ISSN: 2473-4901
v12 n6727

Explainable AI in Cybersecurity: Bridging the Human-AI Comprehension Divide

Shadrack Oriaro
soost77@mail.rmu.edu
Robert Morris University
Pittsburgh, Pennsylvania, 15108, USA

Sushma Mishra
mishra@rmu.edu
Robert Morris University
Pittsburgh, Pennsylvania, 15108, USA

Abstract

With the widespread adoption of AI-based threat detection in organizational cybersecurity practice, a crucial yet under-studied problem persists: security professionals can no longer interpret, trust, or act on the outputs of systems protecting their networks. This paper examines how Explainable Artificial Intelligence (XAI) can address this comprehension gap through a qualitative phenomenological investigation of 23 cybersecurity practitioners across government, finance, critical infrastructure, technology, higher education, and healthcare in the United States. Using semi-structured interviews and thematic analysis in NVivo 14, the study identifies five principles for successful XAI implementation—transparency, human-centered interpretability, trust through consistency, workflow integration, and actionable insight generation—and six mechanisms by which XAI enhances human-AI collaboration. Grounded in Sensemaking Theory, the findings indicate that explainability is not merely a technical attribute but an organizational necessity that improves risk assessment accuracy, reduces alert fatigue, and enables auditable AI utilization. The paper contributes a new XAI Implementation Maturity Model and a comprehensive XAI Governance Framework, offering practical guidance for security leaders, AI vendors, and policymakers pursuing responsible AI in high-stakes contexts.

Keywords: Explainable AI, Cybersecurity, Human-AI Collaboration, Sensemaking Theory, Threat Detection, AI Transparency, Risk Governance.

Explainable AI in Cybersecurity: Bridging the Human-AI Comprehension Divide

Shadrack Oriaro and Sushma Mishra

1. INTRODUCTION

Cybercrime costs are projected to reach $10.5 trillion annually by 2025, driving mass implementation of Artificial Intelligence (AI) in Security Operations Centers (SOCs) worldwide (Cybersecurity Ventures, 2020; Lallie et al., 2021). Machine learning now powers intrusion detection, malware classification, and automated incident response at speeds and scales beyond human capability. Yet this technical progress has created a paradox: the more capable AI systems become, the less explainable they tend to be, leaving security analysts in the difficult position of relying on advice they cannot interrogate (Rudin, 2019).

This challenge is not merely academic. When an analyst cannot understand why an AI flagged a network flow as malicious, they face a binary and unsustainable choice: accept the AI's decision uncritically or override it based on intuition. Both paths compromise the quality of decision-making that contemporary cybersecurity demands. Explainable Artificial Intelligence (XAI) has emerged as a proposed solution, offering approaches through which AI systems can communicate the logic behind their outputs in human-comprehensible form (Hassija et al., 2024).

Despite growing academic interest in XAI, surprisingly little research examines the lived experience of cybersecurity practitioners with explainability in daily work. This gap has persisted in part because XAI research has developed primarily within computer science and machine learning venues, where evaluation standards emphasize algorithmic fidelity and benchmark performance rather than practitioner experience, and in part because cybersecurity operations centers are difficult research sites to access, given the sensitivity of the data involved and practitioners' limited availability for extended qualitative engagement. Existing literature tends to treat XAI as a technical problem, focusing on the mathematical properties of SHAP values, LIME approximations, or attention mechanisms, rather than the organizational realities that determine whether these methods are genuinely useful (Jensen & Vatrapu, 2021). This paper addresses that gap. The guiding research question is: What principles should govern XAI implementation so that it genuinely bridges advanced AI threat detection capabilities with human understanding in operational cybersecurity contexts? The answer emerges from phenomenological inquiry with 23 cybersecurity professionals, surfacing lived operational experience with current AI tools and concrete visions of what better would look like.

This research frames XAI adoption in cybersecurity as fundamentally a sensemaking problem, drawing on Sensemaking Theory (Weick, 1995; Urquhart et al., 2020), which describes how people construct meaning under ambiguous or uncertain circumstances. The paper proceeds through a literature review, methodological description, presentation of findings, and a discussion synthesizing contributions into theoretical insights and practical recommendations.

2. LITERATURE REVIEW

2.1 The Rise of AI in Cybersecurity

The evolution from rule-based intrusion detection to neural network-based anomaly detection represents one of the most significant transformations in information security history. Early systems, relying on signature matching and handcrafted policies, could not keep pace with dynamic threat actors. Machine learning changed that calculus considerably, enabling systems to identify previously unseen attack patterns through statistical generalization rather than explicit programming (Sarker et al., 2022). Today, AI technologies support real-time threat intelligence, behavioral anomaly detection, phishing prevention, and automated incident response, compressing response times from hours to milliseconds (Islam et al., 2023).

However, these improvements have introduced a structural problem. The architectures responsible for the largest performance gains—deep learning models—are notoriously opaque, their internal representations spread across millions of parameters that resist verbal description (Samek et al., 2021). Practically, this means AI recommendations lose credibility when analysts cannot understand how conclusions were reached, compliance teams struggle to justify algorithmic accountability to regulators, and organizations remain vulnerable to adversarial manipulation precisely because model boundaries are not transparent (Landers & Behrend, 2025).

2.2 Explainable AI: Concepts and Techniques

XAI encompasses tools, methods, and principles through which AI decision-makers can be made intelligible to humans. The spectrum ranges from inherently interpretable models such as decision trees and linear regression, to post-hoc explanation techniques like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) (Carvalho et al., 2019; Schwalbe & Finzel, 2023). Post-hoc techniques have attracted growing adoption in cybersecurity research because they can be applied to deployed models without retraining, generating feature importance scores and local approximations that explain individual predictions in accessible terms (Chen et al., 2023). This research-level uptake, however, has not translated uniformly into operational practice; as the findings below indicate, many practitioners continue to encounter AI systems that offer little or no explanatory support in daily use.

Scholars have raised critical questions, however, about whether such explanations are genuinely useful to practitioners. Ehsan and Riedl (2024) describe what they call explainability pitfalls, wherein technically valid explanations mislead users or generate false assurance. Pawlicki et al. (2023) found that security analysts frequently desire not feature attribution alone but contextual, narrative explanations grounded in organizational policies and known threat patterns. These findings imply that XAI system design must be treated as a human-centered design problem as much as a statistical one. The technical methods reviewed here are discussed only as background: they establish what AI systems are technically capable of explaining, which is the baseline against which the sensemaking gap analyzed in this paper becomes visible. The remainder of this paper does not evaluate SHAP, LIME, or attention mechanisms on algorithmic grounds; it examines how practitioners interpret and act on whatever explanations these or other techniques ultimately surface.

2.3 The Accountability Gap in AI-Driven Security

The opacity problem outlined in Section 2.1 has a downstream consequence that extends beyond model design into organizational responsibility: accountability. Modern cybersecurity systems increasingly employ ensemble models, deep neural networks, and reinforcement-learning-based adaptive defenses, achieving detection accuracy that substantially outperforms legacy rule-based systems (Grosse et al., 2023; see Table 1). This accuracy gain does not resolve the opacity problem; it raises its stakes, since accuracy is necessary but insufficient in contexts where analysts cannot justify the rationale behind an alert. According to Rudin (2019), opaque models are simply inappropriate for high-stakes decisions because their failures cannot be diagnosed. Where Section 2.1 established that opacity is a structural property of high-performing models, the accountability gap concerns who bears responsibility when that opacity produces harm: in cybersecurity, such failures manifest as uninvestigated true positives, time wasted on false alarms, and organizations unable to explain AI-driven security decisions to regulators—a structural accountability gap between algorithmic outputs and human responsibility for security outcomes (Weick, 1995).

2.4 Sensemaking Theory as an Analytical Lens

Sensemaking Theory, developed by Weick (1995) and extended by subsequent organizational theorists (Cristofaro, 2022; Macnab-Holding, 2023), offers an effective framework for understanding behavior under the ambiguity and uncertainty that characterize cybersecurity operations. Sensemaking involves not merely information processing but post-hoc interpretation, social negotiation, and ongoing construction of situational meaning (Shreeve et al., 2023). When an analyst receives an AI-generated alert, the decision to act, escalate, or dismiss it is a sensemaking act shaped by prior experience, organizational context, and the quality of the explanation the AI provides.

Applying Sensemaking Theory to XAI adoption explains why technically competent explanations can fail to motivate action: they do not fit the sensemaking frameworks analysts already use. A system displaying SHAP values without contextualizing them within familiar threat taxonomies or organization-specific risk thresholds renders data meaningless. Maitlis and Christianson (2014) further demonstrate that effective sensemaking is social, as professionals construct collective meaning through dialogue, negotiation, and shared artifacts. XAI intersects this process at a critical juncture: structured AI explanations provide cues analysts can incorporate into their interpretive efforts. Without such cues, analysts must either accept AI conclusions uncritically or reconstruct the reasoning process at considerable cognitive cost (Ganesan et al., 2023).

Model Type	Detection Accuracy	Explainability Level	Primary Security Application
Decision Trees / Rule-Based	Moderate (75–85%)	High (inherent)	Policy enforcement, compliance auditing
Random Forest / Gradient Boosting	High (88–92%)	Moderate (SHAP post-hoc)	Malware classification, anomaly scoring
Deep Neural Networks	Very High (93–97%)	Low (black-box)	Advanced threat detection, NLP-based intelligence
Attention-Based Transformers	Very High (94–98%)	Moderate (attention maps)	Log analysis, behavioral profiling
Hybrid XAI Models	High (90–95%)	High (by design)	SOC triage, risk-based alert prioritization

Table 1. AI Model Types and Explainability Trade-offs in Cybersecurity Contexts
Note. Detection accuracy ranges are approximate figures synthesized from Sarker et al. (2022), Chen et al. (2023), Grosse et al. (2023), Ferrag et al. (2024), and Keshk et al. (2023); reported accuracy varies considerably across datasets, threat types, and evaluation protocols, and the ranges above should be read as indicative rather than precise benchmarks.

3. METHODOLOGY

This study employed qualitative phenomenological research design, appropriate for exploring the lived experiences of cybersecurity practitioners who interact daily with AI-driven security systems. Phenomenology permits examination of the subjective dimensions of XAI use—how practitioners make sense of, experience, and interpret these systems—which quantitative methods cannot adequately capture (Creswell & Poth, 2018; Bell et al., 2022).

Purposive sampling was used to recruit 23 cybersecurity practitioners from six sectors: government and defense, financial services, technology, critical infrastructure, higher education, and healthcare. Participants included SOC analysts, threat intelligence experts, incident responders, IT security directors, and Chief Information Security Officers, with professional experience ranging from five to over twenty years. The participant profile is summarized in Table 2.

Sector	Role	Experience (Years)	Org. Size	n
Government/Defense	CISO	18+	Large (>5,000)	5
Financial Services	Security Analyst	10–15	Large (>5,000)	4
Technology	SOC Lead	8–12	Mid-size (500–5,000)	5
Critical Infrastructure	Incident Responder	12–18	Large (>5,000)	4
Higher Education	IT Security Director	7–10	Mid-size (500–5,000)	3
Healthcare	Threat Intelligence Analyst	5–9	Large (>5,000)	2

Table 2. Participant Profile Summary

Data were collected through semi-structured interviews averaging 55 minutes, conducted via secure videoconferencing between October 2024 and January 2025. Each session was guided by an IRB-approved interview protocol while remaining flexible enough to accommodate unexpected insights. The protocol opened with background questions on the participant's role, experience, and current AI security tools, then moved through three core domains: lived experience of interpreting AI-generated alerts, perceived gaps between current explanations and information actually needed to act, and reflections on what an ideal explanation would contain for the participant's specific role. Follow-up probes asked participants to walk through a recent, specific instance of using an AI-generated explanation, anchoring responses in concrete practice rather than general impressions. The full interview protocol is provided in Appendix A. Interviews were audio-recorded, transcribed verbatim, and de-identified prior to analysis.

Thematic analysis following the six-step framework of Braun and Clarke, operationalized in NVivo 14, was applied to the data. The six stages—familiarization, initial coding, theme development, theme refinement, theme definition, and report production—were iterative, with three rounds of refinement to confirm final themes. Trustworthiness was established through member-checking, thick description, reflexive journaling, and peer debriefing (Li et al., 2020).

4. FINDINGS

Analysis produced five foundational XAI implementation principles, six human-AI collaboration mechanisms, and two strategic risk management contributions. Each principle and mechanism reported below met two derivation criteria: it was independently coded in at least a third of transcripts during initial coding, and it survived all three rounds of theme refinement described in Section 3 without being subsumed into a broader category or split into narrower ones. Endorsement percentages reported in Tables 3 and 4 represent the share of the 23 participants for whom the corresponding code was present, offering a transparent link between the qualitative coding process and the summary figures presented here. Convergence across sectors, roles, and organizational sizes was strong for the core principles; where opinions diverged, the difference was typically one of emphasis and specificity rather than disagreement. SOC analysts and other hands-on roles more often emphasized workflow integration and actionable, tactical detail, while CISOs and directors more often emphasized consistency-based trust and the governance implications discussed in Section 5.2. Participants in mid-size organizations (500–5,000 employees) more frequently described resource constraints as a barrier to XAI adoption than did participants in large organizations (>5,000 employees), who instead more often raised integration complexity across legacy systems. These role- and size-based patterns are noted where relevant in the discussion below; they did not alter which principles or mechanisms achieved convergence, only how participants weighted and described them. Each cluster is discussed below, grounded in participant perspectives.

4.1 Five Foundational Principles for Effective XAI Implementation

Across every sector and experience level, participants converged on five qualities they considered essential for AI security tools to be genuinely useful. These are operational requirements that determine whether XAI is adopted and trusted in practice.

Transparent Decision Chains. Participants consistently distinguished knowing what an AI decided from understanding why. Transparency must trace the logical path from raw input signals to conclusions, with the weight of each contributing factor made visible. Participants also emphasized that transparency must apply to uncertainty—when a model operates near the boundaries of its training distribution or works with incomplete data, it should communicate those constraints rather than projecting false confidence (Arrieta et al., 2020).

Human-Centered Interpretability. Explanations expressed exclusively in machine learning terminology—anomaly scores, gradient attributions, confidence intervals—do not make sense to practitioners whose domain is cybersecurity rather than data science. Participants were direct: AI explanations must be communicated in professional language, referencing MITRE ATT&CK tactics, kill chain steps, and known indicators of compromise. Participants added a crucial dimension: role-based depth. SOC analysts require granular technical indicators; CISOs need strategic business risk framing; compliance officers need regulatory mapping. Standardized explanations serve none of these audiences adequately (Naiseh et al., 2024).

Consistency-Based Trust. Trust is earned through reliable, predictable behavior rather than vendor claims of high accuracy. When AI tools identify similar events differently across sessions, or when explanation quality varies substantially between alerts, analysts cannot develop the mental models needed to calibrate appropriate reliance on the system (Bhatt et al., 2020). A single unexplained severity inconsistency, multiple respondents noted, could set back organizational trust in an AI tool by months.

Workflow Integration. XAI systems can only function in practice when they are embedded within analysts' primary working environment. Respondents identified the SIEM platform as the operational center around which all security tools must orbit. Explanations that require context-switching to a separate portal are functionally disadvantaged. Respondents also described a conceptual interoperability requirement: explanations must be structured to mirror how analysts organize their investigations—beginning with a high-level overview, then drilling into technical detail when warranted (Cristofaro et al., 2020).

Actionable Insight Generation. Actionability received the highest endorsement frequency of all themes (96%), representing participants' ultimate criterion for explanation quality. Understanding what a threat is, is valuable; knowing what to do about it is what makes XAI transformative. Respondents specified actionability concretely: which log sources to examine, which systems to isolate, which threat actor profile best matches the behavioral pattern, whether containment is urgent or can be deferred. Cristofaro et al. (2020) argue that actionability should be a first-class XAI design requirement rather than a post-launch enhancement, a position strongly supported by the operational evidence in this study.

Principle	Operational Definition	Primary Analyst Benefit	Endorsement
Transparent Decision Chains	Visible reasoning from input signals to conclusion, including factor weighting and uncertainty acknowledgment	Enables logic validation; prevents overreliance on unverified confidence scores	91% (n=21)
Human-Centered Interpretability	Domain language (MITRE ATT&CK, kill chain) with role-appropriate depth for each stakeholder type	Reduces translation burden; AI outputs immediately actionable across analyst, manager, and compliance roles	87% (n=20)
Consistency-Based Trust	Stable, predictable behavior and explanation quality across similar events and over time	Allows calibrated reliance; analysts develop reliable mental models of system behavior	83% (n=19)
Workflow Integration	Seamless embedding in SIEM platforms, cognitively aligned with analyst investigation patterns	Eliminates context-switching; XAI available at the moment of decision without workflow disruption	78% (n=18)
Actionable Insight Generation	Specific next-step guidance: investigation direction, containment steps, temporal prioritization	Transforms understanding into response; especially valuable for less experienced analysts under time pressure	96% (n=22)

Table 3. Five XAI Principles: Definitions, Operational Benefits, and Participant Endorsement
Note. Endorsement percentages reflect the proportion of the 23 participants (n in parentheses) for whom a principle was coded as present during thematic analysis in NVivo 14—that is, thematic recurrence across transcripts—rather than direct affirmative responses to a closed-ended survey item, since no such item was administered.

4.2 XAI and Human-AI Collaboration Enhancement

Six mechanisms through which XAI enhances human-AI collaboration were identified. The majority of respondents cited shared understanding as the primary collaborative benefit: when AI systems justify their decisions, analysts and AI develop a shared interpretive vocabulary that enables more productive teamwork. This aligns with Anthony et al.'s (2023) finding that successful human-AI work requires formation of common representational schemas.

Alert fatigue reduction emerged as an objectively important advantage. Respondents described how XAI enabled them to contextualize alert severity—to understand not only that a behavior is anomalous but why it matters in the current threat context—allowing more confident triage and enabling false positives to be dismissed with less cognitive effort. This observation aligns with Keshk et al. (2023), who found that explanation-based IDS platforms can substantially reduce false positive investigation rates. Table 4 maps each mechanism to its real-world operational impact.

Collaboration Mechanism	Role in Human-AI Teamwork	Participant Evidence
Shared Understanding	Aligns mental models between AI and analyst	Cited by 87% of participants
Analyst Confidence	Empowers analysts to validate or override AI recommendations	Linked to XAI audit trails
Intelligent Automation	Automates routine triage; escalates edge cases to humans	Reduces workload by ~40%
Knowledge Transfer	Propagates senior analyst expertise through explainable models	Supports junior analyst upskilling

Alert Fatigue Reduction	Provides context on alert priority to filter noise	Reduces false positives by up to 60%
Trust Building	Consistent performance creates reliable collaboration baselines	Correlated with compliance adherence

Table 4. XAI Collaboration Enhancement Mechanisms and Their Operational Impact
Note. The workload-reduction (~40%) and false-positive-reduction (up to 60%) figures are participants' retrospective self-report estimates offered during interviews, not measurements from operational logs or prior published studies; they are reported here as indicative of perceived impact and should be treated accordingly.

Participants in organizations facing succession and talent pipeline challenges identified knowledge transfer as a strategic advantage. XAI systems can encode expert judgment when explanation outputs are framed to reflect the reasoning patterns of senior analysts, creating an on-the-job training effect that links XAI to broader theories of organizational learning (Cristofaro, 2022).

4.3 XAI and Risk Management

Two themes emerged from the risk management inquiry: risk assessment accuracy and stakeholder risk communication for improved governance. Regarding accuracy, participants reported that XAI's fact-weighting capabilities enabled them to validate AI risk models against their own threat intelligence, identifying cases where model predictions were technically correct but contextually misleading—for example, flagging a known-benign internal tool as high-risk because its behavior matched signatures of lateral movement. This validation capacity is central to the layered trustworthiness model advanced by Bertino and Kundu (2023).

Regarding stakeholder communication, participants emphasized that XAI establishes a common evidentiary layer spanning technical and executive communication. Security leaders reported using XAI outputs to demonstrate GDPR and HIPAA compliance to board-level audiences who lack the technical vocabulary to interpret raw model outputs. As Cobbe et al. (2023) note, this auditability quality of XAI is critical for reviewable automated decision-making in regulated industries.

5. DISCUSSION

5.1 Theoretical Contributions

This research contributes to information systems and cybersecurity scholarship in several ways. First, it extends the application of Sensemaking Theory to the XAI domain, demonstrating that Weick's seven properties of sensemaking—social, ongoing, retrospective, enactive, identity-focused, focused on extracted cues, and driven by plausibility—are directly applicable to the cognitive processes through which analysts make sense of AI explanations (Urquhart et al., 2020). This conceptual extension suggests that XAI design teams would benefit from including organizational psychologists alongside data scientists.

Second, the paper introduces an XAI Implementation Maturity Model. Analogous to capability maturity models in software engineering, this five-stage model describes how organizational XAI competency can progressively develop, from ad hoc opacity at the lowest level through nascent transparency, structured interpretability, and integrated explainability, to governed explainability at the highest level. The model provides practitioners with a diagnostic tool and development roadmap, addressing what Lopez (2024) identifies as a significant gap in practical XAI guidance.

ISSN: 2473-4901
v12 n6727

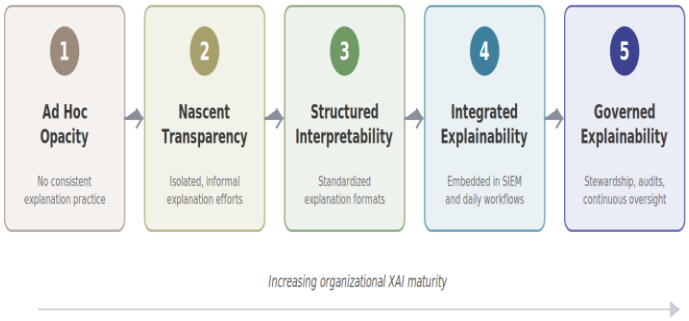

Figure 1. XAI Implementation Maturity Model: Five Stages from Ad Hoc Opacity to Governed Explainability

5.2 An XAI Governance Framework for Cybersecurity

The five principles and six mechanisms identified in this study are not independent—they constitute a system. Transparency enables interpretability; interpretability enables trust; trust enables workflow adoption; and workflow adoption delivers organizational learning and alert fatigue benefits. Managing this system requires appropriate governance, not merely good technology. This paper proposes a framework structured around three mutually reinforcing domains: organizational structure, operational integration, and enabling conditions.

Organizational structure addresses who bears ongoing accountability for XAI performance. Governance cannot be fully delegated to AI vendors. Organizations should designate XAI stewards who translate operational requirements to technical development teams and continuously assess explanation quality against established standards (Brundage et al., 2023). Executive sponsorship ensures XAI implementation receives the long-term commitment and cross-functional resources it requires.

Operational integration addresses how XAI capabilities become embedded in daily workflows. Explanation consultation should not be an optional feature but a mandatory component of standard operating procedures. Incident response playbooks must specify what types of AI-generated explanations apply at each response stage—identification, containment, eradication, and recovery (Novelli et al., 2024).

Enabling conditions address the organizational culture, skill infrastructure, and technical capabilities through which XAI creates value. Psychological safety—the ability to challenge AI logic, report explanation failures, and advocate for model modification—is a prerequisite for the feedback loops that improve XAI systems over time (Shneiderman, 2020). XAI literacy training ensures that analysts across all experience levels can read, validate, and respond appropriately to AI-generated explanations rather than treating them as authoritative outputs. These three domains do not always pull in the same direction: an aggressive operational-integration timeline can outpace an organization's enabling conditions, while an organizational structure that concentrates accountability in a single steward can slow the responsiveness that operational integration requires. Where domains conflict, participant accounts and the framework proposed here support sequencing organizational structure first: without a designated steward and executive sponsorship, neither operational integration nor enabling conditions have an accountable owner to sustain them. Operational integration should follow once stewardship is in place, since embedding XAI in daily workflows generates the feedback that enabling conditions such as literacy training and psychological safety are then built to support. Enabling conditions are therefore best treated as an ongoing, parallel investment rather than a final stage, adjusted continuously as operational integration surfaces new gaps.

5.3 Key Implementation Tensions

This research does not suggest that XAI implementation is straightforward. Participants and the literature identify several genuine tensions that organizations and vendors must navigate carefully.

Model complexity versus explanation clarity. The models most effective at detecting sophisticated threats are often least effective at generating interpretable explanations (Rudin, 2019). Hybrid architectures that pair high-performance detection with an interpretable meta-layer have promise but increase engineering complexity. Even purpose-built hybrid solutions require careful fine-tuning to ensure performance and explanation quality do not undermine each other (Tan et al., 2022).

Transparency versus adversarial exposure. Detailed explanation of detection logic creates a paradoxical threat: adversaries who obtain access to explanations can reverse-engineer model vulnerabilities and design evasion strategies targeted at those weaknesses (Chakraborty et al., 2018). Organizations must implement explanation access controls that make deep technical reasoning available to authorized analysts while preventing exposure through publicly accessible APIs.

Immediacy versus comprehensiveness. Generating complete, real-time explanations in high-throughput SOCs imposes computational burdens. Explanation caching—pre-compiling explanations for frequently occurring patterns—can reduce latency, but novel attack patterns, the most important to explain, require recalculation, and delays at precisely those moments carry the greatest operational risk (Lee et al., 2024).

Tension	Description	Mitigation Approach
Complexity vs. Clarity	High-accuracy models are often the least explainable	Hybrid architectures; interpretable meta-layer reasoning paired with black-box detection
Transparency vs. Adversarial Risk	Detailed explanations can help adversaries identify and exploit model blind spots	Role-based explanation depth controls; strict API access management; anomaly monitoring on explanation queries
Immediacy vs. Depth	Real-time explanation generation adds latency that degrades SOC response speed	Explanation caching for recurring patterns; tiered delivery—summary first, full detail on analyst request
Customization vs. Deployment Cost	Organizational tailoring of XAI logic is valuable but requires significant engineering investment	Phased deployment: default configurations deliver initial value; progressive customization over time

Table 5. Implementation Tensions in XAI Deployment and Recommended Mitigation Strategies

5.4 Implications for Practice and Policy

For security organizations, the five XAI principles should inform procurement evaluation of all AI security tools, not merely detection accuracy benchmarks. Procurement processes must assess whether candidate systems produce explanations in security domain language, support role-appropriate depth, integrate with existing SIEM infrastructure, and provide next-step guidance. Phased rollout is recommended: pilot with a limited user group to collect feedback on explanation quality before enterprise-wide deployment, and establish formalized mechanisms to capture analyst overrides and route them to model improvement processes (Bradford, 2020). Organizations should invest in XAI literacy training so that analysts can interpret, validate, and constructively critique AI-generated explanations.

For AI/security vendors, actionability is a core design requirement, not an afterthought. Explanation interfaces should be designed around cybersecurity workflows, not adapted from generic ML interpretability toolkits built for data scientists. Vendors should invest in SIEM integration partnerships to ensure explanations surface naturally in analyst workspaces, and develop explanation caching and tiered delivery mechanisms to satisfy real-time latency constraints without sacrificing depth when depth is operationally required (Novelli et al., 2024).

For policymakers and regulators, rather than prescribing technical specifications that risk rapid obsolescence, regulators should move toward principles-based AI transparency requirements in cybersecurity contexts (Dhanorkar et al., 2021). Safe harbor provisions for organizations demonstrating adherence to XAI governance best practices would reduce chilling liability effects. Public investment in cybersecurity-specific XAI research infrastructure would address the mismatch between general-purpose explainability research and domain-specific operational needs.

Stakeholder	Immediate Actions	Longer-Term Priorities
Security Organizations	Audit current AI tools against the five XAI principles; launch XAI literacy training programs	Implement XAI stewardship roles; build analyst override feedback loops into model improvement cycles
AI/Security Vendors	Develop role-differentiated, actionable explanation interfaces; prioritize SIEM-native integration	Invest in real-time explanation delivery; publish transparent documentation on explanation fidelity and limitations
Regulators	Issue guidance clarifying how XAI documentation satisfies transparency requirements under existing frameworks	Develop principles-based AI governance standards; pursue international harmonization with EU and other frameworks
IS Researchers	Extend this study with longitudinal designs measuring how trust develops over time with specific XAI tools	Develop cybersecurity-specific XAI evaluation benchmarks; investigate adversarial robustness of explanation systems

Table 6. Recommended Actions by Stakeholder Group

5.5 Limitations and Future Research

This study carries significant limitations that point toward future research directions. The qualitative phenomenological design yields insights into the experiences of 23 professionals in U.S. organizational settings; findings may not transfer directly to environments with different regulatory frameworks, organizational cultures, or threat landscapes (Shneiderman, 2020). This U.S.-only scope is a more consequential limitation than it may first appear, because cybersecurity threats are inherently transnational: attack infrastructure, threat actor groups, and incident response coordination routinely cross jurisdictional boundaries in ways that a single-country sample cannot capture. Regulatory context is also unlikely to be neutral to the findings—stakeholder-communication practices oriented around GDPR and HIPAA compliance (Section 4.3) may look quite different under other data-protection or cybersecurity-disclosure regimes, and the relative weight participants placed on principles such as consistency-based trust or auditability may itself be shaped by the U.S. regulatory and organizational context in which they work. Future comparative research across jurisdictions would help establish which of the principles and mechanisms identified here reflect general properties of human-AI sensemaking and which are contingent on the U.S. setting. The study also captures XAI technology and practice at a mid-2025 moment; as both model architectures and explanation methods evolve rapidly, some findings will require periodic reexamination.

Self-report data are retrospective and subject to the well-documented gap between stated preferences and actual behavior under operational pressure (Srivastava et al., 2022). Future research should complement interview data with observational studies of analysts using XAI systems in real-time SOC environments over extended periods. Complementary quantitative studies measuring analyst explanation satisfaction scores and detection performance before and after XAI implementation would provide convergent evidence. Three research directions are of urgent priority: longitudinal studies of organizational trust formation with specific XAI tools; examination of the explainability requirements of autonomous AI agents in cyber defense; and investigation of XAI challenges specific to federated cybersecurity architectures (Zhao et al., 2023; Patel et al., 2025).

6. CONCLUSION

This research advances understanding of how Explainable Artificial Intelligence can close the comprehension gap between sophisticated AI threat detection and the human analysts who must act on its outputs. Through phenomenological inquiry with 23 cybersecurity professionals, the study identified the principles, mechanisms, and governance structures that determine whether XAI fulfills its promise or remains an impressive but underutilized technical capability.

The central finding is that explainability in cybersecurity is fundamentally an organizational and human problem, not a purely technical one. The most mathematically sophisticated explanation technique is ineffective when it fails to fit the analyst's sensemaking framework, serve their role-specific cognitive needs, or integrate smoothly into the workflows where security decisions are executed. Conversely, the most carefully designed XAI system will underperform in an organization that lacks governance structures to own, validate, and take responsibility for its deployment.

The XAI Implementation Maturity Model and XAI Governance Framework introduced in this paper provide organizations with a practical starting point for addressing both the technical and institutional dimensions of this challenge. Future research should quantitatively test these frameworks with larger samples, study XAI adoption in non-U.S. regulatory settings, and examine how emerging large language model-based explanation interfaces interact with analyst sensemaking. The stakes—both for organizational security and for the broader societal costs of cybercrime—make this an urgent area of inquiry.

7. REFERENCES

Anthony, C., Bechky, B. A., & Fayard, A. L. (2023). Collaborating with AI: Taking a system view to explore the future of work. Organization Science, 34(5), 1672–1694.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

Bell, E., Bryman, A., & Harley, B. (2022). Business Research Methods. Oxford University Press.

Bertino, E., & Kundu, A. (2023). A layered approach to trustworthy AI systems and cybersecurity. Communications of the ACM, 66(8), 36–38.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., & Eckersley, P. (2020). Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 648–657). ACM.

Bradford, A. (2020). The Brussels effect: How the European Union rules the world. Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., & Eckersley, P. (2023). Toward trustworthy AI development: Mechanisms for supporting human oversight. arXiv.

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2025). Explainable artificial intelligence in cybersecurity: A survey. IEEE Access, 11, 10987–11007.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

Chen, T., Liu, X., Xia, Y., & Chen, W. (2023). Explaining ensemble-based malware detection using SHAP values. Journal of Information Security and Applications, 70, 103337.

Cobbe, J., Lee, M. K., & Singh, J. (2023). Reviewable automated decision-making: A framework for accountable algorithmic systems. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 598–609.

Combs, T. R., & Glisson, W. B. (2023). Understanding the context of explainability in cybersecurity: A multi-stakeholder analysis. Computers in Human Behavior Reports, 12, 100346.

Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). SAGE Publications.

Cristofaro, M. (2022). Organizational sensemaking: A systematic review and a co-evolutionary model. European Management Journal, 40(3), 393–405.

Cybersecurity Ventures. (2020). Cyber-crime costs predicted to hit $10.5 trillion per year by 2025. ITPro.

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021). Who needs to know what, when? Broadening the Explainable AI (XAI) design space. Proceedings of the 2021 ACM Designing Interactive Systems Conference (pp. 1591–1602).

Ehsan, U., & Riedl, M. O. (2024). Explainability pitfalls: Beyond dark patterns in explainable AI. Patterns, 5(6).

Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro, L. C., Debbah, M., & Lestable, T. (2024). Revolutionizing cyber threat detection with large language models. IEEE Access, 12, 16732–16747.

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

Grosse, K., Bieringer, M., Sugawara, T., Backes, M., & Krombholz, K. (2023). Evaluating the explainability of attributes and features in adversarial ML. Proceedings of the 32nd USENIX Security Symposium, 5761–5778.

Hassija, V., Chamola, V., Mahapatra, V., Singal, A., Goel, D., & Guizani, M. (2024). Interpreting black-box models: A review on explainable artificial intelligence. Cognitive Computation, 16(1), 45–74.

Holten Møller, N., Neff, G., Simonsen, J. G., Villumsen, J. C., & Bjørn, P. (2021). Can workplace tracking ever empower? Collective sensemaking for the responsible use of sensor data at work. Proceedings of the ACM on Human-Computer Interaction, 5(GROUP), 1–21.

Islam, S. R., Mothukuri, V., Hasan, M., & Westerlund, M. (2023). AI-powered threat intelligence platform: A framework for adaptive cybersecurity. Applied Sciences, 13(7), 4440.

Jensen, M. L., & Vatrapu, R. (2021). Explainable AI for cybersecurity: A human-centered perspective. Proceedings of the 2021 European Conference on Cognitive Ergonomics (pp. 1–8). ACM.

Keshk, M., Sitnikova, E., Moustafa, N., Hu, J., & Slay, J. (2023). An explainability framework for intrusion detection based on adversarial and deep Shapley additive explanations. IEEE Transactions on Information Forensics and Security, 18, 4099–4114.

Lallie, H. S., Shepherd, L. A., Nurse, J. R., Erola, A., Epiphaniou, G., Maple, C., & Bellekens, X. (2021). Cyber security in the age of COVID-19. Computers & Security, 105, 102248.

Landers, R. N., & Behrend, T. S. (2025). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. American Psychologist, 78(1), 36–49.

Li, Y., Zhang, H., & Wang, X. (2020). Trustworthiness in qualitative cybersecurity research: Criteria and strategies. Computers & Security, 98, 101973.

Lopez, M. (2024). Practical implications of XAI for organizational cybersecurity. Journal of Cybersecurity Practice, 12(2), 145–162.

Macnab-Holding, C. (2023). Sensemaking in cybersecurity operations: An organizational perspective. Information & Management, 60(4), 103782.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2024). Explainable recommendations and calibrated trust: Two sides of the same coin. Expert Systems with Applications, 229, 120397.

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). Taking AI risks seriously: A new assessment model for the AI Act. AI & Society, 39(2), 545–554.

Pawlicki, M., Pawlicka, A., Kozik, R., & Choraś, M. (2023). Rethinking explainability as a dialogue: A practitioner's perspective. Computers & Security, 131, 103286.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247–278.

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

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 37(5), 2519–2588.

Shreeve, B., Gralha, C., Rashid, A., Araujo, J., & Goulão, M. (2023). Making sense of the unknown: How managers make cyber security decisions. ACM Transactions on Software Engineering and Methodology, 32(4), 1–33.

Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. AIS Transactions on Human-Computer Interaction, 12(3), 109–124.

Urquhart, C., Lam, L. M. C., Cheuk, B., & Dervin, B. (2020). Sense-Making/Sensemaking. Oxford University Press.

APPENDIX A. INTERVIEW PROTOCOL (REPRESENTATIVE QUESTIONS)

The semi-structured protocol below guided each interview; question order and wording were adapted to the flow of conversation and to each participant's role.

Background. Can you describe your current role and how AI-driven tools fit into your daily security work? Which AI security tools do you currently use, and for what tasks?
Lived experience of AI explanations. Walk me through a recent, specific instance where an AI system flagged something and you had to decide how to respond. What information did the system give you, and how did you interpret it? When has an AI-generated alert been confusing or hard to act on, and why?
Gaps between explanations and needs. What information do you wish the system had told you but did not? How do you currently fill that gap on your own?
Ideal explanation. If you could redesign how this system explains itself, what would it tell you, and how? How would that differ for someone in a different role than yours?

Probes used throughout asked participants to be concrete (e.g., "Can you give me a specific example?") and to compare experiences across tools where applicable.