Journal of *Cybersecurity and Privacy*

MDPI

Systematic Review

Connecting the Dots: A Systematic Literature Review of Explainable AI, Cybersecurity, Human-Centered Design and Edge Computing

Gaia Cecchi [1,*], Fabrizio Benelli [1,2], Mario Caronna [1], Giulia Palma [1] and Antonio Rizzo [1]

[1] Dipartimento di Scienze Sociali Politiche e Cognitive, Università degli Studi di Siena, 53100 Siena, Italy; fabrizio.benelli@gmail.com (F.B.); mario.caronna@unisi.it (M.C.); giulia.palma2@unisi.it (G.P.); antonio.rizzo@unisi.it (A.R.)

[2] Doctoral Program in Big Data and Artificial Intelligence, Universitas Mercatorum, 00186 Rome, Italy

* Correspondence: gaia.cecchi2@unisi.it

Abstract

The incorporation of Artificial Intelligence (AI) into cybersecurity has become widespread, largely propelled by the emergence of Generative AI (GenAI) and Large Language Models (LLMs). While these technologies promise to revolutionize threat detection, they introduce profound challenges regarding explainability, trust, and deployment feasibility in resource-constrained environments. Current research often exhibits a form of technological determinism, prioritizing algorithmic performance over the operational realities of Security Operations Centers (SOCs). This paper presents a hybrid qualitative Systematic Literature Review (SLR) and Mapping Study, adhering to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 guidelines. Our research questions are narrowly focused, seeking to explore how four key domains intersect: (1) Explainable AI (XAI) methods; (2) cybersecurity operations; (3) human-centered design; and (4) the constraints inherent to edge computing. From an initial corpus of 385 records drawn from Scopus and OpenAlex (spanning a search window from 2014 to 2025, with relevant findings heavily clustered in the 2020–2025 period), included studies were evaluated using a quality assessment protocol adapted from Kitchenham's guidelines, scoring each study on a 0–24 scale across four dimensions (Venue Quality, Methodological Rigor, Dataset Realism, and Depth of XAI/Human Validation). The results reveal a significant "validation gap": while 63% of studies claim human-centric relevance, only ~22% incorporate empirical validation with human operators. Furthermore, we identify a critical trade-off between the reasoning power of cloud-based LLMs and the privacy requirements of Edge security. We conclude by proposing a research agenda for "Cognitive SOCs", emphasizing the need for Small Language Models (SLMs), standardized human-centric metrics, and robust hallucination detection mechanisms.

Keywords: cognitivesecurity; edge AI; explainable AI (XAI); generative AI; human-in-the-loop (HITL); large language models (LLMs); security operations center (SOC); systematic literature review

Academic Editor: Martin Gilje Jaatun

Received: 30 March 2026
Revised: 14 May 2026
Accepted: 15 May 2026
Published: 19 May 2026

1. Introduction

Signature-based defenses have proven increasingly inadequate against polymorphic malware and zero-day exploits [1]. Moreover, traditional perimeters are porous to insider threats, which leverage authorized access to evade detection, necessitating multimodal AI

approaches capable of analyzing user behavior beyond simple log anomalies [2]. As Omar and Zangana [3] demonstrate, the static nature of traditional security cannot match the dynamism of modern threats. Consequently, the adoption of Artificial Intelligence (AI) and Machine Learning (ML) has transitioned from a research-oriented endeavor to an operationally critical component of network defense [4] and adaptive access control strategies that transcend static policy enforcement [5], necessitating machine learning frameworks capable of real-time adaptation against evolving threat vectors [6]. As highlighted by recent comprehensive reviews, AI has become pivotal in automating threat recognition and incident response across diverse security domains [7].

Yet, this reliance on Deep Learning (DL) introduces a precarious trade-off: higher accuracy comes at the cost of interpretability. The resulting "Black Box" phenomenon leaves security analysts blind to the model's internal logic [8]. To mitigate these risks, the field has historically relied on foundational post-hoc frameworks such as Local Interpretable Model-agnostic Explanations (LIME) [9], which provides local surrogate explanations, and SHapley Additive exPlanations (SHAP) [10], based on game-theoretic Shapley values. However, modern cybersecurity demands extend beyond simple transparency, requiring resilience against Adversarial XAI attacks—where explanations can be manipulated to hide malicious intent [11]. In high-stakes domains like critical infrastructure, where a single false positive triggers expensive shutdowns, blind trust is unacceptable [12]. This is particularly acute in Internet of Things (IoT)–Health environments, where end-users require granular "security metadata" to maintain situational awareness rather than relying on opaque system assurances [13]. Zhang et al. [14] rightly argue that without Explainable AI (XAI), advanced models remain practically undeployable; operators simply cannot validate, and therefore cannot trust, automated decisions.

1.1. The Generative AI Paradox

The term "paradox" is used here to denote the simultaneous co-occurrence of enhanced defensive capabilities and expanded attack surfaces introduced by the same technology, as documented by Ahi and Valizadeh [15] and Abuadbba et al. [16].

Large Language Models (LLMs) have introduced a dual-use dynamic to cybersecurity operations. While they substantially enhance the capabilities of defensive teams by automating log synthesis and threat intelligence [17], they simultaneously lower the barrier to offensive exploitation. Abuadbba et al. [16] note that the distinction between offensive and defensive AI capabilities is narrowing, enabling less experienced actors to leverage LLMs for generating polymorphic code and targeted phishing content [15]. This escalation compels a pivot toward *Cognitive Security* [18]. The research objective is increasingly oriented not merely toward detection accuracy, but toward the realization of the "Cognitive SOC": an environment where AI systems not only detect threats but mimic human reasoning processes to support analysts in complex decision-making [19]. Building upon this premise, we formally define a Cognitive SOC as a socio-technical ecosystem, where AI-driven automation and human cognitive processes are tightly integrated. Unlike traditional SOCs that rely on deterministic rule-matching and isolated alert generation, a Cognitive SOC leverages multimodal AI (including GenAI and XAI) to dynamically contextualize threats, provide faithful reasoning for automated decisions, and adapt its interface to the cognitive load of the human operator. This aligns with emerging research roadmaps for Self-Adaptive Systems (SAS), which envision Generative AI not just as a tool, but as an orchestrator of the monitoring and planning loops within autonomous defenses [20].

However, realizing this vision of autonomous defense is not without significant hurdles. Despite these benefits, the transition to GenAI-generated explanations introduces

some operational risks. It should be noted that the stochastic nature of LLMs introduces non-deterministic behavior, whereby identical security telemetry may yield inconsistent explanations across different inference runs. Furthermore, high sensitivity to prompt engineering can inadvertently bias the AI's reasoning, while the massive computational overhead of these models poses a critical barrier to scalability in resource-constrained Edge environments.

1.2. The Missing Links: Humans and Edge

Current scholarship frequently suffers from an algorithmic tunnel vision, ignoring the operational realities of the Security Operations Center (SOC). This disconnect primarily manifests across two critical dimensions: human cognitive limits and infrastructural constraints.

On the human front, SOC analysts are subject to severe alert fatigue [21], a condition where high volumes of low-priority alerts diminish the capacity to accurately triage critical events. Paradoxically, an AI system that merely detects threats without providing a contextual explanation can actually contribute to increased analyst cognitive load and burnout. To counter this, recent frameworks propose a transition to "Next-Generation SOCs," where AI and Machine Learning (ML) are integrated holistically across detection, response, and threat intelligence layers to automate routine triage and prioritize critical incidents [22].

Beyond the cognitive burden placed on human operators, the deployment of these AI systems faces stringent infrastructural barriers. Data sovereignty laws (e.g., the General Data Protection Regulation (GDPR)) and strict latency constraints often prohibit offloading sensitive network telemetry to centralized cloud models. Consequently, the industry increasingly demands *Edge AI*—systems capable of executing, and explaining, inference locally [23].

1.3. Contribution and Scope

Existing surveys [24,25] predominantly isolate algorithms from their deployment contexts, neglecting the friction between model complexity and human usability. We aim to address this gap through this Systematic Literature Review (SLR), adhering to PRISMA 2020 guidelines. We interrogate the literature through three lenses:

- **RQ1:** How do XAI techniques adapt to the cognitive demands of modern SOCs?
- **RQ2:** Do current evaluation metrics account for Human Factors such as trust and alert fatigue?
- **RQ3:** What are the tangible trade-offs between LLM capabilities and Edge/IoT resource constraints?

We distilled an initial corpus of 385 records down to 27 specific high-quality primary studies (2020–2025). We found that existing research streams predominantly enhance information processing, system robustness, and detection quality, but provide limited guidance on how decision pathways, escalation actions, and accountability should be structurally designed in AI-enabled contexts.

The paper proceeds as follows: Section 2 details the review methodology; Section 3 synthesizes the findings; Section 4 critiques open challenges; and Section 5 concludes.

2. Research Methodology

Scientific rigor demands replicability. Consequently, this Systematic Literature Review (SLR) adheres strictly to the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) 2020 guidelines [26]. We designed the protocol to isolate and synthesize empirical evidence at the precise nexus of XAI, Cybersecurity, Human Factors, and Edge Computing, filtering out noise to focus on the signal. A completed PRISMA 2020 checklist is provided

as Supplementary Material (Table S1). This review was not pre-registered on PROSPERO or any other prospective register.

2.1. Search Strategy and Data Sources

We executed the literature search in **November 2025**, scoping the observation window from **January 2014 to November 2025**. We selected this decade-long aperture to map the complete evolutionary arc: from the nascent application of Deep Learning to the arrival of Generative and Explainable AI. However, as illustrated in Figure 1, a clear temporal pattern emerged during the identification phase. While the search spans eleven years, the density of relevant findings is heavily skewed toward the post-2021 era. It is worth mentioning the public release of GPT3.5 in November 2022 and the wide adoption starting from the following 2023.

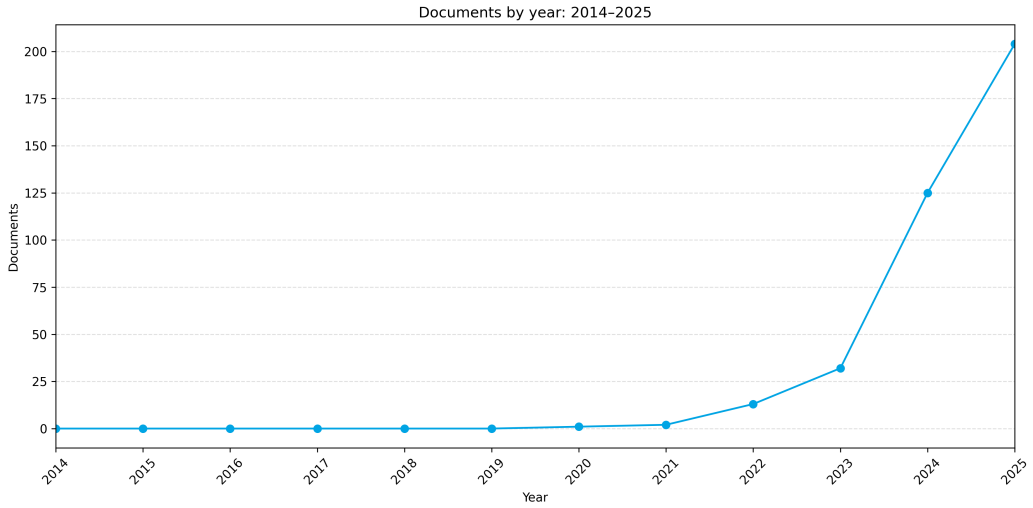

Figure 1. Yearly distribution of scientific publications included in the initial screening (2014–2025). Data from Scopus (search performed November 2025). The trend highlights a sharp increase in research activity following the mainstream adoption of GenAI technologies.

To mitigate indexing bias and ensure longitudinal depth, we queried two complementary databases:

1. *Scopus:* Selected for its rigorous indexing of high-impact Computer Science venues (including IEEE, ACM, and Springer).
2. *OpenAlex:* Included to capture high-velocity pre-prints and open-access repositories, ensuring coverage of recent developments often sequestered behind traditional paywalls.

2.2. Search Query Construction

The search strategy was designed to mandate an intersection between four conceptual domains: (1) Explainability, (2) Security context, (3) Human-centricity, and (4) Deployment constraints. The final boolean syntax required the simultaneous presence of elements from each cluster:

> *(“Explainable AI” OR “XAI” OR “explainability”) AND (“cybersecurity” OR “SIEM” OR “security information and event management”) AND (“human centered” OR “Human-in-the-Loop” OR “user behavior” OR “alert fatigue”) AND (“edge computing” OR “on-device” OR “local model” OR “privacy-preserving”)*

The query was executed on the **Title, Abstract, and Keywords** fields (`TITLE-ABS-KEY` operator in Scopus; equivalent indexed fields in OpenAlex). Search performed on 20 November 2025.

The decision to mandate the *simultaneous* presence of all four conceptual domains in the boolean query reflects a deliberate methodological trade-off between *precision* and *recall* [27,28]. A conjunctive four-domain constraint maximizes precision, ensuring that every retrieved record genuinely operates at the intersection under investigation, thereby reducing noise in the synthesis. This comes at the cost of recall: studies addressing, for instance, XAI for cybersecurity without an explicit edge or human-factor dimension are systematically excluded. We consciously accept this trade-off, as our research objective is not to produce an exhaustive census of XAI techniques in security, but to map the *specific operational nexus* where explainability, human cognition, and deployment constraints converge—a niche that existing broader surveys (e.g., [14,24]) do not address in their full interdisciplinary depth, precisely because they isolate algorithmic performance from human and deployment factors.

2.3. Inclusion and Exclusion Criteria

We applied rigorous filtering criteria—Inclusion Criteria (IC) and Exclusion Criteria (EC), detailed in Table 1—to the initial corpus. Priority was given to peer-reviewed articles providing empirical evidence or novel architectural frameworks. Conversely, theoretical position papers lacking specific security applications or human-centric validation were systematically excluded. The details of our Inclusion and Exclusion criteria are presented below.

Table 1. Inclusion and exclusion criteria (IC/EC).

Ref	Inclusion Criteria	Exclusion Criteria
IC1	Published between 2014 and 2025.	Pre-2014 publications.
IC2	Q1–Q2 Journals and top-tier conference proceedings.	Grey literature (blogs, tutorials, non-technical policy).
IC3	Explicit focus on Cybersecurity or Security Information and Event Management (SIEM) operations.	General AI applications in non-security domains.
IC4	Integration of XAI with Human Factors.	Pure ML optimization lacking human-centric context.
IC5	Validation via real-world datasets OR novel conceptual frameworks.	Studies relying exclusively on synthetic/toy experiments.
IC6	Manuscripts written in English.	Non-English manuscripts.

2.4. Study Selection Process

The selection workflow was managed using *Zotero (v. 7.0.30)* for reference management and *Rayyan* web application [29] for the blind screening phase. As illustrated in the PRISMA flow diagram (Figure 2), the process followed a standard four-stage funnel:

1. **Identification:** Deduplication of results yielded 385 unique records.
2. **Screening (Blind Mode):** Researchers independently screened titles and abstracts to eliminate selection bias. Disagreements were resolved via consensus, reducing the pool to 36 candidate papers.
3. **Eligibility (Full-Text):** Full-text versions were assessed against the IC/EC. Nine papers were excluded due to insufficient technical depth, weak validation methodologies, or retraction.
4. **Inclusion:** The final synthesis corpus comprises 27 selected studies.

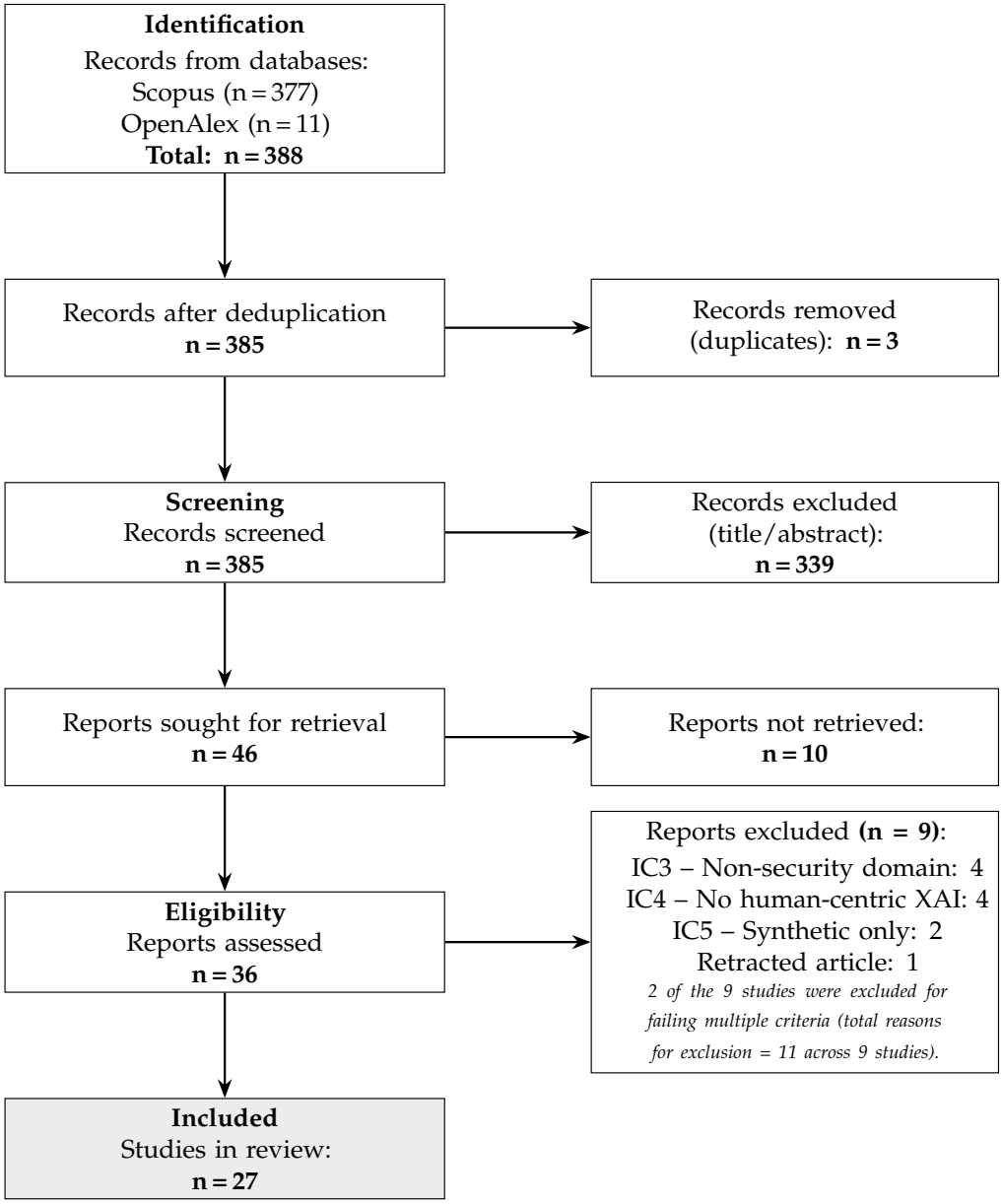

Figure 2. PRISMA 2020 Flow Diagram. Records identified from Scopus and OpenAlex ($n = 388$); after deduplication and four-stage screening, 27 studies were included in the final synthesis. Exclusion reasons at eligibility stage are mapped to criteria IC3–IC5 (see Table 1).

To ensure the reliability of the selection process, two independent reviewers performed a blind screening of the initial 385 records using the Rayyan platform. The inter-reviewer agreement was **97.7%**, with a **Cohen's Kappa of 0.86**, indicating 'almost perfect' agreement according to standard biometric criteria. Disagreements ($n = 9$) were resolved through consensus sessions involving a third senior researcher, yielding the final corpus of 27 studies. Specifically, to ensure reproducibility, ambiguous studies that spanned multiple conceptual domains within our taxonomy were classified by applying a strict priority rule: the *Cognitive Ergonomics* dimension (i.e., how the explanation is presented to and validated by the human operator) was assigned hierarchical precedence over the underlying algorithmic mechanism. For instance, studies proposing hybrid post-hoc feature attribution, but utilizing an LLM for the final user interface, were systematically categorized under 'Generative Interpretability'.

Following the finalization of the included corpus, the analysis shifted from study selection to systematic information synthesis. Subsequently, a standardized data extraction

framework was applied to this filtered corpus. We manually extracted and coded information across six primary variables: (1) Operational context and scope, (2) AI and XAI architectures utilized, (3) Data provenance and dataset realism, (4) Presence and type of empirical human validation, (5) Quantitative and qualitative evaluation metrics, and (6) Explicitly identified research gaps. Reference management was conducted via Zotero. The extracted data was then synthesized and tabulated to facilitate the cross-study comparisons presented in Tables 2 and 3.

Table 2. Summary of selected studies. Studies are categorized by operational context, AI approach, data source, and level of human validation.

Ref. & Year	Context & Scope	AI/XAI Approach	Dataset	Human Validation/Factor
Palma et al. [17] (2025)	Log Analysis/Vulnerability Detection (SME focus)	LLM-generated Explanations (Natural Language) compared with SHAP/LIME. Prompt Engineering and structured Parsing.	Real (Proprietary dataset of corporate logs, 1317 records, not publicly available).	Human-in-the-Loop (Feedback loop for refinement), Qualitative analysis of the stability of explanations.
Akshya et al. [30] (2025)	NIDS in Autonomous Vehicles & IoT (UAV, V2X)	SHAP (Kernel SHAP) for Feature Importance and decision transparency.	Mixed: Synthetic (SUMO/OpenStreetMap) + UNSW-NB15 (Benchmark).	N/A (Technical evaluation of XAI via plots and feature ranking, no user study).
Alansary et al. [1] (2025)	Zero-Day Attacks Detection/NIDS/IoT/IoMT	N/A (Review). Discusses XAI (SHAP, LIME) as a future necessity and to mitigate Poisoning in FL.	N/A (Review). Analyzes CICIDS2017, UNSW-NB15, NSL-KDD.	N/A (Mentions Human–AI collaboration as future work).
Casino [18] (2025)	Cognitive Security (Intersects Cybersecurity, Digital Forensics, HCI)	Conceptual Framework. Redefines Cognitive Security using Learning Theories, Agentic AI, LLMs, and Ethics/XAI integration.	N/A (Conceptual paper and Taxonomy).	Theoretical analysis of Human–Computer Interaction (HCI), Cognitive Biases, and Sociocultural factors.
Gupta et al. [31] (2025)	Smart Cities Cybersecurity (IDS, Malware, Fraud, Traffic Analysis)	N/A (Review). Reviews LIME, SHAP, Anchors, Counterfactuals for Smart Cities.	N/A (Review). Cites historical datasets like KDD99, Drebin.	Conceptual analysis on "Public Trust", "Fairness" and "Human-Centric AI".
Alketbi & Mehmood [32] (2025)	Insider Threat Detection (ITD)/SOC	N/A (Survey). Reviews SHAP, LIME, RuleFit, TreeLIME, DeepLIFT, Grad-CAM, etc.	N/A (Survey). Analyzes datasets like CERT (synthetic), CICIDS2017, UNSW-NB15.	N/A (Theoretical discussion on Cognitive Load and Trust).
Phanireddy [5] (2021)	Identity & Access Management (IAM)/Zero Trust	None (Mentions the "Black Box" problem but no XAI solutions).	N/A (Conceptual review with generic case studies).	N/A (Theoretically discusses User Acceptance and Privacy).
Abuadbba et al. [16] (2025)	Red Teaming (Offensive) & Blue Teaming (SOC/Defensive)	N/A (Position Paper). Discusses the need for interpretability tools to mitigate LLM risks.	N/A (Theoretical analysis based on MITRE and NIST frameworks). Criticizes the lack of realistic benchmarks.	Human-in-the-Loop risk analysis: Over-reliance, De-skilling, need for human oversight.
Kolse [21] (2025)	SOC (Alert Triage & Prioritization/SIEM integration)	XAI as a conceptual component of the FACPF framework (no specific algorithm implemented).	N/A (Conceptual Paper, no empirical data validation).	Feedback Loop via Reinforcement Learning (theoretical) to improve analyst trust.
Zhang et al. [14] (2022)	General Cyber Security (NIDS, Malware, Spam, Phishing, Smart Grid, Healthcare, etc.)	N/A (Survey). Reviews LIME, SHAP, Saliency Maps, Grad-CAM, Counterfactuals, etc.	N/A (Survey). Lists common datasets (NSL-KDD, CICIDS, Enron, etc.) used in literature.	N/A (Theoretical discussion on HCI and Human-Centered AI, no user study).

Table 2. *Cont.*

Ref. & Year	Context & Scope	AI/XAI Approach	Dataset	Human Validation/Factor
Moustafa et al. [33] (2023)	NIDS in IoT & Fog Computing Environments	N/A (Survey). Reviews LIME, SHAP, PDP, ALE, ICE, etc. Proposes XAI architecture on Fog Computing.	N/A (Survey). Criticizes the widespread use of outdated and imbalanced datasets in literature.	Conceptual "User-Centric" Framework (Figure 3) mapping XAI benefits to 5 stakeholder types (Experts, Managers, Regulators, etc.).
Bhusal et al. [34] (2023)	Multi-domain: Log Anomaly Detection (SOC/SIEM), Malware Analysis, Adversarial Image Detection	Comparative evaluation of 9 methods (LIME, SHAP, Integrated Gradient, DeepLIFT, GradientShap, etc.).	Real/Benchmark: HDFS (System Logs), Mimicus (PDF Malware), MNIST/CIFAR-10 (Adversarial Images).	Qualitative interviews with domain experts (SOC analyst, Malware researcher).
Pawlicki et al. [35] (2024)	Intrusion Detection (NIDS)	Integrated Gradients (IG), SHAP, LIME, and ANCHORS.	Real-world datasets: CIC IoT dataset 2023 and CSE-CIC-IDS2018 Intrusion Detection Evaluation Dataset.	None. The paper is a systematic review and a computational/objective experimental evaluation of XAI metrics.
Dietz et al. [36] (2024)	NIDS (Enterprise Networks/IT Operations & SOC)	N/A (Survey & Framework). Suggests: SHAP, LIME, Decision Trees, GAM, EBM.	Qualitative (Interviews with experts) + Systematic analysis of 166 papers (Literature Dataset).	Proto-Personas (6 user profiles derived from real scenarios); User-Centric Analysis.
Park et al. [23] (2025)	Maritime SOC/Log Summary	No explicit XAI techniques like SHAP or LIME are used, but the system proposed is focused on explaining the log.	Real-world. The experiment uses network alert logs collected by deploying Suricata on an actual ship's internal network.	The system integrates an Adaptive Learning Strategy that continuously refines recommendations based on user feedback.
Dehghani et al. [37] (2025)	Trustworthy AI Framework (Biometric, Cybersecurity)	N/A. Mentions self-explainable models and techniques like attribution maps and counterfactuals.	No novel dataset is used or presented as the paper is a review.	None employed in this paper, as it is a review.
Tomičić et al. [38] (2025)	SOC, enhancing Incident Response, SIEM and SOAR	The paper is a review and roadmap; it does not use a specific XAI approach in an experiment.	The paper is an Overview. It does cite other works that use datasets like the IoT23 dataset.	None employed in this paper.
Ahi & Valizadeh [15] (2025)	SOC, enhancing Incident Response, SIEM, and SOAR	It is a review and roadmap; it does not use a specific XAI approach in an experiment.	No novel dataset is used or presented as the paper is a review.	None employed in this paper.
Sayeed et al. [39] (2025)	Cybersecurity Awareness and Social Engineering Attack Prevention	Explainable AI (XAI) techniques are incorporated but not described in detail.	Curated/Real-world text corpus from Hugging Face was used to fine-tune the LLaMA 3.1 AI chatbot model.	The chatbot design prioritizes a human-centered approach (Community Surveys).
Safavi et al. [8] (2025)	Mainly for SOCs, also applies to IDS	SHAP, LIME, and Grad-CAM (Gradient-weighted Class Activation Mapping).	N/A. The paper proposes a conceptual framework.	None. The paper proposes Human Factor concepts as trust mechanisms (HITL supervision).
Barbieri et al. [40] (2025)	CTI, SIEM systems (transforming them with LLM-driven knowledge extraction)	The paper is a survey and synthesis.	N/A. Paper is a literature review (422 manually labeled papers).	Human-in-the-Loop Validation (Human curation and auditing of the LLM's classification output).
Khayat et al. [22] (2025)	SOC, SIEM, IDS, SOAR, APT Detection	The paper is a systematic review and does not use an XAI approach itself.	As it is a review, there are many: SOC, SIEM, IDS, SOAR, APT Detection.	None. The paper is a systematic literature review.
Binbeshr et al. [19] (2025)	SOC, SIEM and Network Security	The paper is a Systematic Literature Review (SLR) and does not use an XAI approach itself.	N/A. Paper is a literature review.	None. The paper is a systematic literature review.

Table 2. *Cont.*

Ref. & Year	Context & Scope	AI/XAI Approach	Dataset	Human Validation/Factor
Swetha et al. [7] (2025)	IDS, Malware analysis, Phishing, Threat intelligence	The paper is a review and proposes a new methodology, Q-PEARL, which makes use of LIME and SHAP.	N/A (Literature Review).	None. The paper did not employ a Human Factor methodology.
Jaigirdar et al. [13] (2025)	IoT-Health Architecture	Not a traditional XAI paper. The approach uses Security Metadata to create a security-aware provenance graph.	Simulated Use Cases, Table-Top Exercises (Qualitative Expert Study).	Table-Top Exercises (Qualitative Expert Study).
Yilmaz et al. [2] (2024)	Mainly on Insider Threat Detection	None. The paper is an overview that identifies Explainable AI (XAI) as a crucial gap.	N/A. The paper is an overview and does not use a primary dataset.	None. The paper is a study and overview.
Pawlicki et al. [24] (2024)	Network Intrusion Detection Systems (NIDS)	The paper is a systematic mapping study that reviews and categorizes numerous XAI techniques (ANCHORS, LIME, etc.).	N/A (Systematic Mapping Study).	None. The paper is a systematic mapping study.

Table 3. Detailed Analysis of Evaluation Metrics and Critical Research Gaps identified in the Selected Studies (N = 27). **QA Score Legend (V/M/D/X):** V = Venue; M = Methodology; D = Dataset; X = XAI/Human Factors (Max 6). Bold numbers indicate the total QA score.

Ref. & Year	Metrics Used	Identified Research Gaps	QA Score
Palma et al. [17]	F1, Precision, Recall, FPR, FNR, Entropy	(1) False positives causing Alert Fatigue; (2) fragility of manual Feature Engineering; (3) high computational costs for SMEs; (4) LLM hallucination risk; (5) lack of quantitative feature attribution in LLMs.	**19** (5/5/4/5)
Akshya et al. [30]	Acc, Latency, Energy, Interpretability	(1) Lack of adaptive real-time detection; (2) opacity of DL models (Black-box); (3) energy/computational efficiency on Edge; (4) vulnerability to adversarial attacks.	**19** (6/6/5/2)
Dietz et al. [36]	Lit. metrics (% w/o code/FPR)	(1) "Black-box" models undermining trust (86% of papers); (2) absence of GUI/Usability (95%); (3) poor reproducibility (No code/data); (4) outdated datasets; (5) privacy ignored; (6) false Positive Rate too high for practical use.	**23** (6/6/5/6)
Alansary et al. [1]	Acc, F1, MCC, ROC-AUC	(1) Data privacy and security (centralized vs FL); (2) vulnerability to Poisoning Attacks in FL; (3) lack of transparency (Black-box); (4) robustness against adversarial examples; (5) latency and computational costs.	**19** (6/6/4/3)
Casino [18]	N/A (Theoretical). Metrics: Consistency, Accuracy, Completeness, Auditability.	(1) Approaches are too technocentric; (2) ethical/explainability aspects barely highlighted; (3) risks of cognitive biases; (4) lack of benchmarks.	**19** (6/6/1/6)
Gupta et al. [31]	N/A (Qualitative)	(1) Lack of public trust in black-box models; (2) ethical risks and hidden Biases; (3) need for XAI standards in Public Administration; (4) XAI latency in critical real-time contexts.	**14** (5/5/1/3)

Table 3. *Cont.*

Ref. & Year	Metrics Used	Identified Research Gaps	QA Score
Park et al. [23]	MTTR, ROUGE, BLEU	(1) Expanding the dataset and applying data augmentation; (2) refining the feedback integration process with automated validation mechanisms; (3) exploring reinforcement learning-based optimization; (4) investigating RAG for external threat intelligence.	**20** (5/6/6/3)
Alketbi & Mehmood [32]	N/A (Theoretical)	(1) High false positives and Alert Fatigue; (2) lack of semantic/psychometric context; (3) scalability of XAI in real-time; (4) lack of XAI evaluation standards; (5) risks of Adversarial XAI.	**17** (6/6/1/4)
Phanireddy [5]	FP, Risk Scores	(1) Privacy and ethical issues in behavioral data collection; (2) high false positives impacting user experience; (3) vulnerability to adversarial attacks; (4) opacity of future Deep Learning models.	**13** (4/5/1/3)
Abuadbba et al. [16]	N/A (Risk Analysis)	(1) Hallucinations and reasoning limits of LLMs; (2) context loss in long tasks; (3) dual-Use risks (advantage for attackers); (4) erosion of human skills; (5) privacy and security of LLM APIs.	**14** (3/6/1/4)
Kolse [21]	MTTD, MTTR, FPR	(1) High false positive rates and analyst burnout; (2) lack of dynamic context in prioritization; (3) privacy issues in Threat Intel sharing; (4) lack of transparency (Trust) in current AI systems.	**11** (1/5/1/4)
Zhang et al. [14]	N/A (Survey)	(1) Vulnerability of XAI to adversarial attacks; (2) lack of evaluation standards for explanations; (3) privacy risks (data leakage via explanations); (4) accuracy/Interpretability trade-off.	**18** (6/6/4/2)
Moustafa et al. [33]	Acc, FPR	(1) IoT datasets are unrepresentative/outdated; (2) high computational cost of XAI on Edge/IoT; (3) explanations are too technical for non-experts; (4) lack of XAI standards and metrics; (5) instability of existing XAI methods.	**19** (6/6/4/3)
Bhusal et al. [34]	Faithfulness, Stability, Qualitative	(1) Inconsistency between different XAI methods; (2) low usability and "actionability" for experts; (3) vulnerability of explanations to attacks and privacy issues; (4) need for mixed evaluation (quantitative + qualitative).	**23** (6/6/5/6)
Pawlicki et al. [35]	MPRT, Complexity, Monotonicity	(1) Lack of consensus on fundamental properties and definitions of XAI metrics; (2) metric duplication and overlapping concepts; (3) limitations in applicability and compatibility of many metrics with diverse models and data types; (4) need for standardization and a unified framework.	**19** (6/6/6/1)

Table 3. *Cont.*

Ref. & Year	Metrics Used	Identified Research Gaps	QA Score
Dehghani et al. [37]	Fairness (DI, Equal Odds)	(1) Lack of standardization in ethical development and Trustworthy AI; (2) data privacy (especially omission of critical attributes like age, sex/gender...); (3) regulation and policy development; (4) sustainable AI development; (5) governance; (6) differences in stakeholders' perspectives.	**16** (4/6/0/6)
Tomičić et al. [38]	Acc (Cited from lit.)	(1) Data Quality (accuracy and relevance of collected data); (2) model Interpretability (lack of transparency in complex AI models); (3) adversarial attacks (AI models themselves are targets); (4) integration complexity; (5) data privacy and legal compliance.	**16** (5/6/0/5)
Ahi & Valizadeh [15]	N/A (Survey)	(1) Data quality (accuracy and relevance of collected data); (2) model interpretability; (3) adversarial attacks; (4) integration complexity; (5) data privacy and legal compliance.	**16** (5/6/0/5)
Sayeed et al. [39]	Acc, Precision, Recall, F1	(1) Developing a fake profile detection system; (2) developing a fully functional mobile application from the ElderConnect API for greater accessibility; (3) building partnerships with academic, government, and cybersecurity organizations to expand reach.	**18** (5/5/3/5)
Safavi et al. [8]	N/A (Framework)	(1) The inherent trade-off between the computational cost of generating explanations and the need for real-time performance; (2) the challenge of seamless integration of the diverse explainability, security, and trust techniques; (3) the need to develop a functional prototype for rigorous validation; (4) the necessity for ongoing research to ensure the framework aligns with evolving regulatory landscapes.	**14** (4/5/1/4)
Barbieri et al. [40]	Classification Accuracy	(1) Explainability of LLM-based CTI Reasoning; (2) federated and privacy-preserving inference for CTI; (3) LLM robustness against adversarial CTI inputs; (4) automated alignment with CTI standards (MITRE, NIST); (5) addressing the problem of LLM "hallucinations" in CTI workflows.	**21** (6/6/5/4)
Khayat et al. [22]	N/A (SLR)	(1) The lack of a holistic and integrated approach that fully exploits AI and ML across all SOC components; (2) lacked a broader SOC architectural framework.	**17** (6/6/1/4)
Binbeshr et al. [19]	N/A (SLR)	(1) Data quality and availability; (2) model interpretability; (3) integration with legacy systems; (4) evolving cyber threats; (5) researching privacy-preserving AI methods.	**21** (6/6/5/4)
Swetha et al. [7]	N/A (Review)	(1) Data scarcity; (2) model interpretability/explainability; (3) vulnerability to adversarial attacks.	**15** (5/5/2/3)

Table 3. *Cont.*

Ref. & Year	Metrics Used	Identified Research Gaps	QA Score
Jaigirdar et al. [13]	Qualitative (Table-top)	(1) Lack of transparency and system security awareness for end-users in IoT-Health systems; (2) the challenge of scalability when applying the security metadata approach to large-scale IoT applications.	**19** (6/5/3/5)
Yilmaz et al. [2]	N/A (Overview)	(1) Insufficient datasets; (2) static access control; (3) collusion attacks; (4) need for research into explainable AI (XAI).	**14** (5/5/1/3)
Pawlicki et al. [24]	N/A (Mapping)	(1) User-centred approaches; (2) domain-specific XAI metrics; (3) interdisciplinary research; (4) context awareness in XAI; (5) interactive and hybrid approaches; (6) bias-free XAI; (7) transparency; (8) ethical code.	**17** (6/6/4/1)

Unlike traditional SLRs that focus solely on empirical evidence, this study adopts a *Systematic Mapping* approach. We included high-impact surveys and conceptual frameworks alongside experimental studies to capture the rapid evolution of definitions and taxonomies in the Cognitive Security domain.

As required by PRISMA 2020 guidelines, Table 4 provides full transparency on the nine studies excluded at the full-text eligibility stage, with explicit mapping to the inclusion/exclusion criteria defined in Table 1.

Table 4 reports the nine studies excluded during the full-text eligibility assessment, along with the specific exclusion criterion applied in each case.

Table 4. Studies excluded at full-text screening stage with exclusion rationale.

Reference	Journal/Venue	Year	Exclusion Reason
Li et al. [20]—Generative AI for Self-Adaptive Systems: State of the Art and Research Roadmap	ACM Transactions on Autonomous and Adaptive Systems	2024	AI applied to non-security domain (Self-Adaptive Systems); no cybersecurity focus (IC3)
Saarela & Podgorelec [25]—Recent Applications of Explainable AI (XAI): A Systematic Literature Review	Applied Sciences	2024	XAI survey without cybersecurity focus; no security-specific application (IC3)
Vultureanu-Albişi et al. [12]—eXING-IoT conceptual framework for explainability integration in next-generation IoT	Connection Science	2025	No real dataset; synthetic experiments only; non-security domain (IC5, IC3)
Peña-Cáceres et al. [41]—Research Trends and Networks in Self-Explaining Autonomous Systems: A Bibliometric Study	Computers, Materials and Continua	2025	Bibliometric study; no real dataset; non-security domain (IC3, IC5)
Sharma & Lashkari [42]—Hybrid Attention-Enhanced Explainable Model for Encrypted Traffic Detection	International Journal of Information Security	2025	Pure ML optimization; no human-centric context or validation (IC4)
Zivkovic et al. [43]—Addressing Smart City Security: Machine and Deep Learning Methodology	Cluster Computing	2025	Pure ML optimization; no human-centric XAI integration (IC4)

Table 4. *Cont.*

Reference	Journal/Venue	Year	Exclusion Reason
Sharma et al. [6]—Adaptive Threat Detection: Leveraging Machine Learning for Real-Time Cybersecurity	International Conference on Advances in Computing, Communication and Materials (ICACCM)	2024	No XAI implementation; pure ML without human-centric context (IC4)
Benzaïd et al. [44]—FortisEDoS: A Deep Transfer Learning-Empowered EDoS Detection Framework	IEEE Transactions on Dependable and Secure Computing	2024	Pure ML/DL; no human-centric XAI validation; highly technical domain-specific scope (IC4)
Gwassi et al.—Cyber-XAI-Block: An End-to-End Cyber Threat Detection & FL-Based Risk Assessment Framework	Multimedia Tools and Applications	2025	Retracted article

2.5. Justification of Corpus Size

The final synthesis corpus of 27 studies requires contextual interpretation to assess its representativeness and adequacy. It is essential to distinguish between two methodologically distinct review typologies: a *mono-domain meta-analysis*, which aggregates large volumes of homogeneous studies to derive statistical effect sizes, and a *systematic mapping study* at the intersection of multiple specialized domains [27]. The present work belongs unambiguously to the latter category: our search query mandates the *simultaneous* presence of four conceptual domains—Explainable AI, Cybersecurity operations, Human-Centered Design, and Edge Computing constraints. This intersection is, by epistemological design, a narrowly scoped domain.

This structural constraint is a deliberate scoping decision that reflects the nascent state of this cross-domain field. Comparable multi-domain mapping studies in adjacent areas report similarly constrained corpus sizes: for instance, Pawlicki et al. [24] conducted a systematic mapping of XAI in deep learning for cybersecurity, retaining a highly focused final set of 19 primary studies, while Binbeshr et al. [19] synthesized the Cognitive SOC literature with a final corpus of 38 studies. Across the systematic review literature in software engineering, Kitchenham and Charters [27] and Petersen et al. [28] explicitly acknowledge that focused mapping studies targeting emerging or interdisciplinary intersections routinely yield small primary corpora, as the research community itself has not yet produced a large body of work addressing all required dimensions simultaneously.

A corpus of this size, when composed exclusively of studies meeting stringent multi-domain inclusion criteria, provides an interesting evidence density to identify systemic patterns—such as the validation gap and the dataset realism crisis—without the noise introduced by peripherally relevant studies. Therefore, while the 27-study corpus limits statistical generalizability in the meta-analytic sense, it accurately maps the current research frontier at this four-domain intersection and constitutes an adequate empirical foundation for the taxonomy and research agenda proposed in this work.

2.6. Quality Assessment

To ensure the robustness of our synthesis, the included studies underwent a Quality Assessment (QA) protocol adapted from Kitchenham's guidelines [27]. Each study was scored (on a scale of 0–24) across four dimensions: (i) Venue Quality, (ii) Methodological Rigor, (iii) Dataset Realism, and (iv) Depth of XAI/Human validation.

Table 5 presents the detailed scoring rubric applied to each dimension. Each dimension is scored from 0 to 6 using a three-level ordinal scale, where 0 indicates absence or critical deficiency, 3 indicates partial satisfaction, and 6 indicates full satisfaction of the criterion. The rubric was defined prior to the screening phase and applied independently by both reviewers to ensure consistency.

Intermediate scores (1–2, 4–5) were assigned when a study exhibited characteristics between the defined anchor points. All borderline cases were resolved through consensus sessions between the two reviewers, with adjudication by a third senior researcher when agreement could not be reached bilaterally.

Table 5. Quality assessment scoring rubric. Each dimension is scored on a 0–6 ordinal scale; intermediate scores (1–2, 4–5) were assigned when a study exhibited characteristics between the defined anchor points, with the final score determined by reviewer consensus. Bold text in the first column is used to highlight the evaluation dimensions for readability.

Dimension	Score 0 (Absent/Critical Deficiency)	Score 3 (Partial Satisfaction)	Score 6 (Full Satisfaction)
(i) Venue Quality	Grey literature, unranked workshop, or non-peer-reviewed outlet.	Q3–Q4 journal or regional/national conference proceedings.	Q1–Q2 journal or top-tier international conference (e.g., IEEE, ACM, Springer venues).
(ii) Methodological Rigor	No explicit methodology; purely anecdotal or descriptive without reproducible procedures.	Methodology partially described; key steps present but insufficient for full replication.	Clear, fully described, and reproducible methodology with explicit experimental design and evaluation protocol.
(iii) Dataset Realism	Synthetic dataset only (e.g., NSL-KDD derivatives, simulated traffic); no grounding in real-world threat scenarios.	Mixed provenance: partially real-world data combined with legacy or synthetic benchmarks.	Real-world operational data or a recent high-fidelity benchmark (post-2020) representative of current threat landscapes.
(iv) Depth of XAI/Human Validation	No XAI component or human-centric evaluation; purely algorithmic performance metrics reported.	XAI technique present but human validation is theoretical or conceptual only; no empirical user study.	Empirical human validation with domain experts, SOC analysts, or end-users; explicit measurement of Human Factors (e.g., trust, cognitive load, task performance).

Only studies exceeding the predefined quality threshold were retained for data extraction. The distribution of these scores is visualized in Figure 3, indicating that the selected literature maintains a medium-to-high quality standard. It is important to note that the scoring mechanism weights Human Factors symmetrically with Venue, Methodology, and Dataset quality. Consequently, a study can achieve a high overall score (e.g., 18–20) through algorithmic and methodological excellence, even while scoring poorly in the empirical validation of Human-in-the-Loop dynamics. This scoring artifact deliberately exposes the current state of the art: technical rigor is highly rewarded, while human-centric validation remains neglected.

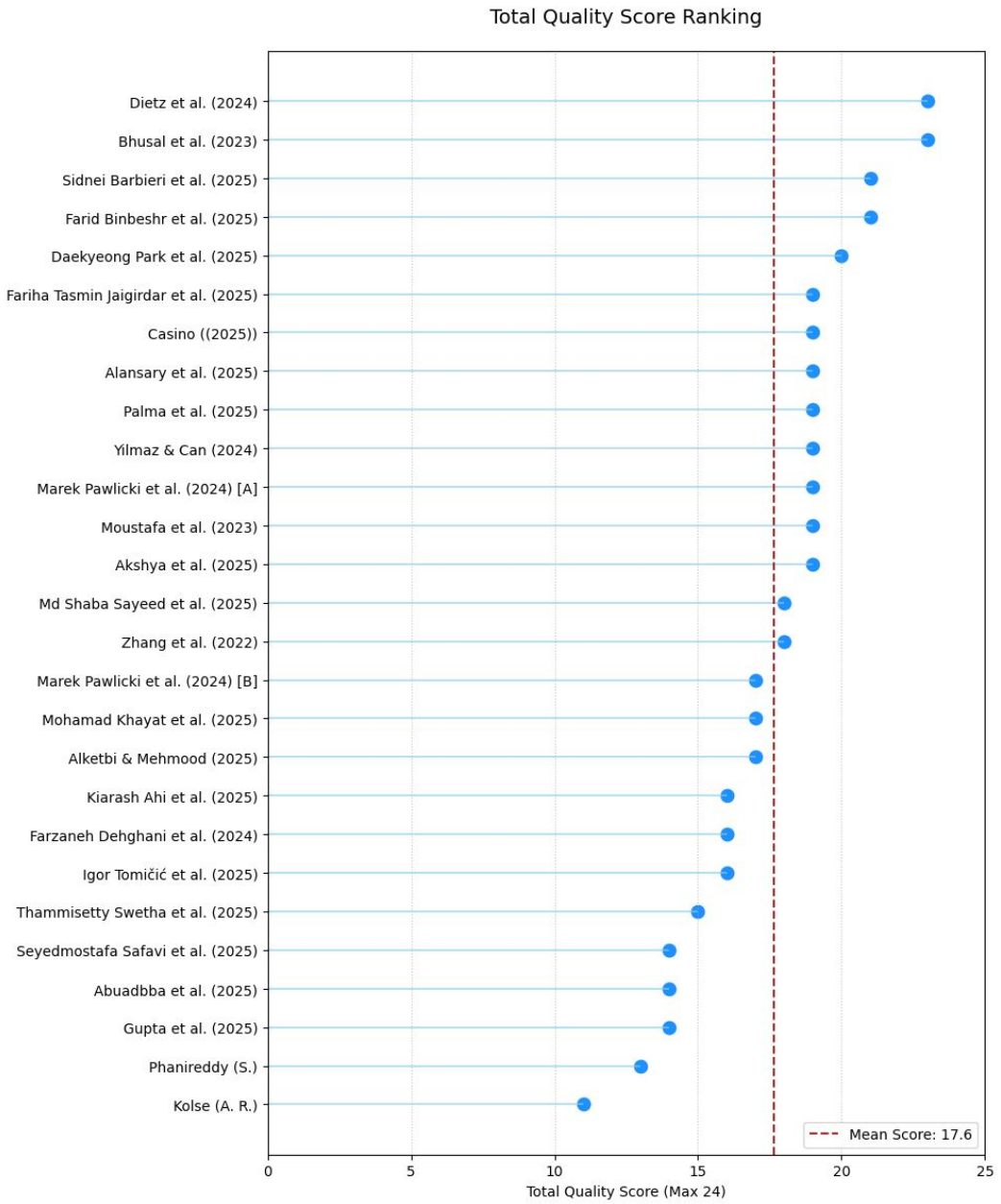

Figure 3. Total quality score ranking. The chart demonstrates that the final corpus meets the rigorous criteria required for a systematic synthesis, with most studies clustering in the medium–high range. References for the studies evaluated in this chart: [1,2,5,7,8,13–19,21–24,30–40].

3. Results and Synthesis

The taxonomy presented operates across three explicit classificatory dimensions: (1) generative capabilities (mapping the transition from diagnostic output to narrative synthesis), (2) algorithmic transparency (categorizing XAI techniques based on feature attribution vs. attention mechanisms), and (3) cognitive ergonomics (evaluating system outputs against human trust and cognitive load). Unlike pre-existing taxonomies (e.g., [14,33,34]), which primarily classify XAI purely by mathematical formulation (local vs. global, model-agnostic vs. specific), our mapping criteria mandate the intersection of computational technique with operational SOC usability. We systematize the findings from the 27 selected studies through a multi-dimensional conceptual taxonomy (see Figure 4). This framework transcends simple categorization, mapping the functional friction between the stochastic

nature of Generative AI, the deterministic requirements of XAI validation, and the stringent cognitive demands placed on cybersecurity operators.

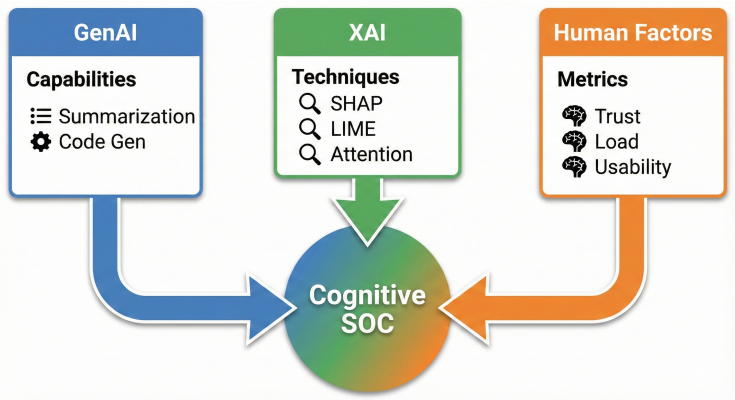

Figure 4. Proposed taxonomy of cognitive security. The schema visualizes the convergence of generative AI capabilities, XAI techniques, and Human Factors, defining the interaction flows within a modern SOC.

The proposed taxonomy can be contextualized against three prior frameworks addressing XAI classification in adjacent security domains.

Zhang et al. [14] organize XAI methods along four algorithmic dimensions: (i) *intrinsic vs. post-hoc*, (ii) *model-specific vs. model-agnostic*, (iii) *local vs. global*, and (iv) *explanation output type* (text-based, visual, model-based, or argument-based). While comprehensive in algorithmic coverage, this framework does not incorporate deployment constraints, cognitive load, or human-operator dimensions, making it unsuitable for characterizing the operational friction of a Cognitive SOC.

Moustafa et al. [33] apply the same foundational XAI taxonomy to the IoT/Edge intrusion detection domain, contextualizing it within resource-constrained deployment scenarios and discussing the trade-off between explainability and detection performance. However, no explicit human operator or cognitive dimension is introduced.

Bhusal et al. [34] represent the closest precedent to a human-centred evaluation framework: they classify post-hoc methods by granularity, supported model type, and explanation type, and evaluate them against both functionally-grounded quantitative metrics (faithfulness, continuity, max-sensitivity, sparsity) and human-grounded qualitative criteria (coherence, actionability, expertise level), validated through structured interviews with security domain experts. Despite this, their framework remains focused on evaluating existing post-hoc methods and does not model the interaction between generative explanation capabilities and analyst cognitive load under SOC operational conditions.

The present taxonomy departs from these frameworks by introducing *Cognitive Ergonomics* as an explicit third classificatory dimension, alongside algorithmic transparency and generative capabilities. Unlike Zhang et al. [14] and Moustafa et al. [33], this tripartite structure is *operational* rather than purely algorithmic: it maps the functional dependencies between what the model generates, how transparently it does so, and whether the resulting output is cognitively actionable for a human analyst under real SOC conditions. Unlike Bhusal et al. [34], it addresses generative LLM-based explanations explicitly and frames cognitive ergonomics not as an evaluation criterion but as a *classificatory dimension* of the taxonomy itself.

As illustrated in Figure 4, the three input domains converge unidirectionally into the cognitive SOC node, reflecting a deliberate design choice: GenAI capabilities (Summarization, Code Generation) and XAI techniques (SHAP, LIME, Attention) are mediated by Human Factor metrics (Trust, Load, Usability) before reaching the operator. This flow exposes a critical gap identified in the corpus: while 63% of studies claim human-centric relevance, only 22% incorporate empirical validation with real SOC operators [21,34,36]. The Human Factors node is thus the least empirically grounded of the three—visually central in the taxonomy, but operationally hollow in current practice.

3.1. Descriptive Statistics

The longitudinal distribution of the selected literature reveals a significant surge in research interest at the intersection of XAI and Generative AI within the cybersecurity domain. As illustrated in Figure 1, over 60% of the identified corpus was published during the 2024–2025 biennium. This growth trajectory correlates with the widespread adoption of Large Language Models (LLMs) following the public release of GPT-3.5 in late 2022 and the consequent urgency to interpret their reasoning processes in mission-critical security operations. These findings mirror broader bibliometric trends in Self-Explaining Autonomous Systems (SEAS), which highlight a consolidated research front focused on transparency in autonomous decision-making [41].

A granular analysis of the experimental methodologies reveals two critical systemic trends. First, regarding *data provenance*, 63% of the analyzed studies rely on synthetic or legacy datasets (e.g., NSL-KDD derivatives), whereas only 37% utilize real-world proprietary logs or modern high-fidelity traffic captures. This discrepancy is further underscored in the Quality Assessment Heatmap (Figure 5), where "Dataset Quality" emerges as the most prominent methodological gap.

Figure 5 makes this asymmetry visually explicit: the red regions (scores 0–1) concentrate almost exclusively in the *Dataset Quality* and *Human Factors* columns, while *Venue Quality* and *Methodological Rigor* remain consistently green across the corpus. Notably, several studies achieving maximum scores (6/6) on Venue and Methodology—such as Pawlicki et al. (2024) and Zhang et al. (2022)—score as low as 1–2 on Human Factors, confirming that publication quality and human-centric validation are structurally decoupled [14,35]. This pattern indicates that the gap is not a function of research quality per se, but a structural blind spot of the field as a whole—precisely what the scoring artifact described in Section 2.6 was designed to expose.

Second, despite the prevailing debate on "human-centricity", empirical validation with end-users remains quite scarce. Only 22% of the papers incorporated human participants (e.g., professional SOC analysts) in their evaluation frameworks. The vast majority of current research relies exclusively on algorithmic proxies (such as fidelity or stability metrics) to quantify explainability.

The consequences of this systemic lack of empirical user testing extend directly to the operational reliability of the proposed models. This "validation gap" introduces significant risks, as illustrated in the Risk of Bias Summary (Figure 6). Both Dataset Quality and Human-in-the-Loop (HITL) dynamics are identified as primary vectors of bias. Without high-fidelity input data and rigorous monitoring of human–AI interaction, there is a high probability that XAI outputs will perpetuate embedded biases, leading to unreliable decision support in live environments.

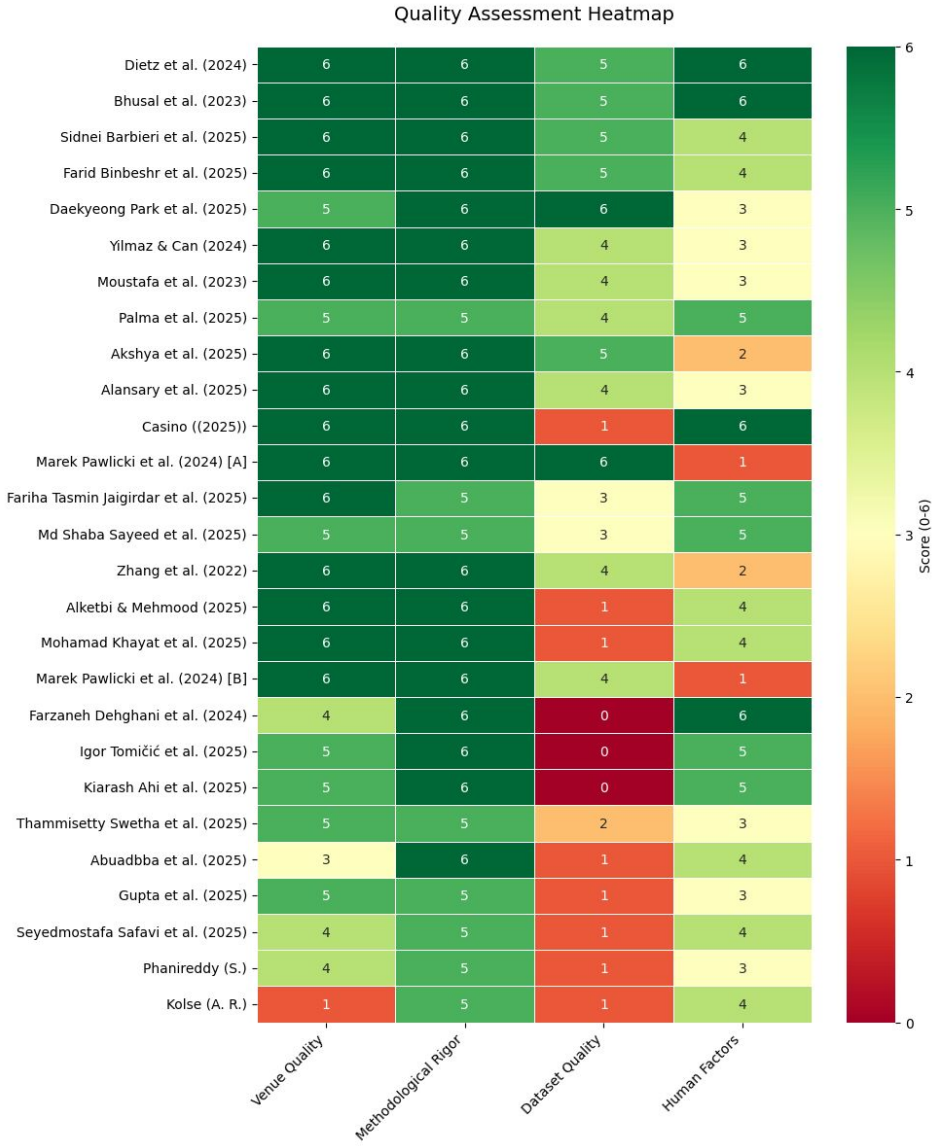

Figure 5. Quality assessment heatmap. The visualization highlights a systemic reliance on low-quality or poorly documented datasets that lack grounding in contemporary real-world threat scenarios. References for the studies evaluated in this chart: [1,2,5,7,8,13–19,21–24,30–40].

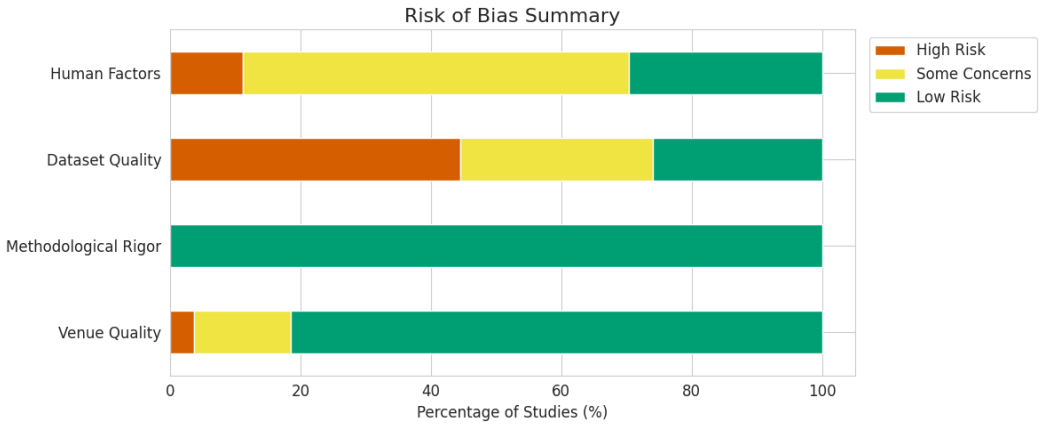

Figure 6. Risk of bias summary. Significant risk concentrations are observed in data provenance and the lack of empirical validation for human–AI interaction dynamics.

Figure 6 quantifies the bias distribution across the four quality dimensions. *Dataset Quality* emerges as the most critical risk vector, with approximately 45% of studies rated High Risk (orange)—the largest such concentration in the corpus. *Human Factors* shows a different but equally concerning pattern: while High Risk is limited (~10%), over 60% of studies fall under Some Concerns (yellow), indicating widespread superficial treatment of human validation rather than outright neglect [21,36]. Conversely, *Methodological Rigor* is almost entirely Low Risk, confirming that technical execution is strong but systematically decoupled from the deployment and human dimensions that determine real-world reliability [1,33].

In terms of publication landscape, *IEEE Access*, *Computers & Security*, and *Neurocomputing* emerge as the most influential venues, reflecting the multidisciplinary rigor—spanning from neural architectures to operational security—demanded by this domain.

This overview ultimately reveals a critical pattern, which we term the *Maturity Paradox*: while the volume of literature on Generative AI for cybersecurity is expanding at an exponential rate, its methodological maturity remains constrained by a persistent decoupling between technical innovation and operational validity, as evidenced by the concentration of low scores in the Dataset Realism and Human Factors dimensions of the QA Heatmap (Figure 5) and the Risk of Bias Summary (Figure 6).

To understand the root cause of this maturity paradox, we must examine the methodological choices driving these studies. A cross-correlation of the Quality Assessment Heatmap and the Risk of Bias Summary reveals that this paradox is driven by a pattern we designate as *Technological Determinism*: a systematic prioritization of algorithmic complexity over deployment realism, as reflected by the observation that studies with the highest computational sophistication consistently exhibit the lowest scores in Dataset Quality and Human Factor dimensions (Figure 5). This suggests that the field is currently optimizing for the reasoning capabilities of LLMs while relying on legacy datasets (e.g., NSL-KDD derivatives) that pre-date current multi-vector attack scenarios by over a decade [33]. Consequently, this synthesis identifies a structural gap between reported algorithmic performance and operational deployability, which undermines the ecological validity of current XAI research for real-world cognitive SOC environments—understood here in the sense originally defined by Neisser [45] as the degree to which experimental conditions reflect naturalistic operational contexts, and subsequently operationalized for machine learning security systems by Arp et al. [46], who identify dataset bias and concept drift as the primary sources of this disconnect in computer security research.

3.2. RQ1: Evolution of XAI Techniques

To avoid conceptual overlap within the cognitive SOC framework, our analysis explicitly distinguishes between three primary paradigms emerging from the literature: (a) *intrinsically interpretable models* (e.g., Generalized Additive Models, Kolmogorov–Arnold Networks), which embed transparency directly into their architectural design; (b) *post-hoc explainability methods* (e.g., SHAP, LIME), which act as external surrogates to approximate the logic of pre-trained black-box models; and (c) *generative explanation systems* (e.g., LLMs), which synthesize narrative reasoning by translating technical attributions into natural language. Regarding intrinsic models, unlike traditional Multi-Layer Perceptrons (MLPs) that require external surrogates like SHAP to approximate their opaque node activations, KANs utilize learnable activation functions on the network's edges. This unique mathematical formulation allows them to be visualized and interpreted as symbolic equations [47]. Consequently, KANs bypass the inherent faithfulness vulnerabilities and explanation instability of post-hoc methods, offering a true "glass-box" alternative that maintains competitive predictive performance for critical SOC operations [48]. Our analysis isolates a clear bi-

furcation in how these implementations have evolved over time. Early methodologies (2020–2022) leaned heavily on **post-hoc feature attribution**; SHAP and LIME became the de facto standards for tabular network data [30,43].

To contextualize the technical trade-offs between these methods—spanning data type, task, computational cost, and robustness—we provide a comparative synthesis in Table 6.

A closer analysis of this landscape reveals a clear temporal evolution in how explainability is implemented. During the first identified phase (the 2020–2022 cohort), methodologies focused almost exclusively on technical 'Feature Attribution'. In this period, methods such as SHAP and LIME became the industry standard because they could point to specific variables in tabular network logs that triggered an alert. Alongside these, techniques such as Integrated Gradients [49] and Grad-CAM [50] broadened the scope of attribution to deep neural networks and visual representations, while Counterfactuals [51] introduced actionable contrastive analysis.

Table 6. Comparative analysis of XAI methods identified in the literature.

Method	Year	Data Type	Task	Comp. Cost	Robustness
SHAP	2017 [10]	Tabular, Log	Global & Local Attribution	High	Medium (Slightly Manipulable)
LIME	2016 [9]	Tabular, Text, Image	Local Attribution	Medium	Low (Local Instability)
Integrated Gradients	2017 [49]	Neural Nets, Log	Feature Attribution	Low	High (Axiomatic)
LLM-Explainer	2022–2023	Unstructured Logs	Narrative Explanation	Very High	Low (Hallucination Risk)
Grad-CAM	2017 [50]	Images, Heatmaps	Visual Explanation	Low	High (Saliency-based)
Counterfactuals	2017 [51]	Tabular	Contrastive Analysis	Medium	High (Actionable)
Intrinsic Models (e.g., KAN, GAM)	2024 [36,47]	Tabular, Time-Series	Intrinsic Explanation	Medium	High (Architectural)

While these post-hoc techniques established the foundation for model transparency, their rigid mathematical nature often proved difficult for non-technical analysts to interpret during high-speed triage. However, our analysis shows a significant paradigm shift in the most recent literature (the 2024–2025 cohort). Research has moved toward 'GenAI-generated explanations', where Large Language Models (LLMs) are used to transform abstract mathematical weights into natural language summaries. While this transition makes security logs more accessible to human operators, it introduces a critical trade-off: the dominant paradigm has shifted from feature attribution methods (e.g., SHAP, LIME), which identify contributing variables in tabular logs, toward generative explanation systems that produce natural language summaries via LLMs. This transition reduces the technical barrier for operators but introduces higher computational costs and an increased risk of unfaithful explanations [16].

This lower algorithmic robustness is a multifaceted challenge that directly impacts the operational reliability of a Security Operations Center. In the earlier cohort of research, robustness was primarily viewed through the lens of mathematical stability; methods like SHAP and LIME, while providing deep insights, often suffer from 'explanation jitter', where minor, non-malicious perturbations in network telemetry can lead to radically different and inconsistent feature attributions. This instability creates a dangerous vulnerability: if an explanation changes significantly due to slight noise in the data, it becomes difficult for an analyst to distinguish between a genuine system anomaly and an artifact of the XAI algorithm itself. Furthermore, these local surrogate models can be intention-

ally manipulated, allowing sophisticated threats to mask their presence by triggering misleading interpretations.

Beyond the vulnerabilities of these mathematical surrogates, the recent paradigm shift introduces entirely new categories of risk. As the field transitions toward Large Language Models, the nature of robustness has shifted from the numerical to the semantic level. While these generative models can produce linguistically consistent and highly persuasive narratives, they often lack a deterministic link to the underlying technical data. This creates a 'factual fragility' where the system might produce a detailed, logical-sounding explanation of a non-existent threat or, conversely, misinterpret a critical attack vector as a benign event. Consequently, the current state of the art faces a paradoxical tension: we have achieved greater descriptive clarity for the human operator, but we have introduced a more volatile and less predictable security foundation that requires constant oversight to distinguish between genuine forensic insights and plausible model hallucinations.

Recent advancements have further pushed these boundaries by integrating hybrid attention mechanisms to interpret even obfuscated data sources, such as encrypted traffic flows, improving both classification performance and explainability [42]. While statistically rigorous, these outputs often fail the semantic test, offering mathematical weighting where analysts require narrative context. This limitation is particularly critical in Insider Threat Detection (ITD), where differentiating between malicious intent and legitimate privilege misuse requires nuanced behavioral explanations rather than simple feature attribution [32]. In contrast, the 2024–2025 cohort signals a decisive shift toward **Generative Interpretability**. Papers such as Palma et al. [17] and Park et al. [23] deploy LLMs not merely for detection, but to synthesize natural language summaries of security logs. This evolution is fundamental: it moves the interface from abstract feature importance to semantic reasoning, aligning system outputs with the cognitive workflow of human operators. This alignment is particularly vital in Smart City infrastructures, where hyper-connectivity amplifies vulnerability, and decision transparency is mandatory for maintaining public trust in automated governance systems [31]. Moreover, this generative capability is revolutionizing Cyber Threat Intelligence (CTI), enabling the automated extraction of actionable insights—or "diamonds"—from vast unstructured datasets, a process previously bottlenecked by human cognitive limits [40].

Synthesizing the evolution of XAI paradigms mapped in this section, a clear trend emerges: a transition from deterministic, post-hoc feature attribution (e.g., SHAP, LIME) to generative, semantic interpretability driven by LLMs. However, this transition introduces a critical trade-off: while generative models significantly reduce the cognitive barrier for human operators by providing natural language narratives, they simultaneously diminish algorithmic robustness, introducing severe risks of hallucinations and explanation instability. A key limitation of the current literature is the systemic over-reliance on these surrogate and generative methods, highlighting an urgent need to pivot toward intrinsically interpretable architectures that can offer both semantic clarity and mathematical faithfulness.

3.3. RQ2: Human Factors and Evaluation Metrics

We found an interesting methodological dissonance that characterizes the current state of the art: the chasm between "Human-Centric" and the empirical reality. As reflected in our Quality Assessment, numerous studies achieve high overall scores due to robust algorithmic architectures, but fail in the "XAI & Human Factors Validity" dimension. While 63% of the analyzed papers prioritize the human operator, rigorous user studies or feedback loops appear in a negligible ~22% of the corpus.

A fundamental issue driving this gap is the widespread absence of a formal definition of what constitutes a "correct" or "faithful" explanation in operational security. To resolve

this ambiguity, we formally define a *faithful explanation* in a cybersecurity context as an output that accurately maps to the true causal mechanism of the model's decision—rather than merely highlighting statistical correlations—and is fully verifiable against raw network telemetry. This necessitates a pivot toward *causal explainability*, which empowers analysts to trace the explicit sequence of malicious actions. Furthermore, establishing operator trust requires rigorous *explanation verification* frameworks and dedicated *faithfulness benchmarking metrics* [34,35] to continuously audit the alignment between the XAI's generated narrative and the underlying data.

A review of the experimental setups reveals a significant discrepancy in data realism and evaluation strategies. Table 7 summarizes the primary datasets identified, while Table 8 categorizes the existing and proposed human-centered metrics.

Dietz et al. [36] confront this abstraction by introducing *Security Personas*, arguing that explanations must be dynamically scoped to the recipient's hierarchy (e.g., Tier-1 Analyst versus Chief Information Security Officer (CISO)). Expanding this inclusivity, Sayeed et al. demonstrate that AI-driven security interfaces can be effectively tailored even for non-technical, vulnerable populations like the elderly, validating the "Human-Centric" premise beyond the professional SOC context [39]. Yet, as detailed in Table 3, this nuance is largely absent from the broader landscape, where frameworks remain fixated on sterile statistical metrics (Accuracy, F1-Score).

Table 7. Primary datasets identified in the selected studies.

Dataset Name	Type	Realism	Key Reference
UNSW-NB15	Synthetic/PCAP	Medium	Akshya et al. [30]
CIC-IDS2017/18	Real-world Traffic	High	Pawlicki et al. [35]
HDFS Logs	System Logs	Medium	Bhusal et al. [34]
Corporate Logs	Real/Proprietary	Very High	Palma et al. [17]
NSL-KDD	Legacy/Synthetic	Very Low	Alansary et al. [1]

Table 8. Taxonomy of human-centered evaluation metrics.

Metric	Focus	Status in Literature
Accuracy/F1	Model Performance	Dominant (80% of papers)
Trust Score	Perceptual Trust	Emerging (Gupta et al. [31])
NASA-TLX	Cognitive Load	Proposed/Rare (Dietz et al. [36])
Actionability	Decision Support	Proposed (Bhusal et al. [34])
Time-to-Comprehend	Efficiency	Proposed in this Study

Exceptions such as Bhusal et al. [34] and Palma et al. [17] distinguish themselves by integrating qualitative dimensions like "Explanation Entropy" and "Actionability". Their superior Quality Assessment (QA) scores underscore a shifting academic standard: rigor in XAI is no longer defined solely by algorithmic performance, but by the tangible utility of the explanation.

Drawing these human-factor evaluations together, a prominent validation gap characterizes the current trend in the literature: while the majority of studies conceptually claim a human-centric focus, they practically evaluate systems using mathematical proxies (e.g., fidelity, accuracy) rather than empirical user studies. The primary trade-off observed here is between the ease of automated quantitative benchmarking and the ecological validity

of qualitative, human-in-the-loop testing. Consequently, a major limitation remains the widespread absence of standardized, deployable psychometric frameworks capable of rigorously quantifying cognitive load and trust in real-world SOC environments.

3.4. RQ3: Edge vs. Cloud Trade-Offs

The literature exposes a sharp tension between LLM capabilities and deployment realities. Research centering on critical infrastructure [23,30,33] consistently demonstrates that cloud-based inference is untenable; strict latency requirements and privacy statutes (e.g., GDPR) prohibit offloading data. Consequently, a sub-discipline of *Small Language Models (SLMs)* and distilled architectures is gaining traction. Furthermore, distributed architectures that leverage fog computing to offload computational tasks while maintaining local interpretability are becoming essential for scalable IoT defense [33]. Park et al. [23] validate that specialized SLMs can function effectively on shipboard networks, delivering necessary reasoning capabilities without the exorbitant overhead of GPT-4 class models. Current evidence suggests that the scalable deployment of cognitive security may increasingly rely on distributed, hierarchical intelligence rather than exclusively on monolithic cloud models.

Table 9 synthesizes the key findings, supporting evidence, and identified gaps for each of the three research questions addressed in this review.

Table 9. Summary of key findings, supporting evidence, and identified gaps per research question.

RQ	Key Findings	Supporting Studies	Identified Gaps
RQ1	SHAP and LIME dominate post-hoc XAI in cybersecurity. LLM-generated explanations are emerging but lack standardization. Gradient-based methods (IG, Grad-CAM) appear in log and traffic domains.	[14,17,33–35]	No systematic comparison across data types. Computational cost for edge not evaluated. No benchmarking of LLM vs. classical XAI explanations.
RQ2	63% of studies claim human-centric relevance, but only 22% validate empirically with real SOC operators. Metrics such as trust and time-to-comprehend are mentioned but not formally defined.	[13,21,34–36]	No standardized human-centric evaluation protocols. Cognitive load and decision confidence lack formal definitions in SOC contexts.
RQ3	Cloud LLMs offer superior reasoning but conflict with GDPR and latency constraints at the edge. SLMs and on-device inference are proposed but not empirically validated in security contexts.	[15,23,30,33]	No empirical benchmarking of SLMs vs. LLMs at the edge. Hallucination detection mechanisms for edge-deployed models remain undefined.

To synthesize the deployment challenges addressed in this final section, the intersection of Generative AI and edge computing exposes a fundamental architectural trade-off: the advanced reasoning and summarization capabilities of cloud-based LLMs directly conflict with the strict latency, bandwidth, and data sovereignty requirements (e.g., GDPR) of

localized SOC operations. The emerging trend towards Small Language Models (SLMs) and fog computing attempts to bridge this divide. Nevertheless, a critical limitation persists: the literature currently lacks robust empirical benchmarking to validate whether these distilled, edge-deployed models can maintain the requisite explanation fidelity and hallucination resistance without the massive parameter counts of their cloud-based counterparts.

4. Discussion and Open Challenges

The synthesis of these 27 studies indicates a domain undergoing rapid transition. While Generative AI has expanded the technical capabilities of defense systems, it has simultaneously introduced new challenges to the reliability and verifiability of automated explanations. We identify three systemic challenges that, if left unaddressed, could significantly limit the operational utility of these advancements. This suggests that the core limitation is not technological but architectural: AI systems are increasingly sophisticated, yet the organizational structures governing how decisions are made, escalated, and owned remain largely implicit.

4.1. The Evaluation Crisis: The Metric Mirage

The most pervasive methodological flaw is a reliance on convenient proxies over operational reality. As Table 3 indicates, nearly 80% of current research evaluates explainability through algorithmic metrics like Fidelity or Stability. However, recent critical examinations suggest that the sheer proliferation of these metrics often leads to confusion rather than clarity, failing to add substantial value to the user's understanding [35]. As Dietz et al. [36] argue, mathematical fidelity does not equate to cognitive utility; a statistically precise explanation that overwhelms a stressed SOC analyst is a failure.

To bridge this validation gap, future evaluation paradigms would strongly benefit from integrating **Human-Centric Metrics**. To effectively structure this integration, we propose an explicit hierarchy distinguishing three levels of evaluation:

- **(a)** ***Conceptual metrics*** (e.g., general notions of "trust" or "actionability"), which provide theoretical design goals but lack direct operational testability;
- **(b)** ***Validated psychometric instruments*** (e.g., the NASA Task Load Index [52]), which offer standardized, reproducible measurements of an operator's cognitive load during controlled simulations;
- **(c)** ***Deployable SOC usability methodologies***, which consist of practical workflows integrated seamlessly into the SIEM dashboard to measure real-world efficacy without disrupting the analyst's task.

To address this, we operationalize for the SOC context the following deployable metrics, building on Hoffman et al. [53]:

- **Time-to-Comprehend (TTC):** The temporal duration from the moment an XAI-augmented alert is presented to the analyst until a definitive triage decision (escalate/dismiss) is made.
- **Decision Confidence:** A qualitative measure of the analyst's certainty in their triage decision, evaluable via mixed methods, combining operational feedback with level (b) instruments like the NASA-TLX [52] or customized surveys administered post-incident.

Future research should measure how XAI impacts these metrics in high-stress SIEM/Security Orchestration, Automation and Response (SOAR) workflows. Ultimately, establishing trust in autonomous decision-making systems necessitates a broader framework of "Responsible AI" that evaluates not just performance, but fairness, accountability, and transparency [37].

4.2. The Hallucination Vector

The reliance on GenAI-generated explanations [17] introduces a new, insidious attack surface: the hallucination. Unlike a statistical false positive in a traditional Intrusion Detection System (IDS), an LLM-generated hallucination constitutes active disinformation within the security perimeter. This introduces a critical distinction between true causal explainability (which accurately maps the model's internal causal mechanisms and true decision-making process) and natural-language rationalization (which generates a plausible, yet fabricated, narrative disconnected from the model's logic) [17]. In a SOC environment, relying only on rationalizations may represent a risk, as a convincing hallucination can corrupt the forensic chain of custody and subvert downstream incident response planning [16]. To mitigate these risks, deploying GenAI in SOCs requires structural safeguards, such as Retrieval-Augmented Generation (RAG) systems, mandatory Human-in-the-Loop (HITL) validation checkpoints, and 'grounding' with the sources of information.

4.3. Adversarial Robustness and Explanation Stability

Beyond hallucinations, the deployment of XAI in adversarial environments introduces the critical vulnerability of explanation manipulation. As demonstrated by Ghorbani et al. [11], neural network interpretations are inherently fragile: sophisticated attackers can apply imperceptible perturbations to input data that leave the model's predictive output intact while radically altering the resulting explanation. In a SOC context, this adversarial manipulation of explanations allows a threat actor to effectively hide a malicious payload by forcing the XAI to highlight benign network features, thereby deceiving the analyst. Furthermore, even in the absence of an active attack, many post-hoc methods suffer from severe explanation instability (or "jitter"), where minor, non-malicious variations in network telemetry produce wildly inconsistent attributions. To counter this, future Cognitive SOC frameworks must move beyond merely generating explanations to actively evaluating their robustness under attack. This requires the integration of rigorous explanation verification frameworks and continuous faithfulness evaluation protocols, designed to audit the stability of the XAI output against baseline normative behavior.

4.4. Ecological Invalidity: Transformers on Archaic Data

A notable dissonance exists between model architecture and data quality. Researchers are deploying state-of-the-art Transformers and Graph Neural Networks upon legacy datasets. Studies such as [1] continue to benchmark against derivatives of NSL-KDD—data that predates modern multi-vector attacks by decades. This represents a threat to ecological validity [33]: state-of-the-art architectures are being evaluated on datasets that do not reflect current attack vectors, limiting the generalizability of reported results to real-world deployment scenarios. Until the community coalesces around dynamic, contemporary datasets, the generalizability of academic models to operational zero-day scenarios remains empirically undemonstrated [33,46].

Despite these pervasive data quality issues, there are encouraging signals that the research community is beginning to address this gap through the adoption of modern, high-fidelity benchmarks. For instance, recent studies within our corpus (e.g., [35]) evaluate metrics using the CIC IoT dataset 2023 and CSE-CIC-IDS2018, demonstrating a necessary shift towards operationally relevant telemetry.

4.5. Limitations of the Study

It is worth noting that the stringency of our search protocol constitutes a primary threat to validity. By mandating the strict intersection of four complex domains (XAI, Cybersecurity, Human-Centered Design, and Edge Computing), the search query acted

as a highly restrictive filter, yielding a final corpus of only 27 selected studies over an 11-year observation window. While this narrow scope successfully isolated research at the exact operational nexus of our investigation, it inevitably introduces selection bias. Studies offering significant advances in human-centered XAI for cybersecurity that did not explicitly target Edge constraints were systematically excluded, resulting in a highly specialized, niche cross-section of the field. Notably, this led to the under-representation of critical complementary areas for operationalizing Edge SOCs, such as the robustness of XAI against adversarial attacks (where explanations are manipulated to conceal malicious payloads) and the deployment of privacy-preserving Federated Learning (FL) frameworks. We acknowledge that isolating Human Factors from these specific domains limits the generalizability of our findings regarding system robustness, and caution must be exercised before extending these conclusions to broader security architectures. Beyond the constraints of our search protocol, the evidence base itself presents inherent limitations. We explicitly acknowledge that the "Cognitive SOC" framework discussed herein remains primarily conceptual and currently lacks formal operationalization. As demonstrated by our findings, the current literature critically lacks reproducible evaluation pipelines and concrete experimental protocols for validating human-centered XAI in operational environments. Thus, the absence of a concrete experimental pipeline in our synthesis is not a methodological omission on our part, but rather a direct reflection of the systemic immaturity we aimed to expose. Our primary contribution lies precisely in mapping these gaps to establish a foundational blueprint for future empirical validation.

5. Conclusions and Future Directions

This Systematic Literature Review highlights that, alongside detection accuracy, the operational integration of automated reasoning into human decision-making workflows has emerged as a critical bottleneck—a gap that current research has yet to systematically address.

To operationalize the next generation of Cognitive SOCs, the literature suggests five strategic imperatives for future research:

1. **Prioritize Intrinsic Interpretability:** Future research should explore architectures in which interpretability is embedded by design, rather than applied post-hoc. Emerging frameworks such as Kolmogorov-Arnold Networks (KANs) [47] represent a promising direction in this regard. These architectures embed interpretability directly into their mathematical formulation, bypassing the faithfulness issues of post-hoc methods while maintaining competitive predictive performance.
2. **The Era of Small Language Models (SLMs):** Privacy and latency constraints often make cloud-based LLMs impractical for certain critical infrastructure. Consequently, there is a growing operational need for model distillation [23]: bespoke, efficient generative models capable of reasoning at the Edge.
3. **Mandatory Human-in-the-Loop Validation:** To move beyond theoretical claims of usability, the literature underscores the necessity of standardized evaluation protocols that treat the human analyst as an active variable in the experimental design, rather than a passive consumer [34,36].
4. **Decision Architecture as a Design Constraint:** Future research should therefore move beyond model-centric perspectives and investigate decision architecture as a critical design layer that enables organizations to translate AI capabilities into effective and accountable action [18,37].
5. **Explanation Verification and Hallucination Mitigation:** To address the factual fragility of LLMs and the critical problem of hallucinations, the deployment of generative explanation systems must be coupled with concrete verification strategies.

Future architectures should integrate Retrieval-Augmented Generation (RAG) [15,17] to actively ground generated explanations in local, verified threat intelligence. Furthermore, systems must employ cross-validation protocols that continuously audit LLM-generated narratives against deterministic rule-based telemetry and explicit faithfulness benchmarking metrics [34,35] to ensure operational reliability.

An AI system deployed in security-critical environments without adequate explainability mechanisms introduces operational risks that may outweigh its potential benefits [8,34].

Ultimately, we hope that the findings and the taxonomy synthesized in this research will provide a practical blueprint for security architects and researchers. By serving as a foundational guideline for the design of next-generation Cognitive SOCs, we aim to inspire the creation of security environments that are not only algorithmically advanced, but genuinely human-centered, facilitating a more effective and verifiable integration between artificial intelligence and human cognition.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jcp6030091/s1, Table S1: PRISMA 2020 Checklist used in the current systematic review study.

Author Contributions: Conceptualization, G.C., F.B., M.C., G.P. and A.R.; methodology, G.C., F.B., M.C., G.P. and A.R.; validation, G.C., F.B., M.C., G.P. and A.R.; formal analysis, G.C., F.B., M.C., G.P. and A.R.; investigation, G.C., F.B., M.C., G.P. and A.R.; resources, A.R.; writing—original draft preparation, G.C. and G.P.; writing—review and editing, G.C., F.B., M.C., G.P. and A.R.; supervision, G.C., F.B., M.C., G.P. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. All extracted data are presented in Tables 2 and 3. The RIS export file is available from the corresponding author upon request.

Acknowledgments: The authors declare that Generative AI tools were used exclusively for the purpose of refining the English language and grammar of this manuscript. The scientific conception, data analysis, literature selection, and final conclusions are the sole work of the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alansary, S.A.; Ayyad, S.M.; Talaat, F.M.; Saafan, M.M. Emerging AI Threats in Cybercrime: A Review of Zero-Day Attacks via Machine, Deep, and Federated Learning. *Knowl. Inf. Syst.* **2025**, *67*, 10951–10987. [CrossRef]
2. Yilmaz, E.; Can, O. Unveiling Shadows: Harnessing Artificial Intelligence for Insider Threat Detection. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 13223–13230. [CrossRef]
3. Omar, M.; Zangana, H. Redefining Security with Cyber AI. In *Advances in Computational Intelligence and Robotics*; IGI Global: Hershey, PA, USA, 2024; pp. 1–20. [CrossRef]
4. Joshi, A.; Kumar, V.; Chauhan, N.; Thakur, G. Leveraging Deep Learning for Enhanced Security in Cloud-Based Critical Infrastructure. In *Advances in Information Security, Privacy, and Ethics*; IGI Global: Hershey, PA, USA, 2025; pp. 389–404. [CrossRef]
5. Phanireddy, S. AI-Driven Identity Access Management (IAM). *Int. J. Sci. Res. Eng. Manag.* **2021**, *5*. [CrossRef]
6. Sharma, S.; Tripathy, P.; Yadav, S.; Mohapatra, S.; Singh, P. Adaptive Threat Detection: Leveraging Machine Learning for Real-Time Cybersecurity. In *2024 International Conference on Advances in Computing, Communication and Materials (ICACCM)*; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
7. Swetha, T.; Kumaran, U.; Meena, V.P.; Hameed, I.A. Leveraging AI for Enhanced Cybersecurity: A Comprehensive Review. *Discov. Appl. Sci.* **2025**, *7*, 373. [CrossRef]

8. Safavi, S.; Abdulnabi, M.S.H.; Rana, M.E.; Alizadeh, S. From Black Box to Trustworthy AI: A Secure Framework for Explainable Cybersecurity Decision-Making. In *2025 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*; IEEE: Piscataway, NJ, USA, 2025. [CrossRef]
9. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 1135–1144. [CrossRef]
10. Lundberg, S.M.; Lee, S. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774. [CrossRef]
11. Ghorbani, A.; Abid, A.; Zou, J. Interpretation of Neural Networks Is Fragile. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 3681–3688. [CrossRef]
12. Vultureanu-Albişi, A.; Bădică, C.; Ivanović, M. eXING-IoT: Conceptual Framework for Explainability Integration in Next Generation-IoT. *Connect. Sci.* **2025**, *37*, 2446338. [CrossRef]
13. Jaigirdar, F.T.; Rudolph, C.; Anwar, M.; Tan, B. Empowering End-Users with Cybersecurity Situational Awareness: Findings from IoT-Health Table-Top Exercises. *J. Cybersecur. Priv.* **2025**, *5*, 49. [CrossRef]
14. Zhang, Z.; Al Hamadi, H.; Damiani, E.; Yeun, C.Y.; Taher, F. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access* **2022**, *10*, 93104–93139. [CrossRef]
15. Ahi, K.; Valizadeh, S. Large Language Models (LLMs) and Generative AI in Cybersecurity and Privacy: A Survey of Dual-Use Risks, AI-Generated Malware, Explainability, and Defensive Strategies. *IEEE Access* **2025**, *13*, 1–8. [CrossRef]
16. Abuadbba, A.; Hicks, C.; Moore, K.; Mavroudis, V.; Hasircioglu, B.; Goel, D.; Jennings, P. From Promise to Peril: Rethinking Cybersecurity Red and Blue Teaming in the Age of LLMs. *arXiv* **2025**, arXiv:2506.13434. [CrossRef]
17. Palma, G.; Cecchi, G.; Caronna, M.; Rizzo, A. Leveraging Large Language Models for Scalable and Explainable Cybersecurity Log Analysis. *J. Cybersecur. Priv.* **2025**, *5*, 55. [CrossRef]
18. Casino, F. Unveiling the Multifaceted Concept of Cognitive Security: Trends, Perspectives, and Future Challenges. *Technol. Soc.* **2025**, *83*, 102956. [CrossRef]
19. Binbeshr, F.; Imam, M.; Ghaleb, M.; Hamdan, M.; Rahim, M.A.; Hammoudeh, M. The Rise of Cognitive SOCs: A Systematic Literature Review on AI Approaches. *IEEE Open J. Comput. Soc.* **2025**, *6*, 360–379. [CrossRef]
20. Li, J.; Zhang, M.; Li, N.; Weyns, D.; Jin, Z.; Tei, K. Generative AI for Self-Adaptive Systems: State of the Art and Research Roadmap. *ACM Trans. Auton. Adapt. Syst.* **2024**, *19*, 1–52. [CrossRef]
21. Kolse, A. Reducing Alert Fatigue in SOC Teams Through Contextual Prioritization and Threat Intelligence Integration. *OSF Prepr.* **2025**. [CrossRef]
22. Khayat, M.; Barka, E.; Serhani, M.A.; Sallabi, F.; Shuaib, K.; Khater, H.M. Empowering Security Operation Center with Artificial Intelligence and Machine Learning: A Systematic Literature Review. *IEEE Access* **2025**, *13*, 19162–19197. [CrossRef]
23. Park, D.; Min, B.; Lim, S.; Kim, B. ATIRS: Towards Adaptive Threat Analysis with Intelligent Log Summarization and Response Recommendation. *Electronics* **2025**, *14*, 1289. [CrossRef]
24. Pawlicki, M.; Pawlicka, A.; Kozik, R.; Choraś, M. Advanced Insights through Systematic Analysis: Mapping Future Research Directions and Opportunities for XAI in Deep Learning and Artificial Intelligence Used in Cybersecurity. *Neurocomputing* **2024**, *590*, 127759. [CrossRef]
25. Saarela, M.; Podgorelec, V. Recent Applications of Explainable AI (XAI): A Systematic Literature Review. *Appl. Sci.* **2024**, *14*, 8884. [CrossRef]
26. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
27. Kitchenham, B.; Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Technical Report EBSE 2007-001; Keele University: Keele, UK; Durham University: Durham, UK, 2007.
28. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic Mapping Studies in Software Engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE 2008)*; BCS Learning & Development: Swindon, UK, 2008; pp. 68–77.
29. Ouzzani, M.; Hammady, H.; Fedorowicz, Z.; Elmagarmid, A. Rayyan—A web and mobile app for systematic reviews. *Syst. Rev.* **2016**, *5*, 210. [CrossRef]
30. Akshya, J.; Sundarrajan, M.; Vijayakumar, R.; Dhanaraj, R.K.; Nayyar, A. Explainable AI-Driven Intrusion Detection for Securing IoT-Enabled Autonomous Transportation Systems. *Clust. Comput.* **2025**, *28*, 884. [CrossRef]
31. Gupta, S.; Singh, J.; Agrawal, R.; Batra, U. Explainable AI and Machine Learning for Robust Cybersecurity in Smart Cities. *Cyber Secur. Appl.* **2025**, *3*, 100084. [CrossRef]
32. Alketbi, K.S.; Mehmood, A. A Comprehensive Survey of Explainable Artificial Intelligence Techniques for Malicious Insider Threat Detection. *IEEE Access* **2025**, *13*, 121772–121798. [CrossRef]

33. Moustafa, N.; Koroniotis, N.; Keshk, M.; Zomaya, A.Y.; Tari, Z. Explainable Intrusion Detection for Cyber Defences in the Internet of Things: Opportunities and Solutions. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 1775–1807. [CrossRef]
34. Bhusal, D.; Shin, R.; Shewale, A.A.; Veerabhadran, M.K.M.; Clifford, M.; Rampazzi, S.; Rastogi, N. SoK: Modeling Explainability in Security Analytics for Interpretability, Trustworthiness, and Usability. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec)*; ACM: New York, NY, USA, 2023. [CrossRef]
35. Pawlicki, M.; Pawlicka, A.; Uccello, F.; Szelest, S.; D'Antonio, S.; Kozik, R.; Choraś, M. Evaluating the Necessity of the Multiple Metrics for Assessing Explainable AI: A Critical Examination. *Neurocomputing* **2024**, *602*, 128282. [CrossRef]
36. Dietz, K.; Mühlhauser, M.; Kögel, J.; Schwinger, S.; Sichermann, M.; Seufert, M.; Herrmann, D.; Hoßfeld, T. The Missing Link in Network Intrusion Detection: Taking AI/ML Research Efforts to Users. *IEEE Access* **2024**, *12*, 79815–79837. [CrossRef]
37. Dehghani, F.; Dibaji, M.; Anzum, F.; Dey, L.; Basdemir, A.; Bayat, S.; Boucher, J.-C.; Drew, S.; Eaton, S.E.; Frayne, R.; et al. Trustworthy and Responsible AI for Human-Centric Autonomous Decision-Making Systems. *arXiv* **2025**, arXiv:2408.15550. [CrossRef]
38. Tomičić, I.; Grd, P.; Bernik, A. Advancing the Threat Intelligence with AI: An Overview, Taxonomy and Roadmap. *J. Univers. Comput. Sci.* **2025**, *31*, 1102–1129. [CrossRef]
39. Sayeed, M.S.; Tamut, H.; Dutta, I.K. ElderConnect: An AI-Powered Platform to Empower Seniors Against Cyberthreats. In Proceedings of the IEEE International Conference on Electro Information Technology, Seattle, WA, USA, 28–30 May 2025; pp. 41–48. [CrossRef]
40. Barbieri, S.; De Souza, F.L.D.S.; Teixeira, M.A.; Marcondes, C.A.C.; Pereira, L.A. Searching for Diamonds: Cross-Domain Opportunities in Cyber Threat Intelligence. *IEEE Access* **2025**. [CrossRef]
41. Peña-Cáceres, O.; Garay-Silupu, E.; Aguilar-Chuquizuta, D.; Silva-Marchan, H. Research Trends and Networks in Self-Explaining Autonomous Systems: A Bibliometric Study. *Comput. Mater. Contin.* **2025**, *84*, 2151–2188. [CrossRef]
42. Sharma, A.; Lashkari, A. Hybrid Attention-Enhanced Explainable Model for Encrypted Traffic Detection and Classification. *Int. J. Inf. Secur.* **2025**, *24*, 97. [CrossRef]
43. Zivkovic, M.; Jovanovic, L.; Jocovic, V.; Nikolic, B.; Zeljkovic, V.; Abdel-Salam, M.; Mravik, M.; Muthusamy, S.; Bacanin, N. Addressing Smart City Security: Machine and Deep Learning Methodology Combining Feature Selection and Two-Tier Cooperative Framework Tuned by Metaheuristics. *Clust. Comput.* **2025**, *28*, 705. [CrossRef]
44. Benzaïd, C.; Taleb, T.; Sami, A.; Hireche, O. FortisEDoS: A Deep Transfer Learning-Empowered Economical Denial of Sustainability Detection Framework for Cloud-Native Network Slicing. *IEEE Trans. Dependable Secur. Comput.* **2024**, *21*, 3389–3403. [CrossRef]
45. Neisser, U. *Cognition and Reality: Principles and Implications of Cognitive Psychology*; W. H. Freeman: San Francisco, CA, USA, 1976.
46. Arp, D.; Quiring, E.; Pendlebury, F.; Warnecke, A.; Pierazzi, F.; Wressnegger, C.; Cavallaro, L.; Rieck, K. Dos and Don'ts of Machine Learning in Computer Security. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security 22)*; USENIX Association: Berkeley, CA, USA, 2022; pp. 3971–3988.
47. Liu, Z.M.; Wang, Y.X.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. KAN: Kolmogorov–Arnold Networks. *arXiv* **2024**, arXiv:2404.19756. [CrossRef]
48. Hassanin, M. KAN-LSTM: Benchmarking Kolmogorov-Arnold Networks for Cyber Security Threat Detection in IoT Networks. *arXiv* **2024**, arXiv:2603.28985. [CrossRef]
49. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2017; pp. 3319–3328. [CrossRef]
50. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
51. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. J. Law Technol.* **2017**, *31*, 841. [CrossRef]
52. Hart, S.G.; Staveland, L. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*; Hancock, P.A., Meshkati, N., Eds.; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183. [CrossRef]
53. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects. *arXiv* **2018**, arXiv:1812.04608. [CrossRef]