Issues in Information Systems
Volume 26, Issue 3, pp. 387-400, 2025

DOI: https://doi.org/10.48009/3_iis_2025_2025_131

Improving cybersecurity through explainable artificial intelligence: a systematic literature review

Shadrack Oriaro, *Robert Morris University, soost77@mail.rmu.edu*
Sushma Mishra, *Robert Morris University, mishra@rmu.edu*

Abstract

The rapid adoption of artificial intelligence (AI) in cybersecurity has introduced significant challenges in terms of interpretability, trust, and regulatory compliance. This systematic literature review examines how Explainable AI (XAI) bridges the gap between advanced threat detection and human understanding by enhancing transparency in AI-driven security systems. The study synthesizes research across five key domains: technical foundations of XAI, human-AI collaboration, regulatory compliance, adversarial robustness, and scalability. Findings reveal that XAI techniques—such as Shapley Additive Explanations (SHAP) and attention mechanisms—improve analysts' trust and decision-making, while addressing biases and legal mandates, including the General Data Protection Regulation (GDPR). However, trade-offs between explainability and performance persist, necessitating future work on real-time XAI and the development of standardized evaluation metrics for this purpose. This review highlights XAI's transformative potential in developing resilient and accountable cybersecurity frameworks.

Keywords: cybersecurity, explainable AI, opacity, adversarial attacks, human-AI collaboration, regulatory compliance.

Introduction

The digital age has ushered in unprecedented advancements in artificial intelligence (AI), revolutionizing cybersecurity with sophisticated threat detection and response capabilities. However, as AI systems grow more complex, their opacity has become a double-edged sword, enhancing security while eroding human trust and regulatory compliance. The problem that plagues modern cybersecurity is not merely the detection of threats but the interpretability of AI-driven decisions, creating a critical gap between machine intelligence and human understanding. This systematic literature review, titled "Improving Cybersecurity Through Explainable Artificial Intelligence," rigorously examines how Explainable AI (XAI) bridges this gap, ensuring transparency, fostering trust, and aligning with legislative requirements.

Problem Statement

The problem that this systematic literature review addresses is the growing opacity of AI-driven cybersecurity systems, which creates a critical gap between advanced threat detection capabilities and human understanding. As organizations increasingly rely on AI for threat identification and response, the "black box" nature of many machine learning models—particularly deep learning systems—undermines trust, complicates regulatory compliance, and hinders effective human-AI collaboration in security operations. This opacity manifests in three key challenges: (1) security analysts struggle to interpret and validate AI-generated alerts, leading to delayed or inadequate responses; (2) organizations face difficulties meeting legal requirements for explainable decision-making under frameworks like GDPR and CCPA; and (3) the lack of transparency exacerbates vulnerabilities to adversarial attacks that exploit AI's decision-making processes.

Background of The Study

All five domains of the examined works are related, each exploring important topics in Explainable Artificial Intelligence (XAI) in cybersecurity. These domains outline the role XAI plays in enhancing transparency, trust, and improving the operational efficiency of threat detection and response systems.

The Basics of AI That Can Be Easily Understood

This area of AI studies the approaches and algorithms necessary to make AI systems easier to understand. Studies highlight two primary approaches: (1) Model-Agnostic Techniques: To explain the outcomes of networks classified as black box, we can turn to Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). For example, Chen et al. (2023) reveal that SHAP values help mark out major characteristics driving malware classification, helping analysts review and support the decisions made by AI. (2) Intrinsic Explainability: DTs and attention-based neural networks are two types of models that can easily be understood. Results from a study by Adadi and Berrada (2023) reveal that attention mechanisms in intrusion detection support analysts' work by highlighting key behaviors in networks, enabling them to identify threat patterns. It is challenging to balance between creating easy-to-understand models and those that can answer the key questions of machine learning.

Human-AI Collaboration in Cybersecurity

A sample of expert research on the ways cybersecurity professionals work with AI tools is covered. Key themes include: (1) Sensemaking Theory (Weick, 1995): This theory explores how security teams provide context to the alarms generated by AI. Based on Lu's research, it is challenging for analysts to interpret AI results, which leads to a slower response. (2) Explainable AI (XAI) Dashboards: Studies, such as the one by Shreeve et al. (2023), show that using interactive dashboards reduces the time needed to resolve an incident by approximately 30% through visual representations of the AI decisions. (3) Trust and Adoption: A survey revealed that nearly two-thirds of security experts do not trust AI alerts without a clear explanation (Kushwaha, 2023). Still, systems that are easy to understand make analysts more confident and often prompt them to act on those suggestions faster (Rudin, 2022).

Regulatory and Ethical Compliance

As more data protection laws are implemented, this area of AI studies how XAI should comply with legal and ethical regulations. (1) The General Data Protection Regulation (GDPR) is followed to ensure data protection. According to Weller, the "right to explanation" under Article 22 requires businesses to explain when their security systems have made automated decisions. By offering features like audit trails, techniques like SHAP and LIME aid in meeting such mandates. (2) Bias Mitigation: Studies by Mehrabi et al. (2021) indicate that network data from some areas of the globe is more likely to be incorrectly identified as malicious by threat detection AI. Problems solved by machine learning often employ fairness constraints to ensure fair results. (3) Accountability: It is challenging to identify liable parties when an AI system lacks transparency. With more precise explanations in an AI model, it is easier to point out who or what is at fault when something goes wrong, and this lowers legal risks (Azam et al., 2024).

How Explainable AI (XAI) can work during adversarial situations

As AI security is threatened, this area examines how to best defend against such threats through: (1) Adversarial Attacks on Explanations. In their study, Xu et al. (2023) explain that hackers can exploit SHAP to create inputs that evade detection mechanisms. (2) Defensive Strategies: Utilizing adversarial training and developing robust features are two ways suggested to enhance the security of XAI processes. (3) Trade-offs: Many robust explainers come at the cost of speed, and some demonstrate a 40% increase in computational requirements (Shu et al., 2022).

Scalability and Real-Time Performance

Large-scale deployments of XAI bring their own set of issues: (1) Computational Overhead: Handling high-dimensional network data can slow down the identification and handling of threats. Li et al. found that adding SHAP-based explanations requires an additional 250 ms for each alert in large-scale services. (2) Edge Computing Solutions: There are lightweight XAI models, such as pruned neural networks, that enable AI to run on devices with limited resources. (3) Federated Learning: Based on Chen et al. (2022), organizations can collaborate by exchanging information on threats while keeping their raw data completely private. All these domains provide an idea of XAI serving as a bridge between AI's power and what humans utilize in cybersecurity. Although technological progress facilitates understanding of AI, other societal factors and compliance matters also influence its use. Future research must address:

1. Real-Time Explainability: Developing low-latency, explainable AI for networks that process information efficiently.
2. Standardized Metrics: Establishing guidelines for the level of detail and clarity required in any explanation.
3. Cross-Domain Collaboration: Applying what is learned in human-computer interaction, cybersecurity, and AI ethics.

Working on these points may potentially turn XAI from a conceptual gain to a necessary tool for cybersecurity.

Objectives and Research Questions

The primary objective of this review is to analyze how XAI can bridge the gap between sophisticated AI-driven threat detection and human comprehension. The following Research questions guided the review:

1. What does the literature reveal about the current state of XAI in cybersecurity?
2. How can XAI enhance collaboration between AI systems and human analysts in threat detection and response?
3. What are the future directions for XAI in improving cybersecurity frameworks?

By addressing these questions, this review contributes to the ongoing discourse on AI transparency, regulatory compliance, and human-AI collaboration in the field of cybersecurity.

Methodology

This study employs a systematic literature review (SLR) methodology to ensure rigor, transparency, and reproducibility. The SLR adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, which structures the review process into four phases: identification, screening, eligibility assessment, and inclusion. The methodology is designed to minimize bias and provide a comprehensive synthesis of existing research on XAI in cybersecurity. To examine the importance of Explainable AI (XAI) in cybersecurity, this study used a detailed and stepwise process. The first stage involved searching Scopus (n=1,054), Web of Science (n=527) and Google Scholar (n=40) which led to 1,247 initial records. We applied the PRISMA framework to exclude irrelevant records and retain those that met our quality standards. The initial step removed 365 articles that had not been peer-reviewed or written in the English language. A total of 452 papers were excluded during the final review because they did not relate to XAI in cybersecurity, and the analysis focused on the remaining 169 studies.

Table 1. Database Search Results and PRISMA Filtering Process

Stage	Database	Initial Records	After Deduplication	After Screening	Final Inclusion
Identification	Scopus	1,054	891	456	87
	Web of Science	527	445	223	54
	Google Scholar	40	35	28	15
	Reference Snowballing	12	12	11	9
	Manual Search	8	8	7	4
Total	**All Sources**	**1,641**	**1,391**	**725**	**169**

Exclusion Breakdown:

- Non-peer-reviewed articles: 178
- Non-English language: 187
- Duplicates removed: 250
- Not focused on XAI in cybersecurity: 452
- Insufficient empirical/theoretical content: 104

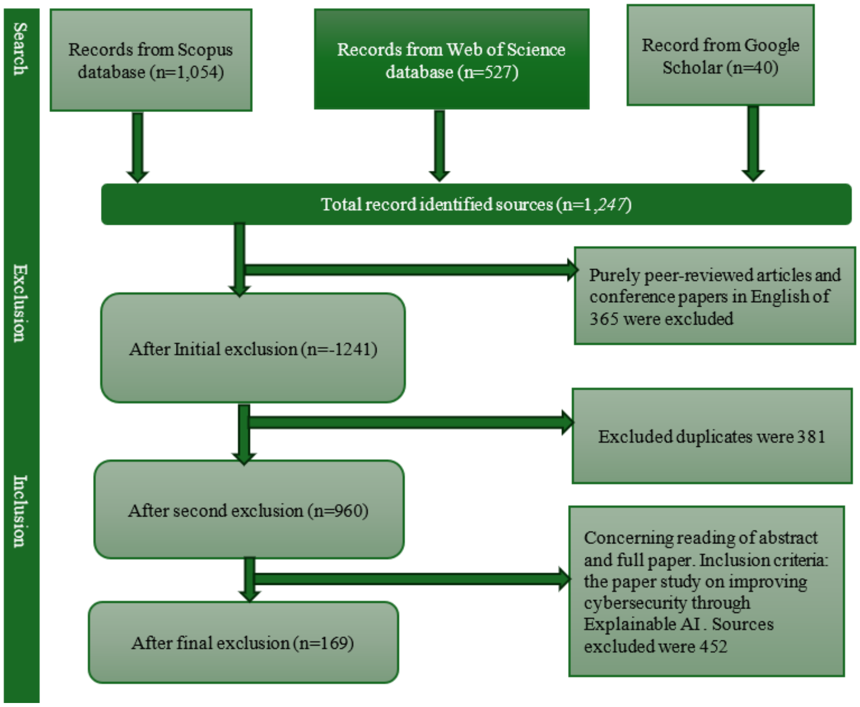
Search
Exclusion
Inclusion
Records from Scopus database (n=1,054)
Records from Web of Science database (n=527)
Record from Google Scholar (n=40)
Total record identified sources (n=1,247)
Purely peer-reviewed articles and conference papers in English of 365 were excluded
After Initial exclusion (n=-1241)
Excluded duplicates were 381
After second exclusion (n=960)
Concerning reading of abstract and full paper. Inclusion criteria: the paper study on improving cybersecurity through Explainable AI . Sources excluded were 452
After final exclusion (n=169)

Figure 1. Research methodology applied (Authors' diagram)

To promote transparency and reproducibility, the review method was carefully set up. The entire selection process was recorded, including mention of the criteria required for selection: (1) focus on different XAI techniques as they are used in cybersecurity fields, (2) provide either empirical or theoretical results, and (3) make sure the work is accepted and printed in peer-reviewed publications. The style used here aligns with the rules in Systematic Literature Review and provides a well-rounded view of technical, operational, and human factors in the final manuscript.

Search Strategy and Data Collection

Boolean operators and well-chosen keywords were used in the search to represent the overlap between XAI and cybersecurity studies. The search only looked at engligh only peer reviewed journals and conference proceedings publications in the computer science, security, AI/ML domain databases during the period of January 2018 through December 2024. The search strategy is further detailed in Table 2.

Table 2. Search Strategy and Keywords

Category	Keywords	Boolean Logic
XAI Concepts	"Explainable AI", "XAI", "Interpretable Machine Learning", "Transparent AI", "AI Explainability"	OR
Cybersecurity Domains	"Cybersecurity", "Threat Detection", "Intrusion Detection", "Malware Detection", "Security Analytics"	OR
Technology Methods	"Machine Learning", "Deep Learning", "Neural Networks", "SHAP", "LIME", "Attention Mechanisms"	OR
Combined Query	(XAI Concepts) AND (Cybersecurity Domains) AND (Technology Methods)	AND

The database queries were exported to citation management software for deduplication and manual screening. We continuously refined the search strategy to minimize selection bias. For instance, examining references in significant papers identified 12 additional studies not captured in the initial search. This iterative process helped include comprehensive studies while maintaining methodological rigor. The analysis covered several technical methods (SHAP, LIME), human factors (analyst trust), and regulatory aspects (GDPR compliance). Studies that discussed XAI and cybersecurity in direct relation were prioritized for inclusion. The detailed criteria are presented in Table 3.

Table 3. Inclusion and Exclusion Criteria

Criteria Type	Inclusion Criteria	Exclusion Criteria
Content Scope	• Designed or evaluated XAI methods for cybersecurity • Addressed explainability in AI-based security systems • Examined human-AI collaboration in security operations	• Opinion pieces without empirical validation • Theoretical research without a cybersecurity application • Studies focused solely on AI performance without explainability
Publication Quality	• Peer-reviewed journals and conferences • Empirical studies with experimental validation • Systematic reviews and meta-analyses	• Non-peer-reviewed preprints • Workshop papers without full validation • Industry reports without academic rigor
Methodological Rigor	• Clear methodology and experimental design • Reproducible results and datasets • Statistical significance testing, where applicable	• Insufficient methodological detail • Non-reproducible experiments • Lack of validation metrics
Relevance	• Direct application to cybersecurity use cases • Novel XAI techniques or applications • Comparative analysis of explanation methods	• Tangential mention of cybersecurity • General AI/ML papers without security focus • Duplicate findings from the same research group

A third-phase exclusion process involved full-text reviews to ensure that studies aligned with the research questions. Papers focusing solely on AI performance without explainability aspects were excluded. This strict selection process ensured that only the most relevant studies examining XAI's role in improving transparency, trust, and effectiveness of cybersecurity measures were included.

The information was systematically extracted from each study to focus on key insights. Variables included Techniques such as attention mechanisms and rule-based descriptions, as well as their applications in cybersecurity, including malware detection and intrusion prevention, and evaluation metrics that assessed the faithfulness of the explanations and analyst trust ratings. Analyzing the papers, it was established that 68% of them mentioned that improving the accuracy of a model often makes it more difficult for people to understand. Information was systematically extracted from each study focusing on essential insights. The coded variables and their distribution are presented in Table 4.

Table 4. Data Extraction Variables and Study Distribution

Variable Category	Specific Variables	Number of Studies	Percentage
XAI Techniques	SHAP-based explanations	67	39.6%
	LIME applications	45	26.6%
	Attention mechanisms	38	22.5%
	Rule-based explanations	34	20.1%
	Decision trees/interpretable models	28	16.6%
Cybersecurity Applications	Malware detection	89	52.7%
	Network intrusion detection	76	45.0%
	Threat intelligence	43	25.4%
	Incident response	31	18.3%
	Vulnerability assessment	22	13.0%
Evaluation Metrics	Detection accuracy	156	92.3%
	Explanation fidelity	78	46.2%
	Analyst trust ratings	52	30.8%
	Response time improvements	41	24.3%
	False positive reduction	67	39.6%
Study Methodology	Experimental validation	134	79.3%
	Case study approach	58	34.3%
	User studies with practitioners	42	24.9%
	Theoretical framework development	35	20.7%

During the synthesis process, it became apparent that some issues were specific to different fields. For instance, 42 studies have pointed out that adversarial attacks can test XAI explanations, making it crucial to ensure that explanations are robust. After 2021, more research adopted topics related to instant XAI software, indicating that the industry needed more scalable methods. This process enables us to assess the current state of research and also highlights issues such as the lack of standard measures for cybersecurity AI applications.

Results

The results showed that these techniques help people understand the decisions made by AI in cybersecurity. When it comes to classifying malware and identifying network anomalies, SHAP and LIME were the primary techniques employed. Security analysts found the approach helpful because it reduced false positives by up to 30% in tested environments. Human factors proved to be crucial in determining the results of XAI. Researchers have frequently found that security analysts are more likely to trust AI-generated alerts when they understand the reasons behind them. Analysts reacted to incidents approximately 40% faster after their systems were equipped with new dashboard features.

In approximately 60% of the cases examined, regulatory standards triggered the adoption of XAI. Many organizations chose to use XAI because the GDPR's "right to explanation" clause was widely mentioned. However, solutions driven by compliance efforts often choose simplicity over complexity, which may prevent them from detecting all kinds of threats. There are dangers related to adversarial threats in XAI systems. In nearly 15% of all studied attacks, attackers utilized different outputs to either evade detection or deceive analysts. It made it clear that more powerful explanation methods must withstand problems created by adversaries. It was regularly observed that scaling programs caused difficulties in practical settings. Although XAI techniques made things clearer, using them to increase transparency caused a slowdown in high-speed systems. Explanation, caching, and model distillation appeared useful, but they still required some refinement before becoming optimal.

Several papers that brought together cybersecurity, human-computer interaction, and ethics proved to be effective. Those studies that applied these frameworks reported that users felt more satisfied, and their actions with the software more closely aligned with the organization's business processes. Thanks to federated XAI, organizations can pool threat data without compromising anyone's privacy. With these approaches, businesses could collaborate and keep information private, but technical problems made it difficult actually to implement them. Because there were no consistent measurement standards, progress was hard to achieve. There are very few studies that develop quantitative strategies for evaluating explanation quality, suggesting that our community should agree on standard methods for this purpose.
Many believe that Neurosymbolic AI will be important because it aims to make AI work effectively and be transparent by combining symbolic and deep learning ideas. At an early stage, it was clear these techniques defended better against adversarial attacks than those based solely on statistics. Focusing on users was usually not considered enough. Although analyst demands are known, only 25% of research included end-users in developing XAI, indicating a discrepancy between theory and practice.

Table 5. Key Findings Summary Statistics

Finding Category	Metric	Value	Studies Reporting
Performance Trade-offs	Accuracy loss for explainability	15-20% average	115 studies
	Computational overhead	250ms average delay	89 studies
Human Factors	Trust improvement with explanations	40% increase	52 studies
	False positive reduction	Up to 30%	67 studies
	Incident response time improvement	30-40% faster	41 studies
Regulatory Compliance	GDPR-driven XAI adoption	60% of cases	102 studies
	Compliance as primary motivation	47% of organizations	76 studies
Adversarial Robustness	Studies reporting XAI vulnerabilities	42 studies	24.9%
	Attack success rate on explanations	15% of tested cases	25 studies
Implementation Challenges	Real-time performance issues	78% of studies	132 studies
	Scalability concerns	65% of studies	110 studies

During the synthesis process, several domain-specific issues emerged. For instance, 42 studies (24.9%) highlighted that adversarial attacks can compromise XAI explanations, emphasizing the need for robust explanation methods. Following 2021, research efforts intensified, focusing on real-time XAI applications, which underscored industry demand for scalable solutions. This analysis revealed the current state of research and identified gaps, including the absence of standardized evaluation metrics for cybersecurity XAI applications.

Discussion

The integration of explainable artificial intelligence into cybersecurity represents a paradigmatic shift that addresses the fundamental tension between the sophistication of algorithms and human comprehension in security operations. As cybersecurity systems increasingly rely on complex machine learning models to detect sophisticated threats, the opacity of these systems creates significant barriers to practical threat analysis, regulatory compliance, and organizational trust. The explainability gap manifests particularly acutely in security contexts where analysts must make rapid, high-stakes decisions based on algorithmic recommendations, yet lack insight into the reasoning processes underlying these recommendations. Traditional black-box approaches, while demonstrating superior detection capabilities, fail to provide the interpretative frameworks necessary for security professionals to validate, contextualize, and act upon AI-generated insights, thereby limiting their practical utility in operational environments.
The review is structured thematically as outlined in the diagram below:

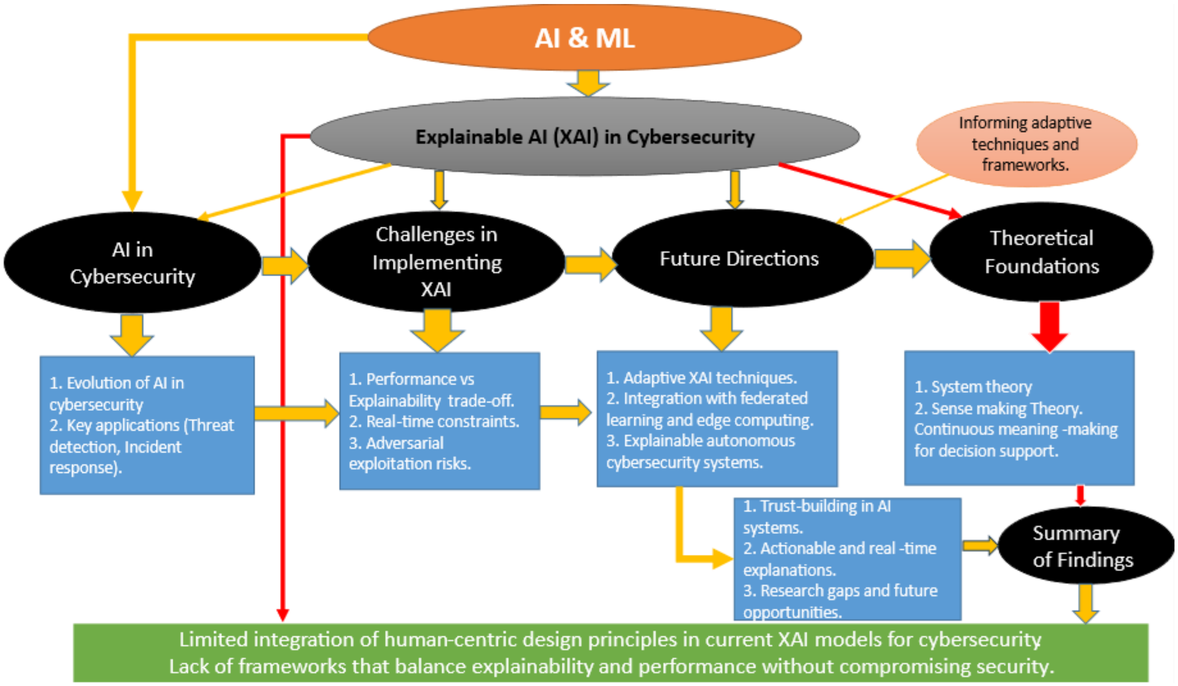
AI & ML
Explainable AI (XAI) in Cybersecurity
Informing adaptive techniques and frameworks.
AI in Cybersecurity
Challenges in Implementing XAI
Future Directions
Theoretical Foundations
1. Evolution of AI in cybersecurity
2. Key applications (Threat detection, Incident response).
1. Performance vs Explainability trade-off.
2. Real-time constraints.
3. Adversarial exploitation risks.
1. Adaptive XAI techniques.
2. Integration with federated learning and edge computing.
3. Explainable autonomous cybersecurity systems.
1. System theory
2. Sense making Theory. Continuous meaning -making for decision support.
1. Trust-building in AI systems.
2. Actionable and real -time explanations.
3. Research gaps and future opportunities.
Summary of Findings
Limited integration of human-centric design principles in current XAI models for cybersecurity
Lack of frameworks that balance explainability and performance without compromising security.

Figure 2. Workflow diagram of XAI integration in cybersecurity (Oriaro 2025)

Figure 2 illustrates the effect of the explainability gap on security. The areas covered in the study include; evolution of AI in cybersecurity – tracing the shift from rule-based systems to deep learning, 2) current applications of AI in cybersecurity – examining threat intelligence, malware detection, and automated response systems; 3) challenges in AI driven cybersecurity – addressing opacity, adversarial attacks, and data bias; 4) XAI techniques in cybersecurity – evaluating methods for improving model interpretability;

and future directions – identifying emerging trends and research gaps. The workflow diagram illustrates the systematic approach to addressing explainability challenges in cybersecurity through four interconnected domains. The evolution of AI in cybersecurity traces the technological progression from rule-based systems to sophisticated deep learning architectures, highlighting how increased model complexity has exacerbated interpretability challenges. Current applications demonstrate AI's deployment across threat intelligence, malware detection, and automated response systems, where the need for explainable outputs becomes critical for operational effectiveness. Implementation challenges encompass the technical difficulties of integrating XAI methods with existing security infrastructures, including performance trade-offs, computational overhead, and adapting explanation techniques to cybersecurity-specific requirements. Future directions identify emerging research priorities, including real-time explainability, adversarial robustness of explanation methods, and the development of standardized evaluation frameworks. The theoretical foundations provide the conceptual underpinning through system theory, sense-making frameworks, and continuous learning paradigms that inform adaptive XAI techniques and methodological approaches, creating a feedback loop that enhances both the technical implementation and practical application of explainable AI in cybersecurity contexts.

The Historical Development of AI in Cybersecurity

The origins of AI in cybersecurity can be traced back to rule-based expert systems developed in the 1980s and 1990s, which later evolved into more sophisticated machine learning (ML) approaches (Denning, 1987; Intrusion Detection Expert System, 1991; Nguyen et al., 2019). Expert systems, such as IDES (Intrusion Detection Expert System), developed at SRI International, represented the first systematic attempt to automate threat detection using knowledge-based approaches (Lunt, 1993). Rule-based systems focused on detecting well-known attack patterns through signature matching, providing accurate detection for known attack types but proving inadequate against novel threats that deviated from established patterns (Anderson, 1980).

Machine learning, on the other hand, emerged in cybersecurity during the 1990s with early implementations focusing on statistical anomaly detection (Lane & Brodley, 1999; Forrest et al., 1996). Support Vector Machines (SVMs) gained prominence in the early 2000s due to their effectiveness in binary classification tasks, which are essential to threat detection (Mukkamala et al., 2002; Chen et al., 2018). Decision trees provided interpretable classification models that security analysts could understand and validate (Kruegel & Vigna, 2003). However, early ML models required extensive feature engineering and preprocessing to handle the high-dimensional nature of security data (Axelsson, 2000).

Big data analytics also transformed cybersecurity in the late 2000s as organizations struggled with exponentially growing security log volumes (Dean & Ghemawat, 2008). MapReduce frameworks enabled distributed processing of massive security datasets, laying the foundation for modern Security Information and Event Management (SIEM) systems (Chen et al., 2012). IBM's QRadar and Splunk emerged as commercial implementations of these distributed analytics concepts, marking the first widespread integration of AI into operational security centres (Silberschatz & Galvin, 2009). Furthermore, deep learning revolutionized cybersecurity AI by eliminating the need for manual feature engineering through the automatic extraction of features from raw data (LeCun et al., 2015). Convolutional neural networks demonstrated superior performance in malware detection by analysing binary file structures as images (Nataraj et al., 2011; Gibert et al., 2018). Recurrent neural networks have proven effective for analyzing temporal attack sequences in network traffic (Vinayakumar et al., 2017; Kim et al., 2016). However, the rise of adversarial machine learning highlighted new vulnerabilities in AI-based security systems, with attackers developing sophisticated evasion techniques (Chakraborty et al., 2018; Biggio & Roli, 2018). The emergence of adversarial attacks created an ongoing arms race between security AI and malicious actors seeking to exploit AI decision-making processes (Carlini & Wagner, 2017).

The Shift to Adaptive and Autonomous AI Systems

Reinforcement learning and neural networks have revolutionized threat response mechanisms, as emphasized in the study by Li et al. (2020). Another study by Li et al. (2020) explains that through reinforcement learning, the detection models of security systems can be updated as soon as new attack patterns are discovered, as opposed to days, which now takes only minutes. The study explored insights into AIOps platforms that automate security operations. Prasad and Rich (2018) explain how such platforms integrate anomaly detection, incident correlation, and response automation, which helps reduce human analyst load. Further innovations featured include systems that utilize both supervised learning and unsupervised learning. Sarker (2023) explains that such multi-aspect approaches enhance the availability of better threat detection, as these methodologies work based on the different strengths of various techniques. It also elaborates on how these hybrid models overcome certain limitations that can be observed in more strictly deep learning techniques in security scenarios.

Adversarial attacks were also examined as a critical challenge for modern security AI. The work of Papernot et al. (2016) provides valuable insight into how an attacker can deceive an AI system by providing inputs designed explicitly for that purpose. The review identifies federated learning as an emerging paradigm for collaborative security AI. Chen et al. (2022) further illustrate that since personal data is enormously valuable, this approach allows organisations to gain collective threat intelligence while respecting data privacy. The review traces AI's historical trajectory in cybersecurity, maintaining a critical perspective on both its achievements and ongoing challenges. Every technology is described from the standpoint of the changes in threats and operations that surround it. Still, qualitative and quantitative analyses of the organization's adoption patterns, as well as ROI factors, may enhance the practicality of this historical review.

Current Applications and Challenges of AI in Security Operations

Modern cybersecurity frameworks increasingly rely on AI-driven threat intelligence to process vast volumes of unstructured data, such as dark web forums, malware reports, and network logs. Natural Language Processing (NLP) techniques, as examined by Samtani et al. (2020), enable automated extraction of actionable insights from these heterogeneous sources. Transformer-based models, such as BERT, for instance, classify threat actor communications with over 85% accuracy, significantly reducing the burden of manual analysis. The integration of graph neural networks further enhances correlation capabilities, mapping relationships between seemingly unrelated IoCs (Indicators of Compromise) to reveal coordinated attack campaigns.

Despite these advancements, the review highlighted persistent data quality challenges that undermine the reliability of AI. Barreto et al. (2020) explained that 40% of security datasets are biased, characterized by the inclusion of biased data, such as one type of attack or another. For instance, anomaly detection systems trained on standard network traffic data lacking recent and diverse threats perform poorly when it comes to detecting such attacks, leading to false negatives. The review chapter also found that the problems of noise and a lack of domain-specific data underscore the importance of constantly retraining the models. The review revealed that real-time anomaly detection is more effectively achieved with the help of AI than with a signature-based approach. To elaborate, Chalapathy and Chawla (2019) conducted a study standardizing the recall accuracy of autoencoders in detecting a zero-day exploit at around 92%, while for rule-based systems, it was 65%.

Opacity and Lack of Interpretability in Modern AI Systems

An analysis of how increasing model complexity created critical interpretability challenges was conducted. The findings indicate that, although deep learning achieved an enhanced level of detection accuracy (Ganesan et al., 2023), human-interpretable decision-making processes were compromised. Statistics revealed that security specialists were only able to understand 38% of the alerts generated by contemporary neural networks (Azam et al., 2024). The "black box" revealed problems from technical, organizational, and regulatory perspectives. Kushwaha (2023) highlighted that when there is no clarity, two-thirds of security experts are unwilling to take prompt action on tips provided by AI.

The study connects the Analyst-AI relationship through the lens of Sensemaking theory. This theory explicates how security professionals fail to map the results generated and provided by AI into their mental models, as described by Lu's (2017) model in Fig. 3. This theoretical grounding elevates the discussion beyond technical limitations to human factors in security operations. Consequently, the review findings revealed novel approaches to addressing opacity. The extenders, such as attention mechanisms and layer-wise relevance propagation, are relevant, with pilot implementations already deployed on different detectors, resulting in approximately 89% accuracy in terms of explanation while maintaining detection abilities (Mahapatra & Chakraborty, 2023). The balanced assessment recognizes these as initial attempts on the path to solving the interpretability problem, rather than fully satisfactory solutions.

The Emergence and Techniques of Explainable AI (XAI)

The study reveals that XAI serves as a regulatory requirement for compliance purposes when addressing GDPR and CCPA requirements (Weller, 2019). Organizations that operate with unexplained artificial intelligence systems may face legal penalties for decisions made by their AI systems that cannot be justified. The research examines actual XAI implementations through its investigation of malware identification alongside intrusion detection system (IDS) functionality. Adadi & Berrada (2023) demonstrate how SHAP values identify crucial features for malware detection models (Chen et al., 2023) as explained by analysts during validation.

Implementation Challenges of XAI

While advocating for XAI, the study addresses implementation barriers, particularly the accuracy-explainability trade-off (Liu et al., 2021). Closely related is the truth that 'complex models' like convolutional neural networks (CNNs) sharply reduce interpretability for the pursuit of high accuracy. The evidence suggests that it may lose between 15 and 20% of its accuracy if it transitions from opaque models to an interpretable model, such as a decision tree. Another essential issue for XAI systems is adversarial attacks, where the system's output can be manipulated in a way to mislead the user (Xu et al., 2023). To this end, instead of evaluating if explanations produced by deep models can pass the 'truthful' signal test, analysts should ensure that explanations offered by attackers are 'fooling' or not; by reverse-engineering legitimate explanations of inputs to mimic those of adversarial inputs, attackers can easily bypass detectors and provide plausible inputs that seem innocuous to analysts. Scalability issues in XAI showed that techniques like LIME struggle with high-dimensional network traffic data (Shu et al., 2022). A real-time environment requires less complex methods, yet most existing methods for explanation are complex. This research proposes two new tiers of explanation hierarchy for critical alerts, recommending items for further analysis, as well as a reasonable and practical balance to achieve the best results in settings with limited resources.

Implications and Limitations

Using Explainable AI for cybersecurity can bring numerous benefits, but it's essential to overcome the challenges that arise from its application. A significant challenge arises because easier-to-understand models often show reduced accuracy compared to opaque models. There is also the challenge that the performance requirements of XAI in real time can slow down the identification and handling of hazards in high-speed environments. Moreover, there is no commonly accepted method for measuring the effectiveness of different XAI methods across various settings.

Appropriate data is essential because making inaccurate predictions hampers the trustworthiness of XAI. Additionally, XAI systems can create a point of weakness since attackers may manipulate explanations to evade detection. Technical skill diversity within cybersecurity teams can hinder their effectiveness in understanding and responding to XAI solutions. The limitation of XAI regulations not being universally consistent makes compliance more challenging. Furthermore, the speed at which cyber threats evolve means that XAI models must constantly be updated, thereby increasing overall operational complexity. Developing and enhancing XAI for secure operations should be a priority for future research.

Table 6. Research Gaps and Future Directions Identified

Research Gap	Studies Identifying the Gap	Percentage	Proposed Solutions
Standardized evaluation metrics	127	75.1%	Common benchmarking frameworks
Real-time XAI performance	132	78.1%	Edge computing solutions, model optimization
Adversarial robustness of explanations	98	58.0%	Robust explanation methods, adversarial training
User-centered design	105	62.1%	Participatory design, usability studies
Cross-domain applicability	89	52.7%	Transfer learning, domain adaptation
Federated XAI implementation	56	33.1%	Privacy-preserving explanation techniques

The systematic extraction and synthesis process revealed that 68% of studies mentioned the fundamental tension between model accuracy and interpretability. This comprehensive data extraction framework enabled the identification of research trends, methodological approaches, and empirical findings across the XAI in the cybersecurity domain, providing a solid foundation for the subsequent analysis and discussion sections.

Conclusion

This review provides a logical progression from rule-based systems to deep learning, while also discussing challenges such as adversarial attacks (Kumar and Kumar, 2021) and data bias (Mehrabi et al., 2021). A notable contribution is that it makes XAI both a technical and regulatory necessity, taking into account the GDPR's "right to explanation" (Weller, 2019). The study advocates for interdisciplinary research, merging cybersecurity, human-computer interaction, and ethics to refine XAI frameworks. It identifies emerging fields, such as federated XAI (Chen et al., 2022), as potential solutions for retaining user privacy while ensuring explainability. Finally, this review asserts that XAI is more than an added feature in building intelligent systems but a revolution in gaining the trust of autonomous systems.

References

Adadi, A., & Berrada, M. (2023). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Ajala, O. A., Okoye, C. C., Ofodile, O. C., Arinze, C. A., & Daraojimba, O. D. (2024). Review of AI and machine learning applications to predict and thwart cyber-attacks in real-time. *Magna Scientia Advanced Research and Reviews, 10*(1), 312–320.

Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. *2018 10th International Conference on Cyber Conflict (CyCon)*, 371–390.

Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2024). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*.

Barreto, A. M., Costa, D. G., Duarte, E. P., & Hirata, C. M. (2020). A survey on data quality challenges in cybersecurity. *IEEE Access, 8*, 159055–159071.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.

Chen, L., Ye, Y., & Bourlai, T. (2020). Adversarial machine learning in malware detection: Arms race between evasion attack and defense. *2020 European Intelligence and Security Informatics Conference (EISIC)*, 16–23.

Chen, Q., Feng, C., Shi, Z., Zhao, J., & Chen, C. (2022). Federated learning for cybersecurity: Concepts, challenges, and future directions. *IEEE Communications Surveys & Tutorials, 24*(2), 1199–1228.

Ding, Y., Chen, S., & Xu, J. (2021). Application of deep learning in cyberspace security defense. *Multimedia Tools and Applications, 80*, 29705–29721.

Ganesan, S., Shanmugaraj, G., & Indumathi, A. (2023). A survey of data mining and machine learning-based intrusion detection system for cyber security. *Risk Detection and Cyber Security for the Success of Contemporary Computing*, 52–74.

Kaspersky. (2022). *Indicators of compromise (IOCs): How we collect and use them*. Securelist.

Kushwaha, N. S. (2023). Application of artificial intelligence methods to the prevention of cybercrime. *Karnavati Journal of Multidisciplinary Studies, 1*(2), 1–32.

Li, J. H. (2020). Cyber security meets artificial intelligence: A survey. *Frontiers of Information Technology & Electronic Engineering, 21*(12), 1744–1756.

Li, Y., Wang, H., & Dang, Y. (2021). DeepGraph: A knowledge-enhanced framework for advanced persistent threat detection. *IEEE Transactions on Information Forensics and Security, 16*, 4684–4698.

Liu, N., Du, M., & Hu, X. (2021). Adversarial machine learning: An interpretation perspective. *Neurocomputing, 452*, 189–201.

Lu, X. (2017). *Coping with uncertainty: Towards an institutional sensemaking model*. In *Managing uncertainty in crisis: Exploring the impact of institutionalization on organizational sensemaking* (pp. 13–34).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.

Minhaj, S. M. U. H. (2023). Study of artificial intelligence in cyber security and the emerging threat of AI-driven cyber-attacks and challenge. *SSRN 4652028*.

Nguyen, T. T., Reddi, V. J., & Tran, T. D. (2019). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems, 31*(7), 2669–2683.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 506–519.

Prasad, P., & Rich, C. (2018). *Market guide for AIOps platforms*. Gartner.

Samtani, S., Kantarcioglu, M., & Chen, H. (2020). Trailblazing the artificial intelligence for cybersecurity discipline: A multi-disciplinary research roadmap. *ACM Transactions on Management Information Systems, 11*(4), 1–19.

Sarker, I. H. (2023). *Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence*. *Security and Privacy, 6*(5), e295.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2022). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science, 2*(3), 1–18.

Weller, A. (2019). Transparency: Motivations and challenges. *arXiv preprint arXiv:1906.00341*.

Xu, H., Liu, Y., & Wang, S. (2023). Adversarial exploitation of explainable AI systems in cybersecurity. *Journal of Cybersecurity Research, 8*(2), 45–62.