Too Much to Trust? Measuring the Security and Cognitive Impacts of Explainability in AI-Driven SOCs

Abstract

Explainable AI (XAI) holds significant promise for enhancing transparency and trust in AI-driven threat detection in Security Operations Centers (SOCs). However, appropriate explanation formats in high-stakes environments remain challenging. To address this gap, we conducted a three-month study combining a survey (N1=248) with interviews (N2=24) to examine (1) how SOC analysts conceptualize AI explanations and (2) which types are perceived as actionable and trustworthy across roles. Findings reveal participants accepted less accurate XAI outputs when explanations were relevant and evidence-backed. Analysts emphasized the importance of understanding the rationale behind AI decisions, preferring contextual depth over simply presented outcomes. Building on these insights, this study reevaluates explanation methods within security contexts and shows that role-aware, context-rich XAI designs aligned with SOC workflows can improve utility. Such tailored explainability enhances comprehension, increases triage efficiency, and supports confident responses to threats.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; • **Security and privacy** → **Intrusion detection systems**; *Cybersecurity and defense*; • **Computing methodologies** → *Explainable AI (XAI)*; Machine learning.

Keywords

Explainable Threat Intelligence, Explainable AI, Security Operation Center, Security Analysts, Threat Intelligence, Alert Fatigue, False Positives, Alert Triage

ACM Reference Format:

. 2025. *Too Much to Trust?* Measuring the Security and Cognitive Impacts of Explainability in AI-Driven SOCs . In *Proceedings of (ACM CHI '26).* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Modern Security Operations Centers (SOCs) ingest a continuous stream of security alerts and notifications from monitoring tools (e.g., intrusion detection systems, SIEM platforms, Endpoint Detection and Response agents, etc.) that signal potential malicious activity[18, 22]. Examples of security alerts include indications of a brute-force login attempt (e.g., repeated failed login attempts from an external IP address), detection of malware activity (e.g., malicious file execution or unusual outbound network connections), or identification of phishing attempts (e.g., suspicious email patterns and known malicious URLs). Such anomalous activities within the network may indicate a security threat, and these alerts require immediate analyst assessment to ascertain their validity, potential impact, and necessary actions.

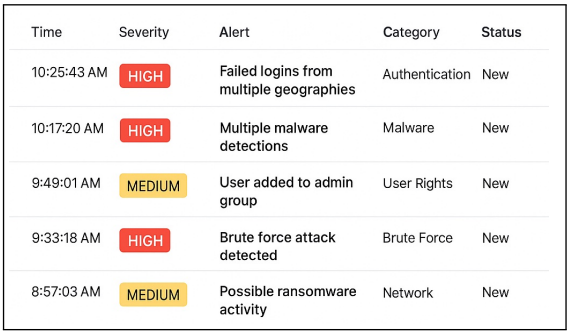
However, while alert detection has become increasingly automated, interpreting and responding to these alerts remains a deeply human, cognitively demanding task [30]. SOC analysts should quickly determine the relevance and severity of alerts under intense time pressure, frequently without clear insights into the underlying reasoning behind the automated detection. This disconnect leads to *alert fatigue*, a well-documented phenomenon marked by cognitive overload, overlooked threats, and inefficient triage workflows. Recent evidence shows that SOCs receive an average of 3,832 alerts per day, 62% of which are ignored; over 70% of analysts report feeling overwhelmed [18, 22, 34]. Effectively managing alerts is, therefore, crucial since missing a real threat alert notification can be detrimental to an organization, whereas chasing too many false alerts can lead to wasting resources [5, 47].

This points to a fundamental challenge between interpretability and usability in high-stakes, time-sensitive domains like security. On the one hand, explainable AI (XAI) techniques promise greater transparency, analyst trust, and insight into the reasoning behind AI-generated alerts [11, 31, 35]. On the other hand, SOC analysts operate in a high-stakes, time-sensitive environment that demands any explanation be fast, context-relevant, and cognitively lightweight to avoid impeding response. We posit that an AI-provided explanation's usefulness highly depends on *who* uses it, *when* it is used, and *why* someone consults it. An entry-level Tier-1 triage analyst, for example, might benefit from a concise summary of why an alert is likely malicious, whereas a Tier-3 threat hunter may require deeper technical details and model rationale [5, 24, 29, 47]. If explanations are too generic or too complex for the situation, they risk confusing the analyst or slowing down the investigation. Unfortunately, existing XAI methods, largely designed for technical debugging or regulatory transparency, fail to account for these contextual factors. As a result, explanations positioned in current research can either overwhelm analysts with extraneous details or, conversely, oversimplify important information, limiting their practical utility in the SOC workflow.

While recent studies have explored Security Operations Centers (SOCs) and their integration with AI/ML tools, none have specifically examined how the explainability of these models influences analyst workflows. This paper addresses this crucial and timely gap by providing the first comprehensive empirical investigation into the cognitive impacts and trust implications of AI-driven explanations within SOC environments.

ACM CHI '26, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX



Time	Severity	Alert	Category	Status
10:25:43 AM	HIGH	Failed logins from multiple geographies	Authentication	New
10:17:20 AM	HIGH	Multiple malware detections	Malware	New
9:49:01 AM	MEDIUM	User added to admin group	User Rights	New
9:33:18 AM	HIGH	Brute force attack detected	Brute Force	New
8:57:03 AM	MEDIUM	Possible ransomware activity	Network	New

Figure 1: Illustrative SOC Dashboard with alerts of different categories - High and Medium.

To address this, we introduce CORTEX- the *COntextual Role and Trust aware EXplanation* framework for the AI model's explanations used in SOC. CORTEX tailors the model's output as well as three key contextual dimensions: the analyst's investigative tier (role), the operational context of the alert, and the current cognitive load of the task. The goal of CORTEX is to deliver the optimal information to the analyst at the right time. We formalize this intuition through the *Trust-Explainability curve*, which models how explanation complexity and timing affect perceived utility and trust across different SOC roles. To evaluate and refine these ideas, we conducted a mixed-methods study over three months. We surveyed (N1=248) SOC professionals, followed by 24 in-depth interviews with analysts (over 810 minutes) across four continents, representing diverse roles, sectors (including managed security service providers, banking, IT, and finance), and years of experience (ranging from 1–7 years of on-the-job experience). The empirical study allowed us to examine the real-world limitations of established XAI methods (e.g., SHAP, LIME, LEMNA [20, 28, 36]) within realistic SOC contexts.

The findings revealed several challenges where generic explanations fall short, informing the design of our context-aware solution, CORTEX. Based on these insights, we propose a set of design recommendations for building more effective explainable security methods grounded in both the qualitative feedback from analysts and quantitative data on triage performance.

Our study is guided by the following research questions:

- **RQ1:** How do SOC analysts conceptualize explanations for AI-driven alerts, and which explanations are perceived as actionable or trustworthy across investigative roles?
- **RQ2:** Under which operational scenarios (e.g., active incident responses, routine monitoring, escalations to leadership) do analysts find explainability beneficial or neutral with respect to decision-making?
- **RQ3:** What technical, organizational, and cognitive barriers impede effective adoption and trust of XAI within SOC workflows?

This research study makes the following key contributions:

(1) **We provide actionable design guidelines grounded in CORTEX,** enabling alerting systems to dynamically adjust explanations based on investigative tier, role, and operational urgency.
(2) **We introduce the *Trust-Explainability curve*,** a novel approach characterizing when and how explainability shifts from being helpful to burdensome in different investigative contexts.
(3) **We present a comprehensive empirical investigation,** including a 248-participant survey and 24 analyst interviews totaling 810 minutes across SOCs in North America, Europe, Africa, and Asia. Full instruments are provided here[1].

2 Background and Motivation

2.1 Explainability Methods in Security

Recent explainable AI (XAI) techniques can be broadly categorized into two classes: *model-agnostic* and *model-specific*, each with distinct implications for interpretability, fidelity, and operational utility [16, 19]. **Model-agnostic methods** operate independently of the underlying learning algorithm and are usually applied post-hoc to generate explanations for any "black-box model." Some of the widely used models include LIME [36] and SHAP [28], which approximate local model behavior and generate feature-importance outputs. Examples of the explanations generated by the above-mentioned models is provided in Table 1. **Model-specific methods**, by contrast, are tightly coupled with the internal structure of a particular model to generate explanations. They exploit architectural details (weights, gradients, or activations) to produce explanations grounded in the model's internal computation [38]. For example, Grad-CAM [39], Integrated Gradients [44], DeepLIFT [41], and Layer-wise Relevance Propagation [7] are used in deep learning, especially vision and medical tasks [15]. Beyond these, a few methods have been tailored for security analytics, such as LEMNA for intrusion detection explanations [20] and XNIDS [46] for network anomaly interpretation. While these security-specific XAI methods align explanations with security telemetry semantics, they remain insufficiently responsive to real-time SOC demands and analyst heterogeneity. Additionally, they assume a one-size-fits-all explanation format, overlooking context-specific needs across different operational scenarios and roles.

2.2 Explainability and Human-centered Research in Security

Existing literature underscores the theoretical promise of XAI for security [10], yet real-world deployments in SOCs remain rare and methodologically underexamined. Many studies prioritize model-centric evaluation metrics like accuracy or robustness without accounting for how explanations are actually perceived, trusted, or utilized by SOC analysts [6, 40]. This leaves a critical blind spot: **an explanation that looks good in theory may be impractical or even counterproductive if it does not fit analysts' cognitive workflows**. Empirical studies suggest that analysts seek features like uncertainty quantification, attribution trails, and causal narratives [35]. However, these desiderata are infrequently incorporated into real-world systems, reflecting a persistent disconnect between academic prototypes and deployment-ready solutions. Furthermore, the cognitive demands placed on analysts are seldom considered when designing explanation objectives or formats, leading to explanations that are either too abstract or too burdensome to be

[1] https://anonymous.4open.science/r/ACM-CHI-2026-BD0B/README.md

Explainability Model	Explanation of Network Log Example
Network Log Example: `Jan 15 10:41:17 host01 sshd[5021]: Failed password for invalid user 'admin' from IP 192.168.3.45 port 49538 ssh2`	
LIME [36]	Highlights keywords or features such as `"Failed password"`, `"invalid user"`, and `"admin"` as key contributors using simplified decision rules based on perturbed examples.
SHAP [28]	Quantifies feature contributions numerically, assigning high Shapley values to features like `"invalid user"`, `"admin"`, and IP address `"192.168.3.45"`, indicating their influence on the alert.
LEMNA [20]	Explains the detection using a local nonlinear approximation that highlights unusual credentials (`'admin'`) and login failures from the IP (`192.168.3.45`).
XNIDS [46]	Uses feature importance derived from neural networks specialized for network events, emphasizing repeated access attempts, unusual usernames (`'admin'`), and rare IP addresses (`192.168.3.45`).
Grad-CAM [39]	Visually emphasizes important segments of the log (e.g., `'admin'` user and IP `192.168.3.45`) through activation maps produced by neural networks.
Integrated Gradients [44]	Traces incremental contributions of log elements (username `'admin'`, phrase `"Failed password"`, IP `192.168.3.45`) directly attributing them to the alert decision.
DeepLIFT [41]	Highlights significant differences from normal activity such as `"invalid user"`, username `'admin'`, and suspicious IP address `192.168.3.45`, identifying their strong relative contribution.

Table 1: Explanations from common XAI models on a network log example.

actionable. This gap motivates a shift toward human-centered explainability approaches in security.

Recent research on SOC AI/ML tools reveals areas for improvement in alert handling and analyst workflows. Kersten et al. [24] introduced an Alert Investigation Support System for Tier-1 analysts, Mink et al. [30] found ML systems often lack context-rich, actionable explanations, and Yang et al. [47] highlighted inefficiency due to excessive irrelevant alerts. Maxam and Davis [29] underscored flexible, exploratory threat-hunting over rigid approaches, and Alahmadi et al. [5] emphasized reliable, explainable, and contextually valuable alerts. Yet none explicitly examines the explainability of AI/ML models in SOCs. Our study fills this gap by empirically investigating how model explanations impact analysts' workload, trust, and decision-making. Given SOCs' reliance on AI-driven detection, poor explanations risk undermining trust, efficiency, and security, making context-aware explanation frameworks imperative.

SOCs are fast-paced and cognitively demanding environments [5, 24, 47]. Analysts should continuously synthesize incomplete threat intelligence, evolving attack indicators, and organizational protocols under significant time pressure. In this context, the utility of an explanation depends not only on its correctness but also on its *relevance, brevity, and timing* with respect to the analyst's task [11, 43]. From a cognitive load perspective, an explanation that is too verbose or poorly timed imposes an extraneous cognitive load on the analyst's limited working memory [35]. Traditional explanation mechanisms, such as saliency maps or ranked feature contributions, often fail to meet these criteria. Prior research in human-computer interaction highlights that role-sensitive, context-aware design significantly improves decision quality under pressure [14, 15].

2.3 Human-centered SOC Practice

Modern SOCs are tiered operations with distinct decision horizons and escalation paths, see Table 2. Tier-1 analysts handle high volumes of alerts requiring rapid response, making concise summaries and clear confidence indicators critical to avoid excessive verbosity or insufficient context. Tier-2 analysts need richer narrative-driven explanations and visualizations to correlate multi-system events and identify attack paths. Tier-3 analysts require detailed, auditable explanations, including model internals and explicit feature rankings, to support retrospective analysis and compliance.

Tier	Role	Responsibilities
1	Triage	Alert validation, severity assessment, escalation
2	Incident Response	Investigation, event correlation, attribution
3	Threat Hunting	Proactive hunting, forensics, rule refinement

Table 2: SOC Analyst Tiers and Responsibilities

3 Understanding the SOC Workflow

Security Operations Centers (SOCs) are mission-critical units responsible for continuous monitoring, detection, triage, and response to cyber threats across an organization's infrastructure [5, 30]. Analysts in these centers operate under stringent cognitive, organizational, and time constraints. Analysts in these environments operate under stringent cognitive, organizational, and time constraints, meaning that an explanation's usefulness depends not only on its correctness but also on its relevance, brevity, and timing within the analyst's workflow [11, 43]. An overly verbose or poorly timed explanation can impose extra cognitive load on an analyst's limited working memory, hindering decision-making. This context sets the stage for why explanations must be tailored to the analyst's role and situation in SOC settings.

Most SOCs implement a *tiered* model of analyst roles, generally comprising three levels: Tier-1 (alert triage), Tier-2 (incident response), and Tier-3 (threat hunting and forensics). Each tier's responsibilities and information requirements differ significantly, as do the time constraints under which analysts operate [24, 30, 47].

Tier	Observed Log Patterns / Features	Key Analyst Actions and Explanation Requirements
1	• Repeated VPN login failures quickly followed by a successful login from suspicious external IP (198.51.100.23). • Alert for large data transfer from IP `198.51.100.23` to internal host `10.0.0.72`.	• **Rapid Triage**: Validate or dismiss large volumes of routine alerts. • **Immediate Severity Assessment**: Decide if repeated failures followed by success is malicious or a benign user error. • **Escalation Criteria**: If pattern is clearly suspicious (e.g., external IP not in whitelist), escalate to Tier-2. • **Explanation Needs**: Concise, high-level indicators that support quick decision-making.
2	• Same suspicious IP (198.51.100.23) identified in multiple alerts. • Correlated events from NetFlow logs showing repeated connections to internal host `10.0.0.72`.	• **Incident Response**: Correlate logs across platforms (firewall, IDS, SIEM) to confirm malicious activity. • **Preliminary Attribution**: Check threat intel feeds to see if `198.51.100.23` is tied to known campaigns. • **Cross-Platform Correlation**: Investigate whether additional hosts are affected. • **Explanation Needs**: Richer context (e.g., timeline views, causal relationships) to support deeper analysis.
3	• Potential advanced threat exploiting stolen credentials and moving data internally. • Large-scale or unusual transfers indicating exfiltration attempts.	• **Threat Hunting**: Search for hidden indicators of compromise and signs of lateral movement. • **Forensic Analysis**: Examine affected hosts (e.g., `10.0.0.72`) at a system level (disk, memory, logs). • **Rule Refinement**: Update detection signatures and machine learning models to catch similar tactics in the future. • **Explanation Needs**: Detailed, model-internal insights for hypothesis testing, plus interactive querying to explore attack vectors.

Network Logs: VPN Authentication and Network Flow Events

```
2025-04-12 14:05:31,AuthService,INFO,VPN login attempt, user=jsmith, ip=198.51.100.23, result=failure
2025-04-12 14:06:15,AuthService,INFO,VPN login attempt, user=jsmith, ip=198.51.100.23, result=success
2025-04-12 14:07:00,NetFlow,INFO,Data transfer initiated, src=198.51.100.23, dst=10.0.0.72, size=10485760 (10MB)
```

\\

Table 3: Explanations from XAI Models using Network Log Example

See Table 2 for a summary of the tiers and their responsibilities and Table 3 for the role each tier plays in threat alert management and mitigation.

Most SOCs implement a tiered analyst model comprising three levels – Tier-1 (alert triage), Tier-2 (incident response), and Tier-3 (threat hunting, forensics). Each tier has distinct responsibilities and decision horizons, as well as different time pressures and information requirements (Kersten et al.[24]; Yang et al.[47]; Mink et al.[30]). Tier-1 analysts focus on high-volume initial alert handling: validating incoming alerts, performing quick severity assessments, and deciding on immediate dismissal or escalation. They typically deal with streams of SIEM/XDR/EDR alerts and must make split-second judgments, so they require concise evidence (e.g., key indicators or confidence scores) to support rapid decisions on whether an alert is benign or malicious (Yang et al.[47]). Tier-2 analysts handle alerts escalated from Tier-1, carrying out in-depth incident investigation and correlation across multiple data sources (e.g., linking firewall logs, IDS alerts, and host events) to confirm an attack and scope its impact. They may perform preliminary attribution (checking threat intelligence to see if an attacker's IP or tactics match known campaigns) and coordinate containment. For these tasks, Tier-2 analysts benefit from richer, context-driven explanations – for instance, narrative descriptions or visual timelines that connect multi-system events and highlight causal relationships, rather than isolated indicators. Tier-3 analysts engage in proactive threat hunting and deep forensics, dealing with advanced threats (e.g., APTs, zero-day exploits, or lateral movement within the network) often after the immediate incident. Their investigations demand

high-fidelity insights into model internals and system state: they prefer detailed, auditable explanations (e.g., explicit feature rankings, anomaly rationales) and even probabilistic reasoning about detection logic. Tier-3 experts also value interactive capabilities to drill down into supporting evidence, refine detection rules, or query model behavior. In short, the level of technical detail and context needed in explanations increases from Tier-1 to Tier-3. Table 2 summarizes the three SOC tiers and their primary responsibilities.

Table 3 provides an illustrative scenario highlighting these role-based differences. In this example, a series of suspicious VPN login attempts followed by a large data transfer triggers alerts. A Tier-1 analyst would perform rapid triage – noticing the repeated login failures and subsequent success from an unfamiliar external IP – and must quickly decide if this pattern is benign (e.g. a user typo) or malicious. They might simply see a high-level flag that the external IP is not on an allowlist and then escalate the alert. A Tier-2 analyst, upon escalation, would investigate the incident more thoroughly: they could correlate the VPN alert with network flow logs showing the data transfer from that same external IP to an internal host, check if the external IP is associated with known threat actors, and determine the scope (whether other hosts were targeted). The Tier-2 explanation needs to provide this broader context – for example, linking the VPN login alert with the data transfer event and relevant threat intelligence. Finally, a Tier-3 analyst might review the case as part of threat hunting or post-incident forensics. They would delve into detailed evidence – examining host *10.0.0.72* for signs of compromise, analyzing how the credentials were obtained, and updating detection rules to catch similar behavior in the future. At this tier, the explanation system should allow exploring the model's reasoning (e.g., why it flagged those log events as suspicious) and enable interactive queries (such as asking for related anomalies in historical data). This scenario underscores how each tier focuses on different aspects of the same security event, and thus, each requires a different form of explanation support.

In practice, SOC analysts' activities are organized around a well-known incident response lifecycle that progresses through phases like alert ingestion, triage, investigation, containment, eradication, recovery, and post-incident review (Yang et al.[47]; Mink et al.[30]). Each phase imposes distinct explainability demands. During the ingestion/triage stage (typically handled by Tier-1), analysts are flooded with alerts and must act under extreme time pressure – here, concise summaries and clear confidence indicators are paramount to avoid overload or delay. In the investigation phase (Tier-2's focus), the analyst needs a deeper narrative that can connect the dots across systems and reveal attack paths, often aided by visualizations or timelines. During containment and eradication, when decisive action is taken to stop an attack, explanations should help validate that the identified threat is indeed the right one to contain – for example, by referencing historical incidents or threat intel that back up the chosen response, reducing uncertainty. Finally, in the post-incident review (Tier-3's realm), the emphasis is on comprehensive, auditable explanations that facilitate learning and compliance – e.g. a detailed breakdown of which features or correlations led the model to detect the threat, so that the organization can improve defenses and satisfy any oversight requirements. These shifting needs across the incident lifecycle mean that an explanation useful in one phase might be counterproductive in another.

Despite this clear heterogeneity in analyst needs, current tools and XAI techniques largely offer one-size-fits-all explanations. Many AI-driven security products provide static or fragmented explanation outputs that are misaligned with the analyst's role or the incident context (Rastogi et al.[35]; Baruwal Chhetri et al.[10]). This mismatch — which we refer to as explanation friction — manifests when an explanation's form or detail level is out of step with what the analyst needs at that moment. For example, a verbose textual explanation that might be appropriate for a Tier-3 forensic analyst could overwhelm a Tier-1 operator who needs a quick summary, whereas a simplistic highlight or score might be dismissed by a seasoned Tier-3 expert as lacking substance (cf. Rastogi et al. [35]). Such friction can lead to wasted time, analyst frustration, or miscalibrated trust in the AI system. The design motivation for our work is to address this gap: by recognizing the role-stratified workflows and dynamic phases of SOC operations, we aim to develop context-aware, role-adaptive explanation strategies. In doing so, the goal is to ensure that each analyst – from Tier-1 to Tier-3 – receives explanations that are appropriately granular, timely, and useful for their specific decision context, ultimately improving both trust and efficiency in AI-assisted security operations.

4 Research Methodology

Our work adopts a sequential explanatory mixed-methods design. We first fielded a questionnaire to identify high-level trends in analysts' experiences with AI-generated alerts and to quantify how explanation style influences trust and triage behaviour. We then conducted semi-structured interviews to enrich and contextualise survey findings, ultimately informing the design of our adaptive explanation framework. This section details the instruments, recruitment, procedures and analyses, following recommendations for transparency in usable-security research [21].

4.1 Survey Instrument (Round 1).

Theoretical grounding. The survey instrument was rooted in the Technology Acceptance Model (TAM) and Applied Cognitive Task Analysis (ACTA). TAM posits that perceived usefulness and ease of use drive technology adoption; ACTA provides structured techniques for eliciting decision-making strategies in complex, time-critical environments. Guided by these theories and by early literature on cognitive load, which notes that working memory can juggle only three to seven information chunks at once, we crafted items to probe how explanation complexity, context, and timing affect analysts' cognitive burden and trust. *Questionnaire.* The final questionnaire contained 35 items grouped into four sections with each item being with multiple-choice and open-ended question. Section 1 captured demographics (tier, years of experience, country, sector, and primary tools). Section 2 asked about current workflow, alert volumes, and perceived alert fatigue. Section 3 presented three explanation formats: minimal, feature-importance bar charts, and contextual narratives, and used Likert scales to measure trust, usefulness, and willingness to act. The expected time to complete the survey was 15-20 minutes.

The dashboards displayed AI-generated alerts based on real dashboard components and comprised sections that used different explanation styles; participants triaged the alert and then answered

questions about the explanation's clarity and helpfulness. Section 4 measured attitudes toward adopting adaptive explainability and included open-ended prompts. The survey was piloted with three SOC analysts, and wording and dashboard usability were refined accordingly. We incorporated two interactive dashboards into the survey via embedded screenshots (shown in Figures 13-14 of the form) to illustrate how explainable threat intelligence might be presented in a real SOC environment. These dashboards showcased features such as *feature attribution*, *confidence scores*, and *attack timelines*, prompting participants to indicate which explanation formats they found most actionable or trustworthy. Additionally, survey items explored participants' opinions on current security tools (e.g., SIEM, XDR, EDR) and any frustrations stemming from insufficient explanations or "black-box" detection logic. We also probed specific challenges in understanding AI-driven alerts, such as interpreting model outputs, confidence intervals, and contextual relevance.

By merging closed- and open-ended items, the survey provided both quantitative frequency data (e.g., how many analysts use SIEM versus XDR and average satisfaction ratings) and qualitative input (e.g., free-text descriptions of alert triage steps and user-defined preferences for particular dashboard layouts). This mixed approach helped validate and refine the concepts that informed subsequent interview protocols. The first wave of responses (N1=248) yielded a broad perspective on analysts' workflows. It demonstrated considerable interest in integrated, explanation-rich security dashboards, thus laying the groundwork for deeper investigations in follow-up interviews. The full survey instrument is provided here [2].

Pilot testing. Pilot participants highlighted question ambiguities and dashboard load times, prompting us to shorten prompts and add tooltips. The survey was hosted on Google Forms.

4.2 Survey recruitment and participation

We disseminated the survey via professional security forums, known professionals in SOC, and professional networking websites for cybersecurity practitioners. Participation was voluntary and anonymous. To reach this specialized population, we also employed snowball sampling [33], asking initial respondents and professional contacts to forward the survey invitation to other qualified SOC analysts. Eligibility required at least one year of SOC experience and regular involvement in alert triage or incident response. While snowball sampling was effective in recruiting hard-to-reach experts, we acknowledge that it could have introduced selection biases and limited representativeness (e.g., over-sampling well-networked individuals).

Respondent profile. In total, respondents represented a wide range of backgrounds: multiple continents (North America, Europe, Asia, Africa), sectors (managed security service providers (MSSPs), banking/finance, government, IT/technology), and roles spanning all SOC tiers (Tier-1, Tier-2, Tier-3), with years of SOC experience ranging from 1 to 7 years.

Ethics and data handling. The study protocol received Institutional Review Board approval (IRB). Respondents viewed an information sheet describing the purpose, procedures, voluntary participation, and data use, and could withdraw at any time. Identifiers were separated from responses. Survey data were hosted on cloud storage accessible only to the research team.

[2]https://anonymous.4open.science/r/ACM-CHI-2026-BD0B/README.md

4.3 Interview Study and Participants

. In the second stage, we conducted 24 semi-structured interviews spanning over 810 minutes with SOC professionals to delve deeper into the how and why behind the patterns observed in the survey. Interview participants were recruited through a combination of outreach on the aforementioned professional platforms and direct referrals. In some cases, survey respondents from Round 1 who indicated willingness to be contacted for follow-up were invited for an interview, complementing additional recruits obtained via snowball sampling. We applied the same inclusion criteria as for the survey (at least one year of SOC experience in an active analyst role). Our sampling strategy intentionally sought variety across key dimensions, SOC role, experience level, geography, and organizational context, to ensure a rich diversity of perspectives. Participants spanned Tier-1 alert analysts (n=9), Tier-2 incident responders (n=11), and Tier-3 threat hunters or senior analysts (n=4), working in sectors including MSSPs, banking, government, and IT, and located across Africa, Asia, Europe, and North America. Years of professional SOC experience among interviewees ranged from 1 to 7 years (median 4 years). The study was conducted over a period of three months.

4.4 Interview (Round 2)

Sampling and recruitment. Survey respondents were given the choice to opt into follow-up interviews upon completing the survey. 24 participants (10 Tier 1, 9 Tier 2, 5 Tier 3) participated in the interview, and a few had earlier participated in Round 1. Those who did not were first given the option to respond to Round 1 verbally during the Round 2 interview. Each participant in Round 2 was compensated with a token honorarium (questions summarized in Appendix).

Interview protocol. A semi-structured guide, informed by ACTA and survey insights, covered four areas: role/SOC workflow, triage and trust strategies, information needs and explanation formats, and reactions to AI-generated explanations. Participants recounted recent alert investigations and critiqued two anonymised AI-generated explanations for the same alert—one minimal, one contextual narrative. Interviews (30–70 mins, mean =34.2, median = 33) were video-recorded, transcribed, and cross-checked for accuracy. Raw data were stored in an AES-256-encrypted institutional repository. Where applicable, participants were also shown examples of synthetic LLM-generated explanations and asked to evaluate their clarity, utility, and credibility. These example explanations were generated using OpenAI latest GPT version [2] from realistic alert contexts (e.g., phishing URL triggers). Each interview lasted 30–70 minutes and was conducted via video conferencing. Interviews were audio-recorded with consent, then professionally transcribed and anonymized. All participants provided informed consent, and our study protocol was approved by the institutional review board (IRB).

4.5 Qualitative Analysis: Reflexive Thematic Coding

In the thematic coding of the interview transcripts (round 2), each interview was segmented into smaller statements focusing on (a) how analysts describe their level of trust in AI-driven alerts and (b) which explanation details they find helpful or overwhelming. Codes such as "`LowTrust`," "`HighTrust`," "`MinimalExplanation`," and "`DetailedExplanation`" were assigned to capture both trust levels and explanation complexity. Once these codes were consolidated, researchers created a table mapping complexity levels (e.g., low, medium, high) to average trust scores. For instance, an analyst who said, "I'd trust the system if it summarizes the threat quickly" was rated `HighTrust` at moderate complexity, but `LowTrust` at extreme detail. The table facilitated comparison across participants and yielded a rough average trust per complexity tier. Finally, these averages were plotted to form a "conceptual" Trust-Explainability curve, often showing that insufficient explanation led to skepticism, moderate explanation boosted confidence, and excessive technical detail provoked confusion or disuse. This study highlighted how each analyst's "sweet spot" for explanation complexity varies by experience and role, guiding future user-centric explainable AI designs. We applied reflexive thematic analysis following Braun and Clarke's approach [12]. Initially, we engaged in detailed familiarization with the transcripts, independently generating preliminary inductive codes. Subsequently, the research team discussed these initial interpretations collaboratively, merging similar codes, exploring areas of disagreement, and refining a set of coherent thematic categories through iterative discussion. This collaborative process was explicitly designed to deepen analytic reflexivity rather than achieve coding consensus or reliability, aligning with Braun and Clarke's position that coding quality in reflexive TA is not dependent on multiple coders or reliability metrics [12].

Codes were organized and managed using Excel spreadsheets, enabling systematic development. The research team iteratively clustered related codes into broader thematic groups, ensuring each emergent theme was well-supported by the data and clearly delineated from others. Regular analytic memos were maintained throughout, capturing ongoing theoretical insights, reflective observations, and justification for analytic decisions. This process remained inductively driven and grounded in participants' responses, allowing the data rather than pre-existing frameworks to shape thematic development.

The final analytic output was a set of salient themes that captured SOC analysts' nuanced experiences and perspectives on explainable AI in their workflows (see Section 5).

4.6 Methodological Rigor and Triangulation

To enhance the credibility and validity of our findings, we incorporated multiple triangulation strategies [17]:

(1) **Data:** We drew participants from multiple sectors (finance, healthcare, technology), a range of SOC roles and experience levels, different age groups and genders, and four continents to ensure that our observations were not idiosyncratic to a single context.
(2) **Method:** We cross-compared results from the qualitative interviews with the quantitative survey findings, examining where they converged or diverged. This helped confirm certain patterns (e.g., the importance of context in explanations was strongly emphasized both in survey responses and interviews) and revealed nuances where one method complemented the other.
(3) **Investigator:** Multiple researchers were involved in the data analysis process. In addition to the independent coding and cross-checking described above, co-authors not directly involved in the initial interviews reviewed the coding scheme and thematic interpretations, providing an external check on our analysis.
(4) **Theory:** We interpreted the data through multiple conceptual lenses (e.g., considering cognitive load implications as well as trust and usability perspectives) to verify that our conclusions held under different theoretical frameworks. These efforts are in line with best practices in security usability research [24, 29, 47] and helped ensure that our findings are robust and well-substantiated from several angles.

Our approach is inspired by prior security usability studies [4, 13, 37] and aligns with the methodological expectations of user studies in security and HCI research.

4.7 Ethical Considerations

Given the sensitive nature of security operations data and the professional settings of our participants, we took careful measures to protect participant privacy and abide by ethical research standards. All study participants provided informed consent after being briefed on the study's purpose and procedures. The online survey was conducted anonymously, no personally identifiable information was collected, aside from broad demographic indicators (e.g., region, sector) needed for analysis. For the interviews, all recordings and transcripts were stored in secure, access-controlled repositories accessible only to the research team. We pseudonymized the interview data: each participant was assigned a code or generic identifier (such as "Analyst-A") in all notes and reports. Transcripts were stripped of any organizational names or specific details that could reveal a participant's identity or employer. In reporting our results, we either paraphrased quotes or used generalized descriptions to convey participants' points while preserving confidentiality. These protocols, along with IRB approval and oversight, ensured that the study adhered to applicable ethical guidelines. The participants were compensated with $15 gift card for their time, and participants were informed that they could withdraw at any time without consequence.

5 Research Findings

Through interviews with SOC professionals, we uncovered several misalignments between the explainability features envisioned by designers and the realities of analysts' operational needs. These themes, supported by participants' quotes, highlight gaps in how current explainable AI (XAI) designs align with SOC workflows and address our research questions (RQ1–RQ3) explicitly.

Finding 1: Analysts Prioritize Actionable Insights Over Model Internals (RQ1, RQ2)

Analysts consistently gravitated toward **concise incident summaries, indicators of compromise (IOCs) [27], and clear recommendations for next steps, rather than detailed model reasoning or low-level feature importance**. As one analyst put

it, *"I'd focus on the indicators, like where the email came from, and then want to see immediate actions; diving into deep algorithm explanations slows me down."* Another analyst stressed, *"During triage, I ignore lengthy explanations. What I need most are straightforward next steps."*

Nonetheless, analysts recognized the value of succinct explanations that clarify why an alert was triggered. As another analyst noted, *"Quickly seeing why something was flagged, like identifying a known malicious domain, boosts my confidence considerably."*

Finding 2: Contextual Explanations Enhance Trust and Decision-Making Clarity (RQ1, RQ3)

Participants in our study provided examples reflecting this delicate balance. **Simply presenting a classification with confidence scores is not enough; analysts want to know *why* those scores make sense in context**. One SOC manager (equivalent to tier 3) explained that an explanation would be more meaningful *"if I have some context about how that percentage was generated"* (Analyst D), such as which log data or past incidents support a model's 92% confidence in a phishing alert.

In current tools, this context is often lacking; as another participant noted, a popular detection platform *"doesn't have the organizational perspective… if that is there then it is like wonders"* (Analyst A), underscoring how missing contextual information (e.g., asset value, user role, historical baselines) limits the usefulness of explanations. **Without contextual grounding, analysts are hesitant to trust AI outputs**. Thus, effective explainability design should incorporate threat intelligence and organizational context (e.g., links to similar past incidents or known bad indicators) to enhance credibility.

In our interviews, participants reacted positively when explanations included such context (for instance, referencing prior phishing campaigns or known malicious domains), because it helped them validate the alert against their environment's reality.

Finding 3: Managing Explanation Detail to Prevent Information Overload (RQ2, RQ3)

Our findings indicate that more explanation is not always better; there is a practical limit to how much detail analysts can absorb during an investigation. Current XAI prototypes often bundle numerous elements, feature importance scores, multiple visualizations, uncertainty metrics, and compliance checks, intending to be comprehensive. However, analysts reported that some of these are low priority or even distractions when an alert first comes in, so providing everything by default can be counterproductive.

In our study, participants rarely mentioned using the fine-grained feature contribution graphs or the "prediction uncertainty" fields in a real-time setting, focusing instead on high-level insights that matter for triage. One tier-2 analyst noted that they would *"look at the key analysis and immediate actions first,"* whereas things like detailed feature scores or regulatory compliance info *"wouldn't be what I look at… first"* (Analyst B). Likewise, another practitioner admitted they would gloss over secondary visual aids: the "graphical stuff" was just "OK" , nice to have, but not essential compared to core incident data (Analyst E). This gap suggests that **explainability tools should prioritize clarity over quantity**. Including every possible detail may overwhelm users, slowing them down rather than helping. Designers should instead identify which explanation components truly reduce an analyst's uncertainty or investigation time, and present those prominently, with options to drill down into more detail only if needed. Streamlining the explainability interface to align with analysts' natural triage workflow can prevent cognitive overload.

Finding 4: One Size Does Not Fit All: Experience Level and Role Matter (RQ1, RQ3)

The utility of an explanation can differ based on an analyst's experience and role. What a junior tier-1 analyst finds illuminating might feel obvious or extraneous to a senior incident responder and vice versa. Our interviews reflected this divergence. Several participants with 3+ years of SOC experience felt that certain explainability aids (like step-by-step remediation guidance or basic attack descriptions) were less relevant to them, though they acknowledged such features could be invaluable for newcomers. As one participant noted, an explainable dashboard could serve as a learning aid for junior staff by essentially *"talking to [them] like another colleague"* (Analyst D) – providing guidance that a more seasoned teammate might offer – whereas more experienced analysts preferred the system to augment their speed, not reiterate fundamentals.

Additionally, access-level differences in SOC teams mean that not everyone sees the same data, which can affect explainability needs. *"Based on their access, the information they can see changes… we want to add all of that into a summarized version [for everyone]"* (Analyst C), explained one analyst who had worked across tiered roles. This highlights that explainable outputs should be tailored – or at least adaptive – to the user's role and permissions. **A possible approach is tiered explanations: a high-level summary and key actionable items by default (useful to all), with deeper technical breakdowns available for those who need them**. Failing to account for the target user's expertise and scope of view is a misalignment, as a one-size-fits-all explanation interface will likely leave someone unsatisfied, either bored by trivial details or lost without sufficient context.

Finding 5: Calibrating Analyst Trust via the Trust-Explainability curve" (RQ1, RQ3)

Finally, our study reinforces that there is a non-linear relationship between the amount (and type) of explanation provided and the analyst's trust in the AI system. Simply adding more detail does not guarantee greater trust or satisfaction; in fact, poorly aligned explanations can backfire. We observe a conceptual Trust–Explainability curve" (Figure **??**) in SOC contexts: providing no rationale for an AI-generated alert yields low trust (analysts are uneasy acting on a black-box output), and providing a minimal but clear explanation (e.g., highlighting that an email was flagged because the sender's domain is rare and the URL was previously blacklisted) can significantly boost confidence and willingness to act. However, additional details show diminishing returns beyond a certain point and can even reduce trust if they introduce confusion or doubt.

Participants in our study provided examples reflecting this delicate balance. One analyst cautioned and was highly optimistic of explanations that if they provide extra information that is *"60%"*

accurate, it could still save their time almost 60% of the time, and we will figure out those times with experience with the tool (Analyst F). Others indicated that overly verbose or abstruse justifications might lead them to question the system's reliability or simply ignore the explanation in favor of their investigation. The sweet spot, according to our findings, is where explanations are accurate, context-rich, and directly relevant to the decision at hand. At this optimal point on the curve, analysts gain confidence in the AI's conclusions without feeling undermined or overwhelmed by details. In other words, **explanations should be transparent enough to answer the critical question "Why should I trust this alert?" yet streamlined enough to support fast decision-making**. This refined understanding of the Trust–Explainability curve, grounded in real-world input, suggests that XAI designers should carefully calibrate the depth and presentation of explanations to align with analysts' trust and efficiency needs.

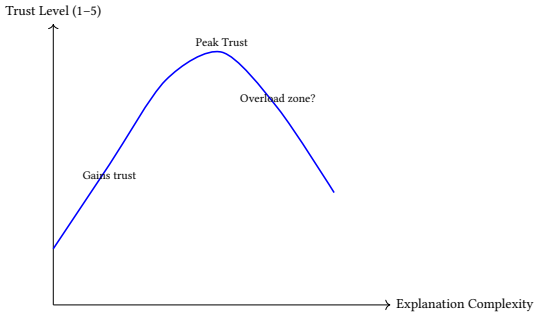


Figure 2: Trust–Explainability Curve, showing trust rising with moderate detail and eventually declining under complexity overload.

6 Design and Methodological Implications

Our findings point to a foundational misalignment between current explainability paradigms and the operational needs of Security Operations Center (SOC) environments. Existing explainable AI (XAI) approaches typically assume a singular, static user perspective and focus on post-hoc model transparency, often centered on global feature importance or local attribution scores. In contrast, SOC analysts operate within high-pressure, role-differentiated ecosystems, where the utility of an explanation is shaped by context, access level, time constraints, and incident complexity. These implications are derived directly from our empirical findings and grounded in current human-AI interaction literature, threat intelligence frameworks, and AI auditing practices.

Role-Specific Explanation Personalization. Our 24 interviews show that the explanatory needs of analysts are strongly modulated by their tier and responsibilities. Tier-1 analysts favored concise outputs like summaries, IOCs, and actions. Tier-2 analysts preferred structured, causal narratives that linked alerts across systems or stages of the kill chain, while Tier-3 analysts required transparency into model internals, uncertainty estimates, and system behavior over time. This heterogeneity is consistent with cognitive load theory and role-based system design [23, 48] and suggests that a single explanation format cannot serve all users effectively. Future systems should dynamically tailor explanation generation using user metadata (e.g., tier, experience, and access level) and support different explanatory "views" of the same alert or decision point.

Multi-Stage Attack Narratives. Current explanation methods overwhelmingly focus on single-decision justification, but our data show that SOC analysts reason across sequences of alerts. Analysts often need to understand whether multiple alerts represent separate incidents or phases of a coordinated campaign. However, this higher-order explanatory context is typically absent. We recommend extending XAI systems to generate narrative explanations that reflect multi-step adversarial behavior. Drawing on the MITRE ATT&CK framework [42], systems should infer connections between alerts and surface them as coherent explanations of attack progression. Architecturally, this could be implemented via sequence-aware encoders or graph-based clustering over alert logs, followed by natural language or visual rendering of event chains.

Contextualization through Organizational and Threat Intelligence. Trust in explanations was repeatedly linked to their ability to reflect local organizational knowledge and external threat intelligence. Participants were more confident in explanations that referenced historical behavior ("this user has never accessed this resource at this hour") or surfaced matches to previously confirmed threats ("this domain was used in a known phishing campaign"). To support such contextualization, explanation generation pipelines should incorporate data from internal security baselines and external threat feeds. Retrieval-augmented generation (RAG) [26] and context-aware embedding methods can be used to condition explanations on relevant past incidents or behavior norms, improving precision and perceived relevance.

Interactive Feedback for Explanation Alignment. We find that analysts are not passive consumers of AI outputs, they often wish to challenge, annotate, or suppress explanations. Several participants requested the ability to provide quick feedback on explanatory relevance or accuracy, both to calibrate trust and to improve future outputs. This suggests a need to integrate lightweight feedback mechanisms (e.g., thumbs-up/down ratings, "helpful/not helpful" tags, or inline flags). These can feed into preference learning or retraining loops, enabling XAI systems to adapt to user norms over time [27]. Additionally, explanation query interfaces, e.g., "Explain why this alert was prioritized over another" or "What feature made this suspicious?", can support active exploration and foster deeper engagement [25].

Provenance and Explanation Auditability. Finally, analysts in senior and reporting roles emphasized the importance of traceable and reproducible explanations. When alerts feed into compliance processes, executive briefings, or legal reports, explanations should carry provenance metadata: which model generated them, when, under what configuration, and using which training data. We recommend that XAI systems log all explanatory outputs alongside digital signatures, model version identifiers, and inference timestamps. Such metadata ensures forensic traceability and supports emerging standards in AI accountability and auditability [17].

7 CORTEX

Derived from the empirical findings in Section **??**, we now present CORTEX, CORTEX- *COntextual Role and Trust aware EXplanation* framework that dynamically adjusts the granularity and format of explanations to align with an analyst's cognitive bandwidth, the incident severity, and their investigative role. In practice, this means providing concise, high-level justifications during time-constrained, high-pressure triage moments and richer, detailed explanations during in-depth investigations. By tailoring explainability to context, CORTEX aims to support appropriate trust in the AI system, informing the analyst just enough to calibrate trust without overloading them. This adaptive strategy is grounded in theories of cognitive load, which is presenting information in digestible chunks to avoid overwhelming the user. It also aligns with the idea of situated cognition that knowledge and explanations are most meaningful when presented in the context in which they are used.

7.1 Trust-Explainability Curve

Trust-Explainability curve. Most prior studies have overlooked the realities of SOC workflows and the diversity of analyst roles [35]. Explanations are rarely evaluated across the full spectrum of analyst tiers (see Table 2) or integrated into actual SOC processes. This lack of deployment challenges limits effectiveness and can even erode trust. A context-agnostic explanation strategy might result in *miscalibrated trust*: Tier-1 analysts could place blind faith in an AI's assessment if given an authoritative-looking explanation they don't fully understand (potential *misuse* of automation), whereas Tier-3 experts might ignore using the AI's advice if the explanations seem too simplistic or irrelevant to their needs [32]. To address this gap, we introduce a structured, empirically grounded model of explanation utility, the **Trust–Explainability curve**, that explicitly captures how explanation complexity and timing affect user trust and decision efficacy across different SOC roles. This curve aims to align user trust with system reliability while considering how the content or complexity of explanations influences that trust. It posits that for each type of SOC analyst, there is an optimal level of explanation detail that maximizes the analyst's calibrated trust in the AI; *too little explanation can breed skepticism or misunderstanding, while too much can cause confusion and delay.*

Explanation friction. To describe the mismatch between an explanation's form and the analyst's immediate cognitive context, we introduce **explanation friction**. It arises when an explanation modality fails to scale with the situation's urgency or the user's expertise level. For example, a verbose textual explanation that might aid a Tier-3 forensic analyst can become an impediment to a Tier-1 triage analyst who needs to make a split-second decision. It can be measured via indicators such as prolonged alert handling times, excessive back-and-forth queries for clarification, or analyst confusion in user studies.

7.2 CORTEX Framework

7.2.1 *Core Principles.*

1. Role-Sensitivity. Building upon the tiered SOC model, CORTEX accommodates different investigative roles (e.g., Tier-1 vs. Tier-3) by delivering only the level of technical detail relevant to that role. For instance, a Tier-1 analyst sees high-level incident summaries and straightforward action recommendations, while a Tier-3 threat hunter can drill down into model weights, deep technical logs, or advanced correlation graphs.

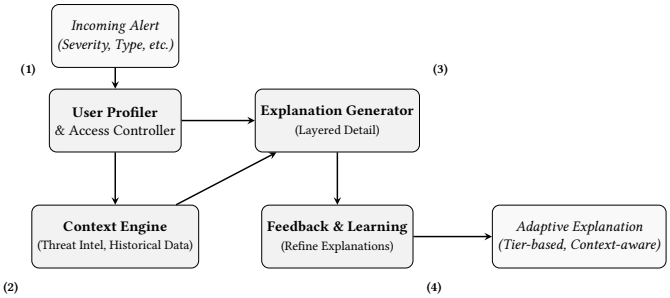


Figure 3: Architecture of the CORTEX framework, integrating user tier, context, feedback, and provenance into adaptive explanation workflows.

2. Contextual Adaptation. CORTEX dynamically adjusts its explanatory detail based on:

- *Alert Severity:* High-urgency incidents (e.g., active data exfiltration) trigger concise, actionable explanations, while lower-severity scenarios allow richer narratives and optional forensics.
- *Operational Phase:* During rapid triage, the framework offers minimal but critical cues. Later, in forensic or root-cause analyses, deeper layers (e.g., historical correlation, model provenance) become readily accessible.
- *Analyst Feedback:* The system tracks user interactions (e.g., which panels are expanded and which details are skipped) to learn personal preferences. Repeated behaviors guide the interface to emphasize or omit certain explanation layers in future alerts.

By leveraging real-time cues, CORTEX addresses *explanation friction* directly, avoiding static, one-size-fits-all descriptions that might overwhelm some analysts or under-inform others.

3. Trust-Aware Explanation Depth. Reflecting insights from our Trust-Explainability curve, the framework carefully meters how much detail is revealed at once.

CORTEX thus modulates transparency *progressively*: essential indicators (e.g., suspicious file hashes, blacklisted IP addresses) appear immediately, while deeper rationales, such as Shapley value plots [28] require an explicit role that typically requires advanced context.

7.2.2 Framework Architecture. Figure 3 illustrates CORTEX's conceptual design, consisting of four primary modules:

(1) **User Profiler and Access Controller:** Maintains information on each analyst's tier, privileges, and usage patterns. This module gates the level of detail provided, ensuring that explanations align with role-based permissions (e.g., Tier-3 can see rule-engine logs that Tier-1 cannot).
(2) **Context Engine:** Integrates external threat intelligence, historical incident data, and organizational knowledge to inform relevant references in the explanations. By grounding each alert

in real-time context, CORTEX enhances perceived credibility and fosters trust.

(3) **Explanation Generator:** Dynamically assembles explanations with graduated complexity, ranging from short text summaries to deeper causal chains. It factors in current alert severity, user preference, and the *explanation friction* thresholds to avoid overwhelming the analyst.
(4) **Feedback and Learning Module:** Observes how analysts interact with each explanation layer (e.g., which sub-panels are opened, how often advanced details are viewed). Over time, it refines default explanation levels on a per-user or per-role basis.

7.2.3 Example Workflow. To illustrate how CORTEX operates in practice, consider a Tier-1 analyst investigating a suspicious outbound connection. The system initially presents:

- A concise alert summary: "Outbound traffic to known malicious IP `198.51.100.23` flagged as suspicious."
- Immediate recommended steps: "Quarantine host `host01`; escalate for deeper investigation."

Should the analyst escalate to Tier-2 or Tier-3, the interface seamlessly expands, revealing a timeline linking the alert to prior failed login attempts, any correlated intelligence from MITRE ATT&CK patterns [42], and optional model-specific details (e.g., local feature attributions). This layered explanation strategy directly mitigates the risk of **explanation friction**, enabling time-pressed Tier-1 staff to act swiftly while affording advanced analysts a richer investigative canvas.

7.2.4 Addressing Explanation Friction and Trust Calibration. As detailed in Section 7.1, *explanation friction* can impede effective SOC operations if not carefully managed. CORTEX counters this by providing graduated forms of transparency, ensuring no single user is inundated with irrelevant details. Guided by the Trust-Explainability curve, the system aims to deliver just enough rationale to maintain appropriate trust, where an analyst neither underestimates the system's reliability nor overestimates its certainty.

8 Discussion and Recommendations for Future Research

This study re-conceptualizes explainability in Security Operations Centers (SOCs) not merely as an issue of model transparency but as a dynamic cognitive interface design challenge situated within complex, role-stratified workflows, including AI-driven decisions. Our empirical findings indicate that when explanation utility is largely integrated into AI-driven SOC environments, it should be inherently role-sensitive, context-dependent, and dynamically modulated by analysts' cognitive workload and task demands. This reframing encourages a new generation of XAI methods: not as static post-hoc interfaces but as adaptive, decision-aligned scaffolds embedded into operational workflows [3]. Below, we outline the key implications of our findings (F1- F6) and research questions (RQ1- RQ3), combining cognitive, human-centered, and system-level insights.

8.1 Reframe Explainability as Cognitive Support for Security Reasoning

Explainability in cybersecurity is most effective when it functions as cognitive scaffolding for investigative tasks such as triage, diagnosis, escalation, and containment (see Section **??**). Our findings support a shift from traditional feature attribution to task-aligned, human-centered explanation generation [9]. Rather than merely highlighting that the "failed login" has a high feature weight in a classifier, a cognitively supportive explanation should synthesize contextual cues, organizational priors, and threat mappings (e.g., STIX/TAXII [1, 8], MITRE ATT&CK [42]) to provide actionable insight:

Moving from Feature Attribution to Cognitive Explanation

Current XAI Explanation:
`Jan 15 10:41:17 host01 sshd[5021]: Failed password for invalid user admin from 192.168.3.45 port 49538 ssh2`

Cognitively Supportive Explanation (CORTEX):
- Evidence of repeated login attempts across hosts from IP **192.168.3.45**.
- Organizational knowledge that **admin** is not a valid user on **host01**.
- Correlation with a known brute-force campaign reported via external threat intelligence (e.g., MISP).

Explainability in cybersecurity is most effective when it functions as cognitive scaffolding for investigative tasks such as triage, diagnosis, escalation, and containment (see Section **??**). Our findings support a shift from traditional feature attribution to task-aligned, human-centered explanation generation [9].

8.2 Design Narrative and Multi-Stage Explanation Models

Our findings (F2, F5) show analysts prefer explanations that reveal narratives, tracing how alerts evolve across time and systems. Current XAI systems produce static outputs that lack the temporal reasoning and causal continuity essential to real-world incidents. Consider this log sequence:

Multi-Stage Attack Narrative with Causal Continuity

Current Alert Format:
Jan 15 10:41:17 host01 sshd: Failed password for admin from 192.168.3.45
Jan 15 10:42:08 host01 su: 'su root' succeeded for admin on /dev/pts/0
Jan 15 10:43:10 host01 netcat: Outbound connection to 45.76.23.91:8080

Narrative Explanation (CORTEX):
"Initial brute-force access on **host01**, followed by **privilege escalation** via su root, and culminating in potential **data exfiltration** using netcat. This sequence maps to ATT&CK T1059.003 (Command and Scripting Interpreter: Windows Command Shell)."

8.3 Operationalize Progressive Disclosure and Tier-Adaptive Interfaces

Explanations should dynamically scale with user roles and time sensitivity. Tier-1 analysts benefit from summary-level justifications (e.g., IOC matches), while Tier-3 analysts prefer drill-downs with model internals [16, 45].

- **Initial View:** Label + alert summary + threat score (supports F1, F3).
- **Expanded View:** Log timelines, asset criticality, detection rationale (supports F2, F5).
- **Expert View:** Feature attributions, model provenance, attack graph overlays (supports F4, F6).

Role-Adaptive Explanation Views

Scenario: High-severity phishing alert targeting finance department via email with suspicious attachment.

Initial View (Tier-1): `Label: Spearphishing Threat Score: 91/100 Summary: Email from external sender with known-malicious domain. Attachment auto-sandboxed. Immediate action: isolate device.`

Expanded View (Tier-2): `Log Timeline: Inbound email received at 08:32, opened at 08:34, macro execution at 08:35. Asset Criticality: Target device belongs to finance team lead. Detection Rationale: Language similarity with prior phishing templates + anomalous login sequence.`

Expert View (Tier-3): `Feature Attribution: 92% confidence from sender domain anomaly, 86% from attachment hash match. Model Provenance: XGBoost-based classifier trained on 2023 phishing corpus. ATT&CK Mapping: T1566.001 (Phishing: Spearphishing Attachment), T1204.002 (User Execution: Malicious File).`

8.4 Evaluate Explanation Utility via Cognitive and Operational Metrics

Traditional evaluation metrics (fidelity, completeness) fall short in high-stakes environments. We propose cognitive-aligned evaluation frameworks assessing:

- **Task Performance:** Reduced triage time, false positives (supports F1, F3).
- **Cognitive Load:** Minimization of unnecessary attention shifts (supports F3).
- **Trust Calibration:** Alignment of perceived vs. actual model reliability (supports F6).

Evaluating Explanation Effectiveness in a Simulated SOC

Setup: Analysts triage 10 alerts using two interfaces: one with basic explanations (baseline) and one withCORTEX's adaptive, tiered explanations.

Metrics:

- `Average triage time`: Reduced from 4.1 min (baseline) to 3.3 min (CORTEX).
- `NASA-TLX Cognitive Load`: Decreased by 18% on average forCORTEX group.
- `Trust Accuracy`: Participants usingCORTEX aligned their trust level with model confidence 82% of time, compared to 64% for baseline.

Quote from Analyst (Pilot Study): *"CORTEX let me focus on what matters-I could act fast on the alert, but still had the evidence if I needed for justification."*

9 Conclusion

This paper presents one of the first comprehensive, empirically grounded studies of explainable AI (XAI) within Security Operations Centers (SOCs), emphasizing that explanation utility is highly contingent on context, cognitive workload, and end-user (SOC analyst) role. Through a mixed method study comprising a survey of 248 security analyst respondents and 24 qualitative interviews, we identify that effective explanation must align with operational workflows and investigative depth across Tier-1 to Tier-3 SOC analysts. To address these findings, we introduce the CORTEX framework, which adapts explanation content, format, and depth to the analyst's role and situation using progressive disclosure, provenance tracking, and user feedback. Our contribution reframes explainability from a static transparency mechanism to a dynamic cognitive aid embedded in real-world SOC workflows. Our findings demonstrate that explainability in cybersecurity should be treated as a first-class interface design problem, not a generic add-on to machine learning pipelines. We advocate for an approach to XAI that is context-aware, role-sensitive, interaction-driven, and operationally grounded, transforming explanations from static monologues into adaptive, forensic, and collaborative tools for real-world defenders.

References

[1] 2022. Introduction to TAXII. https://oasis-open.github.io/cti-documentation/taxii/intro.html.

[2] 2025. *GPT-4 model OpenAI.* Available at https://chatgpt.com/.

[3] Ashraf Abdul, Jo Vermeulen, Danding Wang, and Brian Y Lim. 2020. COGAM: Measuring and moderating cognitive load in machine learning models. *CHI Conference on Human Factors in Computing Systems* (2020), 1–14.

[4] Burcu Akgul, Robert Gutzwiller, William Streilein, and Wayne G. Lutters. 2022. Exploring Cognitive Load and Decision Making in Security Operations Centers. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS).*

[5] Bushra A Alahmadi, Louise Axon, and Ivan Martinovic. 2022. 99% false positives: a qualitative study of {SOC} analysts' perspectives on security alarms. In *31st USENIX Security Symposium (USENIX Security 22).* 2783–2800.

[6] Giovanni Apruzzese, Michele Colajanni, Luca Ferretti, Alessandro Guido, and Mirco Marchetti. 2018. On the Effectiveness of Machine and Deep Learning for Cyber Security. In *2018 10th International Conference on Cyber Conflict (CyCon).* IEEE.

[7] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* 10, 7 (2015), e0130140.

[8] Sean Barnum. 2012. Standardizing cyber threat intelligence information with the structured threat information expression (stix). *Mitre Corporation* 11 (2012), 1–22.

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, and Javier Del Ser. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[10] Mohan Baruwal Chhetri, Shahroz Tariq, Ronal Singh, Fatemeh Jalalvand, Cecile Paris, and Surya Nepal. 2024. Towards Human-AI Teaming to Mitigate Alert Fatigue in Security Operations Centres. *ACM Transactions on Internet Technology* 24, 3 (2024).

[11] Dipkamal Bhusal, Rosalyn Shin, Ajay Ashok Shewale, Monish Kumar Manikya Veerabhadran, Michael Clifford, Sara Rampazzi, and Nidhi Rastogi. 2023. Sok: Modeling explainability in security analytics for interpretability, trustworthiness, and usability. In *Proceedings of the 18th International Conference on Availability, Reliability and Security.* 1–12.

[12] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.

[13] Bryan Campbell and Xin Huang. 2021. Toward a Framework for Trustworthy AI in Security Operations. In *Proceedings of the New Security Paradigms Workshop.*

[14] John M Carroll. 2003. *Making use: scenario-based design of human-computer interactions.* MIT press.

[15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1721–1730.

[16] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[17] Uwe Flick. 2018. Designing Qualitative Research. (2018).

[18] Gustavo González-Granadillo, Susana González-Zarzosa, and Rodrigo Diaz. 2021. Security information and event management (SIEM): Analysis, trends, and usage in critical infrastructures. *Sensors* 21, 14 (2021), 4759.

[19] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–42.

[20] Qian Guo, Zhi Zhang, Hongyu Hu, and Gang Wang. 2018. LEMNA: Explaining Deep Learning Based Security Applications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 364–379.

[21] Ayako A. Hasegawa, Daisuke Inoue, and Mitsuaki Akiyama. 2024. How WEIRD is Usable Privacy and Security Research?. In *33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, PA, 3241–3258. https://www.usenix.org/conference/usenixsecurity24/presentation/hasegawa

[22] Wajih Ul Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. 2019. Nodoze: Combatting threat alert fatigue with automated provenance triage. In *network and distributed systems security symposium*.

[23] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jessica Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. 1656–1660.

[24] Leon Kersten, Santiago Darré, Tom Mulders, Emmanuele Zambon, Marco Caselli, Chris Snijders, and Luca Allodi. 2024. A Security Alert Investigation Tool Supporting Tier 1 Analysts in Contextualizing and Understanding Network Security Events. In *2024 Annual Computer Security Applications Conference (ACSAC)*. IEEE, 890–905.

[25] Himabindu Lakkaraju and Ece Kamar. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *Commun. ACM* 65, 9 (2022), 92–101.

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[27] Q. Vera Liao, Daniel Gruen, and Steven Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[28] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30.

[29] William P Maxam III and James C Davis. 2024. An Interview Study on {Third-Party} Cyber Threat Hunting Processes in the {US}. Department of Homeland Security. In *33rd USENIX Security Symposium (USENIX Security 24)*. 2333–2350.

[30] Jaron Mink, Hadjer Benkraouda, Limin Yang, Arridhana Ciptadi, Ali Ahmadzadeh, Daniel Votipka, and Gang Wang. 2023. Everybody's got ML, tell me what else you have: Practitioners' perception of ML-based security tools and explanations. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2068–2085.

[31] Azqa Nadeem, Daniël Vos, Clinton Cao, Luca Pajola, Simon Dieck, Robert Baumgartner, and Sicco Verwer. 2023. Sok: Explainable machine learning for computer security applications. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 221–240.

[32] Donald A Norman. 1990. The 'problem'with automation: inappropriate feedback and interaction, not 'over-automation'. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 327, 1241 (1990), 585–593.

[33] Charlie Parker, Sam Scott, and Alistair Geddes. 2019. Snowball sampling. *SAGE research methods foundations* (2019).

[34] Nidhi Rastogi. 2022. Contextual Security: A Critical Shift in Performing Threat Intelligence. *USENIX Enigma 2022* (2022).

[35] Nidhi Rastogi, Devang Dhanuka, Amulya Saxena, Pranjal Mairal, and Le Nguyen. 2025. Survey Perspective: The Role of Explainable AI in Threat Intelligence. *arXiv preprint arXiv:2503.02065* (2025).

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[37] Scott Ruoti, Nathan Andersen, Tingsheng Wu, Daniel Zappala, and Kent E Seamons. 2016. We Are On the Same Page: A Usability Study of Secure Email Using Pairs of Novice Users. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*.

[38] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christoph Anders, and Klaus-Robert Müller. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* 109, 3 (2021), 247–278.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

[40] Vishal Sharma, Arpan Das, and V. N. Venkatakrishnan. 2020. Mimic and Classify: A Meta-Framework for Malware Analysis Using Interpretable ML. In *Network and Distributed System Security Symposium (NDSS)*.

[41] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 3145–3153.

[42] Blake E Strom, Andy Applebaum, Doug P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. 2018. Mitre ATT&CK: Design and philosophy. In *Technical report*. The MITRE Corporation.

[43] Ashley Suh, Harry Li, Caitlin Kenney, Kenneth Alperin, and Steven R. Gomez. 2024. More Questions than Answers? Lessons from Integrating Explainable AI into a Cyber-AI Tool. *arXiv preprint arXiv:2408.04746* (2024).

[44] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.

[45] John Sweller. 2011. Cognitive load theory. *Psychology of Learning and Motivation* 55 (2011), 37–76.

[46] Yifan Xu, Jun Li, Wei Wang, and Yang Liu. 2023. XNIDS: Explainable Neural Network-Based Intrusion Detection System. In *Proceedings of the 32nd USENIX Security Symposium*. 1234–1248.

[47] Limin Yang, Zhi Chen, Chenkai Wang, Zhenning Zhang, Sushruth Booma, Phuong Cao, Constantin Adam, Alexander Withers, Zbigniew Kalbarczyk, Ravishankar K Iyer, et al. 2024. True attacks, attack attempts, or benign triggers? an empirical measurement of network alerts in a security operations center. In *33rd USENIX Security Symposium (USENIX Security 24)*. 1525–1542.

[48] Ying Zhang, Gagan Bansal, Shreya Madan, and Daniel S Weld. 2020. The Effect of Explanation Type and User Expertise on the Understandability of Interpretability Methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

Appendix

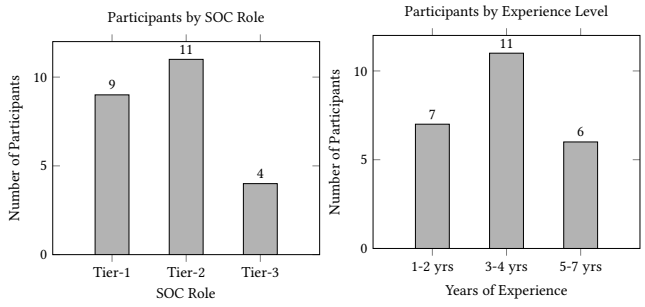


Figure 4: Distribution of study participants in Round-2 by their SOC role and experience level.

Interview Protocol (Round 2)

Participant Background

(1) Briefly describe your role, years of experience, key responsibilities, and your organization's type (e.g., enterprise, government, MSSP).

Alert Investigation and Workflow

(2) Describe your typical workflow upon receiving a security alert. How do you validate and prioritize alerts?

(3) What tools or data sources are essential in identifying genuine threats?

(4) What critical information or features are currently missing or challenging to access?

Explainability Preferences

(5) What types of explanations (confidence scores, attack attribution, feature importance) enhance your confidence when responding to AI-generated alerts?

(6) Which explanation formats (brief text, visual timelines, interactive visuals) best fit your workflow?

Dashboard Feedback (if applicable)

(7) What dashboard features do you find particularly helpful?

(8) Suggest improvements or additions that would better integrate the dashboard into your daily workflow.

Root Cause Analysis and Learning

(9) What key insights do you prioritize during root-cause analysis, and how do you document them to enhance future detection?

(10) How can explainability features further support continuous learning after incidents?

Analyst	Complexity Level	Trust Score (out of 10)	Representative Quote
Analyst 1	Medium	7	"The key analysis gives a good breakdown, immediate actions are very clear. Feature attribution might be beneficial when investigating deeper incidents but is less useful for quick triage."
Analyst 1	Low	5	"Sometimes I'm not part of the panel with detailed access, I'm constrained from accessing deeper explanations."
Analyst 2	Medium	6	"Initially, classification was manual, so explanations would have helped justify automated rules clearly."
Analyst 2	High	8	"Immediate actions and mitigation steps are highly useful; they clearly show what should be done next based on past data."
Analyst 3	High	9	"The immediate action steps are critical, they reduce panic and clarify exactly what's required next. Knowing the exploit weakness is also extremely helpful."
Analyst 3	Medium	6	"Feature attribution and prediction uncertainty might not be as directly useful in urgent responses, but could help guide deeper investigations later."
Analyst 4	Medium	7	"Immediate action box is extremely useful, speeds up mitigation significantly."
Analyst 4	Low	4	"Chatbots generally give generic responses; they're less useful unless integrated tightly with real incident histories and context."
Analyst 5	Medium	7	"Impact assessment and immediate action boxes would help quickly prioritize and act on threats."
Analyst 5	High	8	"Feature attribution clearly explains why an alert is flagged, making it easier to communicate risk."
Analyst 6	Medium	6	"Feature attribution and historical analysis are very helpful for quickly understanding repeat threats."
Analyst 6	Low	5	"Sometimes it's not clear where to start investigation, more detailed context upfront would help reduce confusion."
Analyst 7	High	9	"Immediate actions tied to specific indicators of compromise drastically speed up responses, this is what analysts need immediately."
Analyst 7	Medium	7	"The graphical timeline and visual features greatly enhance understanding of how an attack evolved."
Analyst 8	Medium	6	"Visualization of geographical logs is critical for quickly identifying anomalous login attempts."
Analyst 8	High	8	"Having clear indicators of compromise upfront greatly reduces the time spent manually correlating logs."
Analyst 9	Medium	7	"Historical contextualization is beneficial; seeing similar past alerts and how they were handled helps trust new alerts."
Analyst 9	Low	5	"Generic mitigation steps without context can be less trusted, need more specific evidence behind why these recommendations are given."
Analyst 10	High	8	"Detailed attribution at feature level (like SHAP values) allowed precise manipulation to test robustness, this boosts trust."
Analyst 10	Medium	6	"Local vs global explanations matter, global explanations are less actionable than local, immediate, scenario-specific insights."

Table 4: Interview data supporting the Trust–Explainability Curve (summarized).