classification, intrusion detection, phishing identification, and threat intelligence processing [1,2].

Although these methods can achieve strong predictive performance in specific settings, their complexity may complicate interpretation. Model outputs can also be affected by adversarial examples, distribution shifts, and data quality limitations [10,11].

In high-consequence cybersecurity environments, these challenges strengthen the case for combining automated analysis with monitoring, validation, and appropriate human control. Explainability may contribute to this process but does not independently guarantee model reliability or operational safety [7,13].

Transformer models

Transformer architectures support the analysis of sequential and contextual information and have been applied to security-related text, event sequences, and other structured or unstructured data.

Relevant applications include phishing detection, security log analysis, malicious communication classification, and threat intelligence extraction. Research on LLM security illustrates both the opportunities and the vulnerabilities associated with advanced language-based architectures [9].

Their effectiveness depends on the task, dataset, training procedure, and evaluation conditions. Attention mechanisms and generated explanations should not automatically be interpreted as faithful representations of model reasoning [12,13].

Reinforcement learning security systems

Reinforcement learning investigates how agents can learn policies through interactions with an environment and feedback about their actions.

In cybersecurity research, reinforcement learning has been explored for adaptive defense, resource allocation, intrusion response, and security policy optimization.

Yu et al. [22] examined explanation approaches for deep reinforcement learning-based security applications, highlighting the relevance of interpretability when security systems recommend or perform adaptive actions.

However, many proposed approaches are evaluated in simulated or constrained environments. Real-world deployment raises additional concerns regarding unsafe exploration, reward specification, changing threat conditions, and accountability for automated actions.

Explanations may support the evaluation of recommended actions, but security-critical automation also requires operational safeguards, policy constraints, and mechanisms for human intervention [15,22].

AI-powered security operations center automation

AI-assisted Security Operations Center tools can support alert prioritization, event correlation, incident triage, threat intelligence analysis, and documentation.

Automation may reduce repetitive workloads and help analysts manage substantial volumes of security information. However, automated prioritization can also propagate errors when underlying models are unreliable or when relevant contextual information is missing [2,15].

Explainability may help analysts understand why alerts receive particular risk classifications or response recommendations. Its practical usefulness should be assessed through measures such as investigation accuracy, analyst workload, decision time, and the detection of erroneous recommendations.

Autonomous incident response and human oversight

Autonomous incident response systems aim to identify threats and initiate defensive actions with varying degrees of human involvement. Potential responses include isolating compromised endpoints, restricting suspicious connections, or escalating security incidents [15,20].

The consequences of incorrect actions can be significant, particularly when automated responses interrupt essential services or affect critical infrastructure.

An explainable incident response architecture should therefore distinguish between advisory recommendations, human-approved actions, and narrowly constrained automated responses. Explanation generation, action logging, escalation procedures, and human override capabilities can support accountability, but their effectiveness must be evaluated in operational settings [7,22].

This distinction is especially relevant to national security applications, where the consequences of false positives and inappropriate automated actions may extend beyond individual information systems.

Large language models in cyber threat intelligence

LLMs can assist with threat report summarization, security documentation, vulnerability information retrieval, and the interpretation of unstructured security content [9].

These capabilities may support analysts, particularly when combined with reliable retrieval systems and established verification procedures.

However, LLM-generated content can contain factual inaccuracies, unsupported recommendations, or responses influenced by malicious instructions. The NIST Generative Artificial Intelligence Profile identifies risk management considerations associated with generative AI systems [23].

Prompt injection, information leakage, and inappropriate automation are additional concerns. Liu et al. [24] examined prompt injection attacks and defenses, providing evidence of the security challenges associated with processing untrusted instructions.

Accordingly, the use of LLMs in cybersecurity requires output verification, access controls, monitoring, and carefully defined operational responsibilities. Natural-language explanations should not be assumed to be faithful simply because they appear coherent.

Emergence of generative AI threats

Generative AI introduces new capabilities that can be used for legitimate security activities or malicious purposes. Its implications include scalable content generation, synthetic impersonation, manipulation of AI-assisted workflows, and new forms of information deception [9,23].

AI-generated phishing

Generative AI can assist malicious actors in producing grammatically fluent and contextually tailored phishing messages. Such messages may be more difficult to identify using simple indicators such as spelling errors or unusual formatting.

Effective defense requires combinations of authentication controls, content analysis, behavioral indicators, user awareness, and incident reporting. Explainable classification systems may support investigations by identifying suspicious characteristics, although attackers may adapt when detection mechanisms are exposed [1,9].

Deepfake-enabled attacks

Synthetic audio, video, and image-generation technologies can facilitate impersonation, fraud, and deceptive communications.

In security-sensitive environments, manipulated media may complicate identity verification and the assessment of digital evidence. Chandra et al. [25] examined technical approaches to digital content transparency and the reduction of risks associated with synthetic content.

Explainability can potentially help analysts interpret detection results, but confidence scores and feature explanations are not substitutes for independent verification and provenance assessment.

Prompt injection attacks

Prompt injection occurs when an attacker attempts to influence an AI system through malicious instructions embedded in content that the system processes [24].

LLM-assisted security tools may be exposed to these attacks when analyzing external documents, emails, logs, web content, or retrieved threat intelligence.

Defensive approaches include restricting privileges, separating trusted instructions from untrusted content, validating tool use, and monitoring potentially consequential actions. Explanations may assist investigation but cannot replace architectural security controls [23,24].

Synthetic misinformation and influence operations

Generative AI can facilitate the production of misleading text, manipulated images, synthetic audio, and other deceptive content [9,25].

Such capabilities raise concerns for information integrity and may affect public trust, institutional communication, and security-related decision-making.

However, misinformation detection should not be conflated with conventional network intrusion detection. The two domains involve different evidence, evaluation criteria, and operational risks.

Explainable misinformation analysis may help users and analysts understand the indicators behind a classification, but explanations must not be presented as definitive proof of factual truth.

Adversarial manipulation of AI-enabled security systems

Adversarial ML introduces additional risks when attackers deliberately manipulate model inputs, training data, or operational conditions to influence security outcomes.

Goodfellow et al. [26] examined adversarial examples, while Huang et al. [27] established important foundations for understanding adversarial ML in security-sensitive applications. Subsequent research has further examined evasion attacks and the broader evolution of adversarial threats [10,11,19].

These findings demonstrate that explainability should not be treated as a substitute for adversarial robustness. Explanations may reveal suspicious model behavior or support forensic analysis, but explanation mechanisms can also be manipulated or exploited [13,21].

Explainable artificial intelligence techniques for cybersecurity

XAI encompasses methods intended to make ML predictions, model behavior, or decision processes more understandable [4,5].

Explanations may be local, describing individual predictions, or global, characterizing broader model behavior. They may be generated by inherently interpretable models or through post hoc methods applied to complex models [6].

The DARPA Explainable Artificial Intelligence program described by Gunning and Aha [28] reflects the broader research interest in making AI systems more understandable and useful to human operators.

In cybersecurity, explanation quality should be evaluated in relation to the needs of security analysts and the consequences of incorrect decisions.

Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations (LIME), introduced by Ribeiro et al. [29], uses interpretable surrogate models to approximate the behavior of a target classifier around an individual prediction.

In cybersecurity, LIME may help analysts identify features associated with suspicious classifications in phishing detection, malware analysis, or intrusion detection [1,2].

However, explanations can vary depending on perturbation sampling and model behavior. Their fidelity and stability should therefore be assessed before they are used to support high-consequence decisions [13].

SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP), developed by Lundberg and Lee [30], provides feature attribution methods based on principles from cooperative game theory.

SHAP can support local explanations and broader analysis of feature contributions across multiple predictions. Potential applications include intrusion detection, malware analysis, and fraud investigation [1,2].

However, computational requirements, background data selection, feature dependence, and the interpretation of attribution values can affect practical usefulness. SHAP values describe model attribution under particular assumptions and should not automatically be interpreted as causal explanations.

Saliency and feature attribution methods

Saliency and related attribution methods identify input components associated with a model's output.

In cybersecurity, these methods may highlight influential features in executable files, network traffic representations, or suspicious communications.

However, some attribution methods can produce unstable or misleading explanations. Warnecke et al. [13] examined explanation methods for deep learning in security, highlighting the importance of assessing their behavior and limitations in security applications.

Security applications therefore require careful evaluation of explanation sensitivity, fidelity, and robustness to adversarial manipulation.

Counterfactual explanations

Counterfactual explanations identify changes to input characteristics that could alter a model's prediction [5,6].

For example, a counterfactual may indicate which characteristics would need to differ for a suspicious event to receive a lower-risk classification.

These explanations can support analyst understanding and policy evaluation. Nevertheless, cybersecurity counterfactuals must respect technical and operational constraints. Mathematically minimal changes may not correspond to realistic or permissible changes in security environments.

Rule extraction and symbolic explanations

Rule extraction methods approximate or summarize model behavior through human-readable logical conditions [5,6].

Such explanations can support auditing and operational review, particularly when analysts need to understand the conditions associated with security classifications.

However, simplified rules may omit important relationships captured by the original model. Their usefulness therefore depends on fidelity, completeness, and the suitability of the resulting rules for the intended security task.

Comparative assessment of XAI techniques

No single explanation method is universally optimal for cybersecurity applications.

LIME provides model-agnostic local explanations but can exhibit instability [29]. SHAP offers a structured feature attribution framework but may introduce computational and interpretive challenges [30]. Saliency methods can identify influential input components but require robustness assessment [13]. Counterfactual explanations provide alternative-decision perspectives, while rule extraction supports human-readable summaries of model behavior [5,6].

Calzarossa et al. [31] examined an assessment framework for XAI with cybersecurity applications, reinforcing the importance of evaluating explanation quality rather than assuming that understandable outputs are necessarily reliable.

Method selection should therefore reflect the security task, model architecture, computational environment, analyst expertise, and consequences of incorrect interpretation.

XAI Technique	Typical Cybersecurity Applications	Major Advantages	Principal Limitations	Representative Studies
LIME	Phishing detection, malware classification, intrusion detection	Model-agnostic local explanations	Sampling instability, approximation error	Ribeiro et al. [29]; Capuano et al. [1]
SHAP	Malware analysis, intrusion detection, fraud analysis	Structured feature attribution, local and aggregate analysis	Computationally expensive for large models, feature-dependence assumptions	Lundberg and Lee [30]; Guidotti et al. [6]

Saliency / Feature Attribution	Deep learning malware analysis, network intrusion detection	Visual interpretation of influential features, suitable for neural networks	Sensitive to adversarial perturbations, uncertain fidelity	Warnecke et al. [13]
Counterfactual Explanations	Security decision support, fraud analysis, access control, incident investigation	Demonstrates how alternative outcomes could occur, supports analyst understanding	Unrealistic or infeasible counterfactuals for complex security events	Guidotti et al. [6], Barredo Arrieta et al. [5]
Rule Extraction	Interpretable classification, security auditing	Produces transparent logical rules, facilitates governance and auditing	Possible loss of fidelity when approximating complex models	Guidotti et al. [6]; Adadi and Berrada [4]

TABLE 1: Comparative Summary of Explainable AI Techniques for Cybersecurity Applications

AI, Artificial Intelligence; XAI, Explainable Artificial Intelligence; LIME, Local Interpretable Model-Agnostic Explanation; SHAP, SHapley Additive exPlanation

Table *1* summarizes the main characteristics and limitations of the reviewed techniques. The comparison highlights that explanation quality cannot be assessed solely by whether an output is understandable. Fidelity, stability, robustness, computational cost, and analyst usefulness are also relevant [13,31].

Representative studies in explainable cybersecurity

Existing literature provides a substantial foundation for XAI in cybersecurity.

Adadi and Berrada [4] and Guidotti et al. [6] established broader surveys and taxonomies of explanation methods. Capuano et al. [1], Charmet et al. [17], and Rjoub et al. [2] examined XAI within cybersecurity contexts. Reynaud and Roxin [18] further reviewed XAI for cybersecurity systems.

Other studies have investigated more specialized applications. Wei et al. [20] addressed explainable network intrusion detection, while Taheri et al. [32] examined XAI in federated learning-based intrusion detection for connected vehicles. Sarker et al. [16] considered XAI in cybersecurity automation and digital twin environments.

Yapar's discussion of explainability in national security [8] demonstrates that the relationship between XAI, trust, and accountability has already been addressed in previous scholarship. The contribution of the present review must therefore be distinguished from the general proposition that explainability is relevant to national security.

Study	Research Focus	Main Contribution	Relevance or Limitation for This Review
Adadi and Berrada [4]	General XAI	Survey of explanation concepts and methods	Broad technical foundation rather than a dedicated cyber defense framework
Guidotti et al. [6]	Black-box explanation	Taxonomy of explanation approaches	Supports methodological comparison

Capuano et al. [1]	XAI in cybersecurity	Survey of methods and applications	Establishes prior coverage of explainable cybersecurity
Charmet et al. [17]	XAI in cybersecurity	Literature synthesis of cybersecurity applications	Important related review requiring explicit differentiation
Rjoub et al. [2]	Explainable cybersecurity	Survey of XAI techniques and cybersecurity contexts	Substantial overlap in technical subject matter
Yapar [8]	XAI and national security	Discussion of trust and accountability	Establishes prior national security framing
Wei et al. [20]	Network intrusion detection	Explainable deep learning-based intrusion detection	Specialized operational application
Sarker et al. [16]	Cybersecurity automation and digital twins	Taxonomy and future research perspectives	Relevant to integrated cybersecurity architectures
Reynaud and Roxin [18]	XAI for cybersecurity systems	Review of explainability in cybersecurity	Directly relevant recent synthesis
Taheri et al. [32]	Connected vehicle intrusion detection	Explainable federated learning approach	Domain-specific implementation

TABLE 2: Representative Explainable AI Studies in Cybersecurity

AI, Artificial Intelligence; XAI, Explainable Artificial Intelligence

Table *2* demonstrates that the technical foundations and cybersecurity applications of XAI have been examined extensively. The present review does not claim that previous research overlooked explainability, national security, or governance.

Instead, its proposed contribution is an integrated conceptual synthesis connecting five dimensions: explainable security analytics, human interpretation and oversight, operational decision support, governance and accountability, and adaptive feedback and threat intelligence.

This integration provides a structure for identifying dependencies among explanation generation, analyst action, organizational responsibility, and continuous security evaluation.

An explanation method may produce understandable output without improving analyst performance. Similarly, a governance framework may specify accountability requirements without establishing the technical reliability of explanations. The proposed synthesis focuses on these relationships and the need to evaluate them together [7,12,31].

Proposed conceptual framework for explainable cybersecurity

The conceptual framework developed in this review organizes explainable cyber defense into five interconnected dimensions.

The first dimension, explainable security analytics, concerns the generation of interpretable information about threat classifications, model uncertainty, and relevant security indicators.

The second dimension, human interpretation and oversight, concerns how analysts assess explanations, challenge questionable outputs, and retain appropriate decision-making authority.

The third dimension, operational decision support, addresses the use of explanations during investigation, incident prioritization, and response planning.

The fourth dimension, governance and accountability, concerns documentation, auditability, policy compliance, and responsibility for AI-assisted decisions.

The fifth dimension, adaptive feedback and threat intelligence, addresses how operational observations and analyst feedback may inform continued system evaluation and improvement.

The framework draws conceptually on existing research concerning explainable cybersecurity, trustworthy AI governance, and operational security analysis [1,2,7,8,16,18].

Its purpose is not to prescribe a universally optimal technical implementation but to identify relationships that can guide comparative research, system design, and evaluation.

Interpretability and predictive performance

Figure *2* illustrates the conceptual relationship between interpretability and predictive performance in AI-enabled cybersecurity.

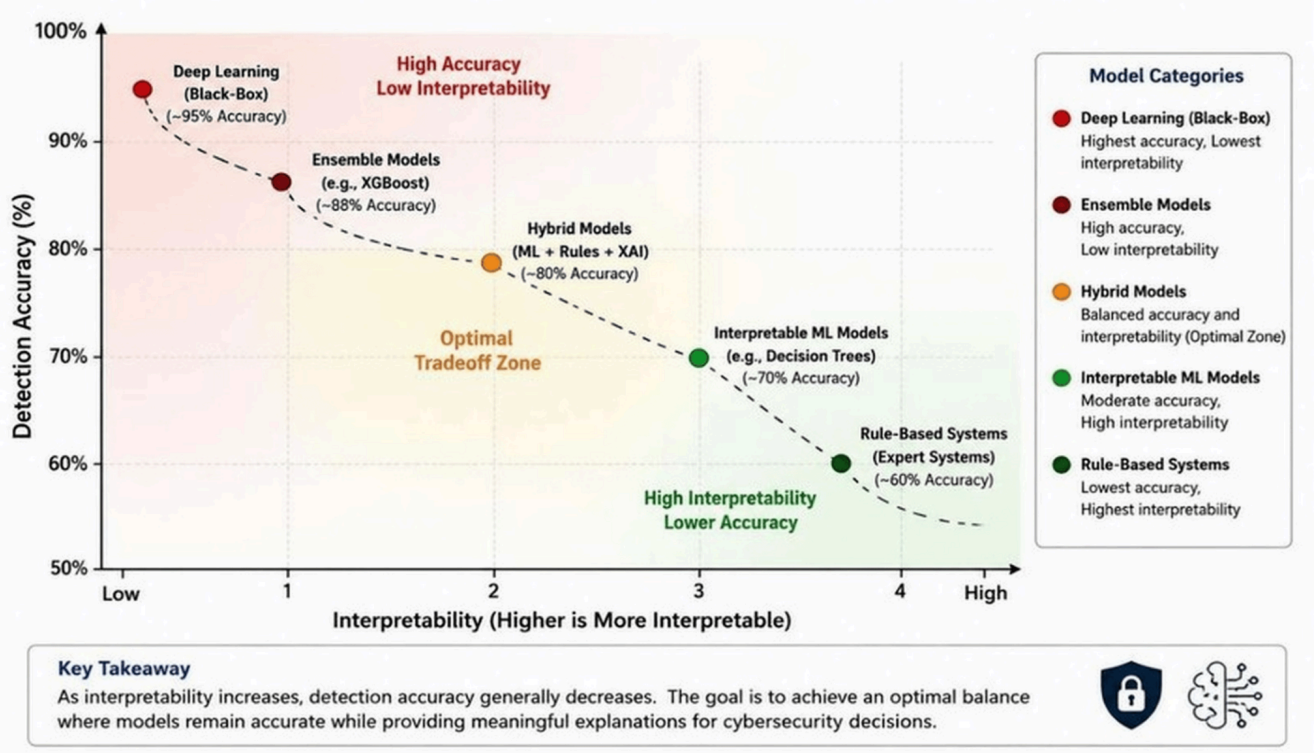

FIGURE 2: Conceptual Relationship Between Interpretability and Predictive Performance in Cybersecurity AI

AI, Artificial Intelligence; XAI, Explainable Artificial Intelligence; ML, Machine Learning

Figure *2* illustrates a conceptual comparison of model interpretability and predictive performance. It supports discussion of model selection rather than establishing an empirically measured or universal trade-off.

Interpretability and predictive performance should be evaluated as separate but potentially interacting characteristics. Their relationship depends on the model, task, dataset, and explanation method [5,12].

An interpretable system is not necessarily less accurate, and a highly accurate system is not necessarily operationally trustworthy.

Building upon these concepts, Figure *3* presents the proposed integration of security data and threat intelligence, AI-assisted analysis, explanation generation, human oversight, operational decision support, governance, and feedback mechanisms. The architecture is conceptual and has not been empirically validated.

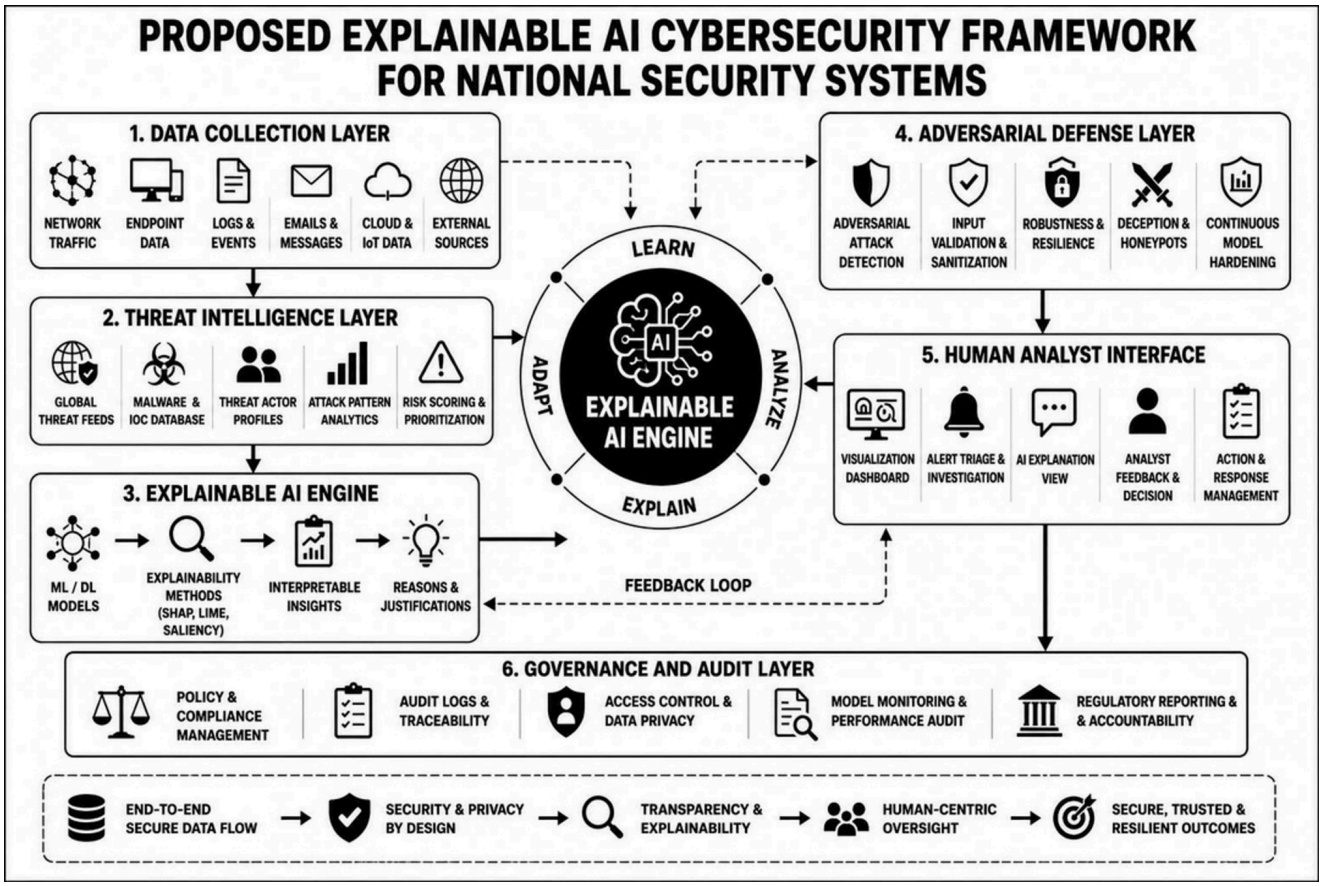

FIGURE 3: Proposed Conceptual Architecture for Explainable AI-Enabled Cybersecurity

AI, Artificial Intelligence; XAI, Explainable Artificial Intelligence; LIME, Local Interpretable Model-Agnostic Explanation; SHAP, SHapley Additive exPlanation

The architecture is intended to illustrate how these functions may interact. Security information is analyzed through AI-assisted detection processes, after which explanations may support analyst interpretation and decisions. Governance mechanisms provide oversight, while feedback processes may inform continued evaluation and refinement [7,8,14].

These relationships are proposed for future investigation and should not be interpreted as demonstrated improvements in accuracy, efficiency, or operational resilience.

Black-box risks in cybersecurity

Figure *4* illustrates selected operational challenges associated with limited model transparency.

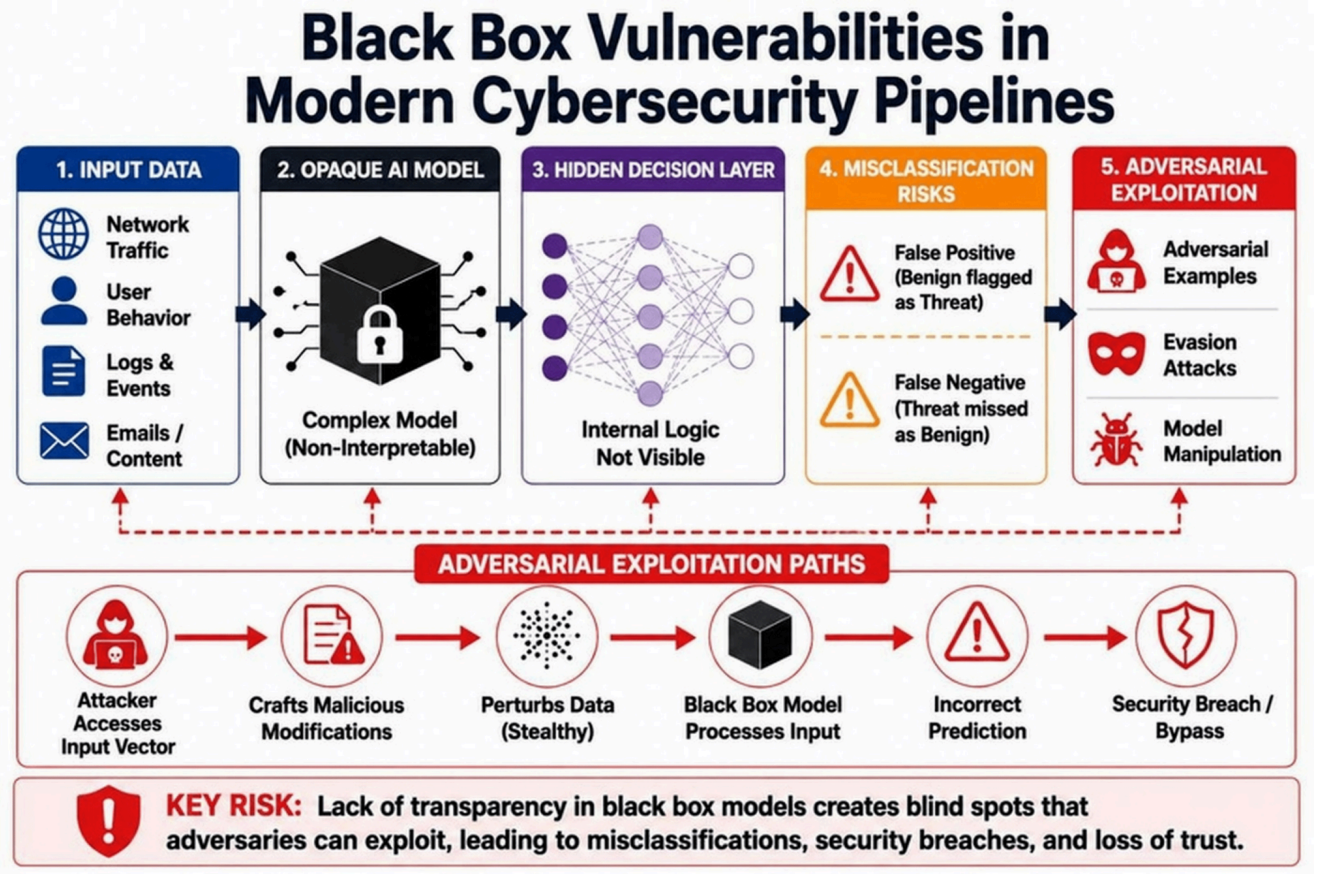

FIGURE 4: The Black Box Problem in AI-Driven Cybersecurity

AI, Artificial Intelligence

Figure *4* conceptually illustrates how limited model transparency may complicate threat interpretation, error investigation, adversarial robustness assessment, and operational accountability.

Black-box systems can make it difficult to determine why particular predictions were generated. However, explainability alone does not prevent false positives, false negatives, adversarial manipulation, or model exploitation [11,13,21].

Effective cyber defense requires the combined use of model evaluation, adversarial testing, access controls, monitoring, analyst oversight, and appropriate governance mechanisms [7,11].

Zero trust architecture and explainable security decisions

Zero Trust Architecture emphasizes continuous verification and the evaluation of access requests according to relevant security policies and contextual information. NIST Special Publication 800-207 provides an established architectural foundation for Zero Trust implementation [14].

AI-assisted risk assessment may contribute to access decisions by evaluating contextual signals, behavioral patterns, and anomalous activity. In such settings, explainability could help administrators understand why a particular access request was classified as suspicious or subjected to additional verification.

However, Zero Trust does not inherently require explainable ML. Many Zero Trust controls can operate through deterministic policies and established identity and access management mechanisms.

The potential contribution of XAI is therefore most relevant when AI-generated risk assessments influence consequential access decisions. Future research should examine whether explanations improve administrator understanding, reduce inappropriate access restrictions, or support the auditing of automated security decisions.

National security implications

XAI may have strategic relevance in security-sensitive environments where automated recommendations influence consequential operational decisions.

Potential applications include critical infrastructure monitoring, government cybersecurity, intelligence analysis, and defensive security operations. In these settings, explanations may assist analysts in evaluating model outputs, documenting decisions, and identifying circumstances requiring additional investigation [7,8].

Nevertheless, evidence from large-scale operational deployments remains limited. It is therefore premature to conclude that XAI is universally necessary or that its adoption automatically improves national security outcomes.

A more defensible conclusion is that explainability is a potentially valuable component of trustworthy AI-enabled cybersecurity when combined with reliable detection methods, operational safeguards, and meaningful human oversight.

Research involving trustworthy federated learning illustrates opportunities for developing collaborative security systems while addressing privacy, robustness, and accountability. Zhang et al. [33] surveyed the challenges and approaches associated with trustworthy federated learning.

Hermosilla et al. [34] compared SHAP and LIME for forensic analysis in intrusion detection, providing application-specific evidence concerning the interpretation of model outputs.

Alketbi and Mehmood [35] reviewed XAI techniques for malicious insider threat detection, identifying opportunities and challenges associated with explanation methods in this specialized security domain.

These studies contribute to the evidence base for explainable cybersecurity but do not independently establish the operational superiority of XAI across national security environments.

Limitations of the review

This review has several limitations. First, the narrative methodology does not provide the exhaustive coverage or reproducibility guarantees associated with a formally conducted systematic review. Although the literature was selected through a structured search and qualitative screening process, relevant publications may have been omitted. The absence of a documented screening count further limits reproducibility.

Second, the rapid development of generative AI, LLMs, and autonomous cyber defense means that emerging findings may extend beyond the primary 2015-2025 search period.

Third, the reviewed literature contains substantial variation in datasets, experimental conditions, explanation methods, evaluation metrics, and application domains. These differences limit direct comparisons and prevent broad conclusions regarding the superiority of particular XAI approaches [13,31].

Fourth, many studies rely on experimental datasets, simulations, or narrowly defined application settings. Evidence from large-scale government, military, and critical infrastructure deployments remains comparatively limited.

Fifth, the proposed conceptual framework has not been implemented or empirically evaluated. Its relationships and potential benefits require testing through comparative studies, analyst-centered evaluations, and operational research.

Finally, the availability of understandable explanations should not be equated with model reliability, explanation fidelity, or improved cybersecurity outcomes. These characteristics require independent assessment [12,13].

Future research should prioritize standardized explanation evaluation, adversarial robustness, human-centered usability testing, comparative framework assessment, and operational validation across security-sensitive environments.

Such work would help establish when explainability provides measurable benefits, which explanation methods are appropriate for particular security tasks, and how technical transparency can be integrated with effective governance and human oversight.

Conclusions

AI has become an important component of modern cybersecurity, supporting threat detection, security operations automation, and threat intelligence analysis. However, the increasing adoption of complex ML, deep learning, and autonomous AI architectures has raised challenges involving transparency, interpretability, accountability, adversarial robustness, and operational oversight, particularly in security-sensitive environments. This narrative review examined the potential role of XAI in addressing these challenges through existing research on interpretable ML, cybersecurity applications, human oversight, and AI governance. Techniques such as LIME, SHAP, feature attribution, counterfactual explanations, and rule extraction may assist security analysts in interpreting model outputs and investigating automated threat classifications. The review proposes a conceptual framework integrating five interconnected dimensions: explainable security analytics, human interpretation and oversight, operational decision support, governance and accountability, and adaptive feedback and threat intelligence. The framework highlights the interaction between technical capabilities, human judgment, organizational responsibilities, and evolving security requirements. Although XAI may provide strategic value in critical infrastructure protection, government cybersecurity, and national defense, current evidence does not establish its universal necessity or consistent superiority over alternative cybersecurity architectures. Existing research relies largely on experimental datasets, simulations, and narrowly defined applications, with limited evidence from large-scale operational deployments. Furthermore, the proposed framework has not been empirically validated. Therefore, explainability should be considered one component of a broader trustworthy cybersecurity strategy incorporating predictive reliability, adversarial robustness, access controls, human oversight, and effective governance, particularly in autonomous incident response and Zero Trust environments.

Future research should prioritize standardized evaluation frameworks assessing explanation fidelity, consistency, robustness, computational efficiency, and usefulness to security analysts. Comparative studies should investigate whether different explanation techniques improve investigation accuracy, decision quality, response efficiency, and operational accountability across cybersecurity applications. Further research should examine the integration of XAI with federated learning, privacy-preserving ML, neuro-symbolic AI, digital twins, large language models, and autonomous security operations, particularly agentic AI and automated incident response. Empirical validation of the proposed framework through prototype implementations, controlled comparisons, analyst-centered experiments, and operational assessments in government and critical infrastructure environments is also necessary. Such studies should compare the framework with existing approaches and determine whether integrating explainability, human oversight, governance, and adaptive feedback produces measurable operational benefits. Additionally, researchers should investigate vulnerabilities arising from explanation mechanisms, including the disclosure of exploitable information and manipulation by attackers, through adversarial testing, explanation robustness assessments, and appropriate safeguards. Collaboration among academic researchers, cybersecurity practitioners, industry organizations, and public-sector institutions will be important in establishing evaluation criteria and addressing implementation challenges. Ultimately, the contribution of XAI to national cybersecurity resilience will depend on demonstrated operational effectiveness rather than interpretability alone, supported by human-centered evaluation, secure automation, and accountable governance.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Bright Duffour

Acquisition, analysis, or interpretation of data: Bright Duffour

Drafting of the manuscript: Bright Duffour

Critical review of the manuscript for important intellectual content: Bright Duffour

Supervision: Bright Duffour

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

Data sharing is not applicable. This work is theoretical; no datasets were generated or analyzed.

Acknowledgements

Artificial intelligence tool ChatGPT was used to assist with language editing, grammar refinement, manuscript organization, and editorial revisions. The authors independently verified all scientific content, interpreted the literature, developed the conceptual framework and figures, and take full responsibility for the accuracy, originality, and integrity of the manuscript.

References

1. Capuano N, Fenza G, Loia V, Stanzione C: Explainable artificial intelligence in cybersecurity: a survey. IEEE Access. 2022, 10:93575-93600. 10.1109/access.2022.3204171
2. Rjoub G, Bentahar J, Abdel Wahab O, et al.: A survey on explainable artificial intelligence for cybersecurity. IEEE Transactions on Network and Service Management. 2023, 20:5115-5140. 10.1109/tnsm.2023.3282740
3. Sommer R, Paxson V: Outside the closed world: on using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy, Oakland, CA, USA. 2010. 305-316. 10.1109/SP.2010.25
4. Adadi A, Berrada M: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018, 6:52138-52160. 10.1109/access.2018.2870052
5. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020, 58:82-115. 10.1016/j.inffus.2019.12.012
6. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D: A survey of methods for explaining black box models. ACM Computing Surveys. 2018, 51:1-42. 10.1145/3236009
7. Tabassi E: Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, Gaithersburg, MD, USA; 2023. 10.6028/NIST.AI.100-1
8. Yapar O: Explainable AI in national security: enhancing trust and accountability. Journal of Emerging Technologies and Innovative Research. 2023, 10:h691-h709 .
9. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y: A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High-Confidence Computing. 2024, 4:100211. 10.1016/j.hcc.2024.100211
10. Biggio B, Roli F: Wild patterns: ten years after the rise of adversarial machine learning. CCS '18: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018, 2154-2156. 10.1145/3243734.3264418

11. Vassilev A, Oprea A, Fordyce A, Andersen H: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. National Institute of Standards and Technology, Gaithersburg, MD, USA; 2024. 10.6028/NIST.AI.100-2e2023

12. Doshi-Velez F, Kim B: Towards a rigorous science of interpretable machine learning [PREPRINT]. Machine Learning. 2017, arXiv:1702.08608. 10.48550/arXiv.1702.08608

13. Warnecke A, Arp D, Wressnegger C, Rieck K: Evaluating explanation methods for deep learning in security. 2020 IEEE European Symposium on Security and Privacy (EuroS&P), Genoa, Italy. 2020, 158-174. 10.1109/EuroSP48549.2020.00018

14. Rose S, Borchert O, Mitchell S, Connelly S: Zero Trust Architecture NIST Special Publication 800-207. National Institute of Standards and Technology, Gaithersburg, MD, USA; 2020. 10.6028/NIST.SP.800-207

15. Kshetri N, Voas J: Agentic artificial intelligence for cyber threat management. Computer. 2025, 58:86-90. 10.1109/mc.2025.3544797

16. Sarker IH, Janicke H, Mohsin AA, Gill A, Maglaras L: Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. ICT Express. 2024, 10:935-958. 10.1016/j.icte.2024.05.007

17. Charmet F, Tanuwidjaja HC, Ayoubi S, et al.: Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications. 2022, 77:789-812. 10.1007/s12243-022-00926-7

18. Reynaud S, Roxin A: Review of eXplainable artificial intelligence for cybersecurity systems. Discover Artificial Intelligence. 2025, 5:78. 10.1007/s44163-025-00318-5

19. Biggio B, Corona I, Maiorca D, et al.: Evasion attacks against machine learning at test time. Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg; 2013. 387-402. 10.1007/978-3-642-40994-3_25

20. Wei F, Li H, Zhao Z, Hu H: xNIDS: explaining deep learning-based network intrusion detection systems for active intrusion responses. Proceedings of the 32nd USENIX Security Symposium. USENIX Association. 2023, 4337-4354.

21. Severi G, Meyer J, Coull S, Oprea A: Explanation-guided backdoor poisoning attacks against malware classifiers. Proceedings of the 30th USENIX Security Symposium. USENIX Association. 2021, 1487-1504.

22. Yu J, Guo W, Qin Q, Wang G, Wang T, Xing X: AIRS: explanation for deep reinforcement learning based security applications. Proceedings of the 32nd USENIX Security Symposium. USENIX Association. 2023, 7375-7392.

23. National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. National Institute of Standards and Technology, Gaithersburg, MD, USA; 2024. 10.6028/NIST.AI.600-1

24. Liu Y, Jia Y, Geng R, Jia J, Gong NZ: Formalizing and benchmarking prompt injection attacks and defenses. Proceedings of the 33rd USENIX Security Symposium. USENIX Association. 2024, 1831-1847.

25. Chandra B, Dunietz J, Awad G, et al.: Reducing risks posed by synthetic content: an overview of technical approaches to digital content transparency. NIST AI 100-4. National Institute of Standards and Technology, Gaithersburg, MD, USA; 2024. 10.6028/NIST.AI.100-4

26. Goodfellow IJ, Shlens J, Szegedy C: Explaining and harnessing adversarial examples [PREPRINT]. Machine Learning. 2014, arXiv:1412.6572. 10.48550/arXiv.1412.6572

27. Huang L, Joseph AD, Nelson B, Rubinstein BI, Tygar JD: Adversarial machine learning. Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence. ACM, New York, NY. 2011, 43-58. 10.1145/2046684.2046692

28. Gunning D, Aha DW: DARPA's explainable artificial intelligence (XAI) program. AI Magazine. 2019, 40:44-58. 10.1609/aimag.v40i2.2850

29. Ribeiro MT, Singh S, Guestrin C: Why should I trust you?": explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY. 2016, 1135-1144. 10.1145/2939672.2939778

30. Lundberg SM, Lee SI: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017, 30.

31. Calzarossa MC, Giudici P, Zieni R: An assessment framework for explainable AI with applications to cybersecurity. Artificial Intelligence Review. 2025, 58:150. 10.1007/s10462-025-11141-w

32. Taheri R, Jafari R, Gegov A, Arabikhan F, Ichtev A: Explainable AI for federated learning-based intrusion detection systems in connected vehicles. Electronics. 2025, 14:4508. 10.3390/electronics14224508

33. Zhang Y, Zeng D, Luo J, Fu X, Chen G, Xu Z, King I: Survey of trustworthy federated learning: issues, solutions, and challenges. ACM Transactions on Intelligent Systems and Technology. 2024, 15:1-47. 10.1145/3678181

34. Hermosilla P, Berríos S, Allende-Cid H: Explainable AI for forensic analysis: a comparative study of SHAP and LIME in intrusion detection models. Applied Sciences. 2025, 15:7329. 10.3390/app15137329

35. Alketbi KS, Mehmood AA: Comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. IEEE Access. 2025, 13:121772-121798.