Revista Científica Sistemática
ISSN: 2675-521

EXPLAINABILITY AS A TRUST INFRASTRUCTURE: XAI FRAMEWORKS FOR AI-SOAR ANALYST DECISION SUPPORT

EXPLICABILIDADE COMO INFRAESTRUTURA DE CONFIANÇA: FRAMEWORKS DE XAI PARA APOIO À DECISÃO DE ANALISTAS EM AI-SOAR

LA EXPLICABILIDAD COMO INFRAESTRUCTURA DE CONFIANZA: MARCOS DE XAI PARA EL APOYO A LA DECISIÓN DE ANALISTAS EN AI-SOAR

Data de submissão: 10/09/2024 **Data de aprovação**: 10/10/2024

https://doi.org/10.56238/rcsv14n6-010

Marcelo Araujo

ABSTRACT

The adoption of artificial intelligence in Security Operations Centers has expanded the capacity to correlate, prioritize, and triage alerts in environments integrated with SOAR platforms. However, triage automation does not eliminate a central limitation of contemporary cyber defense: analysts must understand why a given alert was escalated, why similar events were treated differently, and how the behavior of the monitored environment has changed over time. This article proposes a conceptual explainable artificial intelligence framework for AI-SOAR workflows, organized around three interpretable layers: local explanations, contrastive explanations, and temporal explanations. Local explanations support confidence for immediate action; contrastive explanations strengthen pattern recognition between malicious and benign events; and temporal explanations enhance situational awareness by interpreting sequences, baselines, and attack trajectories. As an extension, the article discusses the use of large language models as explanatory interfaces capable of translating structured XAI outputs into plain-language rationales without replacing human judgment. The article argues that explainability can function as an infrastructure for trust, auditability, and supervision in AI-SOAR escalation workflows.

Keywords: Explainable Artificial Intelligence. SOC. AI-SOAR. Analyst Trust. Cybersecurity.

RESUMO

A adoção da inteligência artificial em Centros de Operações de Segurança ampliou a capacidade de correlacionar, priorizar e triar alertas em ambientes integrados com plataformas SOAR. No entanto, a automação da triagem não elimina uma limitação central da defesa cibernética contemporânea: os analistas precisam entender por que um determinado alerta foi escalado, por que eventos semelhantes foram tratados de forma diferente e como o comportamento do ambiente monitorado mudou ao longo do tempo. Este artigo propõe um framework conceitual de inteligência artificial explicável para fluxos de trabalho de AI-SOAR, organizado em torno de três camadas interpretáveis: explicações locais, explicações contrastivas e explicações temporais. As explicações locais sustentam a confiança para a ação imediata; as explicações contrastivas fortalecem o reconhecimento de padrões entre eventos maliciosos e benignos; e as explicações temporais aprimoram a consciência situacional ao interpretar sequências, linhas de base e trajetórias de ataque. Como extensão, o artigo discute o uso de grandes modelos de linguagem como interfaces explicativas capazes de traduzir saídas estruturadas de XAI em justificativas em linguagem simples, sem substituir o julgamento humano. O artigo argumenta que a explicabilidade pode

Revista Científica Sistemática
ISSN: 2675-521

funcionar como uma infraestrutura de confiança, auditabilidade e supervisão em fluxos de escalonamento de AI-SOAR.

Palavras-chave: Inteligência Artificial Explicável. SOC. AI-SOAR. Confiança do Analista. Cibersegurança.

RESUMEN
La adopción de la inteligencia artificial en los Centros de Operaciones de Seguridad ha ampliado la capacidad de correlacionar, priorizar y triar alertas en entornos integrados con plataformas SOAR. Sin embargo, la automatización de la triagem no elimina una limitación central de la defensa cibernética contemporánea: los analistas deben comprender por qué una determinada alerta fue escalada, por qué eventos similares fueron tratados de manera diferente y cómo ha cambiado el comportamiento del entorno monitoreado a lo largo del tiempo. Este artículo propone un marco conceptual de inteligencia artificial explicable para flujos de trabajo de AI-SOAR, organizado en torno a tres capas interpretables: explicaciones locales, explicaciones contrastivas y explicaciones temporales. Las explicaciones locales respaldan la confianza para la acción inmediata; las explicaciones contrastivas fortalecen el reconocimiento de patrones entre eventos maliciosos y benignos; y las explicaciones temporales mejoran la conciencia situacional al interpretar secuencias, líneas base y trayectorias de ataque. Como extensión, el artículo analiza el uso de grandes modelos de lenguaje como interfaces explicativas capaces de traducir salidas estructuradas de XAI en justificaciones en lenguaje sencillo, sin reemplazar el juicio humano. El artículo argumenta que la explicabilidad puede funcionar como una infraestructura de confianza, auditabilidad y supervisión en los flujos de escalamiento de AI-SOAR.

Palabras clave: Inteligencia Artificial Explicable. SOC. AI-SOAR. Confianza del Analista. Ciberseguridad.

Revista Científica Sistemática
ISSN: 2675-521

1 INTRODUCTION

The burden of alert fatigue is not evenly distributed across the industry. Mid-market organizations — regional hospitals, municipal utilities, community financial institutions, and logistics operators — face the same threat landscape as large enterprises but operate without dedicated 24/7 analyst coverage, purpose-built SOC infrastructure, or the headcount required to manage alert volumes at scale. In these environments, an analyst may be simultaneously responsible for triage, escalation, compliance documentation, and incident response across multiple client organizations. The cognitive cost of operating an opaque automated system in this context is not merely an efficiency concern; it is a patient safety, service continuity, and regulatory accountability concern. For this reason, XAI frameworks designed for AI-SOAR environments must account for the resource constraints and operational realities of organizations that cannot absorb the consequences of misplaced trust in unexplained automation.

A survey of 2,000 security professionals by Vectra AI found that SOC teams receive an average of 3,832 alerts per day, with 62% ignored or left uninvestigated [1]. Separately, a global study by Trend Micro found that 70% of SOC analysts report feeling emotionally overwhelmed by alert volume [2]. Although these figures illustrate the practical scale of alert fatigue, academic research suggests that the problem is not limited to volume, but also involves analysts' calibrated trust in recommendations produced by artificial intelligence systems [3,4]. Security Operations Centers operate in an increasingly complex environment in which endpoints, corporate networks, cloud applications, identity systems, and detection tools continuously generate signals about potential threats. This expansion of telemetry increases visibility into suspicious events, but it also requires analysts to distinguish operational noise, legitimate activity, and potentially malicious behavior under time pressure.

In this context, SOAR platforms, particularly when combined with artificial intelligence models, provide important gains in the automation of repetitive tasks, alert enrichment, preliminary classification, and incident escalation. Nevertheless, automating the triage workflow does not by itself solve the trust problem. A model may correctly prioritize an alert, but the result may remain operationally weak if the analyst does not understand the factors that supported the decision. Likewise, a suppressed event may create risk if the system does not explain why it was treated as benign. For this reason, the integration of AI and SOAR should be designed as analyst decision support rather than as a replacement for human analytical judgment.

Revista Científica Sistemática
ISSN: 2675-521

Explainable artificial intelligence partially addresses this gap by making model outputs more understandable and auditable. XAI methods such as LIME and SHAP were developed to explain predictions made by complex models by identifying which variables contributed to a specific classification [5,6]. In cybersecurity, this function is especially relevant because an alert rarely depends on a single attribute. A possible account compromise may involve an unusual authentication time, atypical geographic origin, prior login failures, access to a sensitive resource, and deviation from the user's behavioral baseline. Explainability allows these elements to be presented in a structured manner, enabling analysts to understand, challenge, or validate the recommendation produced by the system.

The framework proposed in this article treats explainability as trust infrastructure. This formulation means that XAI should not be viewed merely as a visual mechanism or post-hoc transparency feature. In AI-SOAR workflows, explanations must perform an operational function: they must allow analysts to understand the recommendation, assess its consistency, identify evidentiary gaps, and decide whether to accept, enrich, investigate, or reject the escalation. Trust in this context should not be automatic. Research on trust in automation shows that both overreliance and unjustified distrust can undermine human performance in automated systems [7,8]. Therefore, the purpose of XAI is not to persuade the analyst that the machine is always correct, but to create the conditions for informed supervision.

2 THE THREE-LAYER XAI FRAMEWORK

The first layer of the framework is the local explanation layer. This layer answers the question: why was this specific alert classified in this way? Instead of displaying only a risk score, the system should identify which attributes influenced the classification, the direction of that influence, and the technical evidence supporting the recommendation. In an alert involving possible credential compromise, for example, a local explanation may indicate that the classification was influenced by an unusual login time, an uncommon IP address, prior failed authentication attempts, and subsequent access to a sensitive file. This layer addresses the cognitive need for action confidence because it gives the analyst a minimum evidentiary basis for deciding whether the event should be escalated, enriched, or closed.

The second layer is the contrastive explanation layer. While local explanation clarifies an individual case, contrastive explanation compares the alert with similar events that were treated differently. This layer answers the question: why was this event considered suspicious, while another similar event was classified as benign? This type of explanation is essential in SOCs because malicious activity often imitates legitimate behavior. PowerShell

Revista Científica Sistemática
ISSN: 2675-521

execution may be routine during administrative maintenance, but it may indicate risk when it occurs in sequence with an external download, chained script execution, privilege modification, or persistence attempts. Contrastive explanation helps analysts recognize relevant differences between events, strengthening pattern recognition and reducing dependence on isolated algorithmic scores.

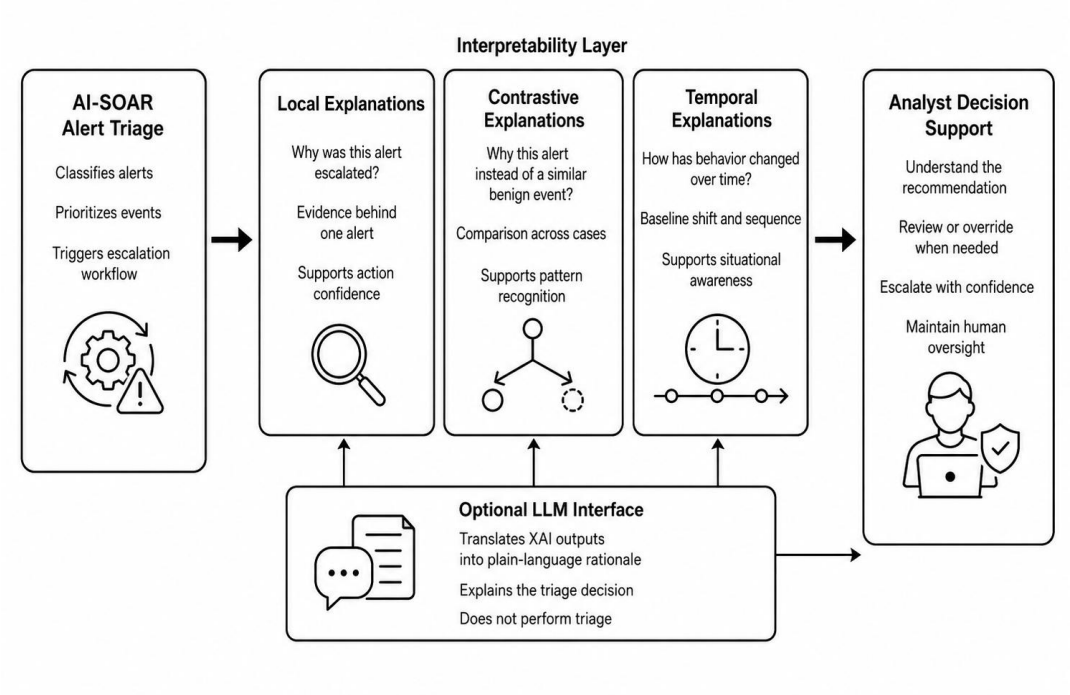
The third layer is the temporal explanation layer. Cyberattacks rarely appear as isolated events; in many cases, they unfold through sequences involving reconnaissance, exploitation, privilege escalation, lateral movement, and actions on objectives. For this reason, approaches based on attack graphs, sequential models, and kill-chain analysis are relevant for incident triage and prioritization [4,9]. Temporal explanations should show how events are connected over time, how observed behavior deviates from the baseline, and whether the sequence is compatible with an attack trajectory. This layer addresses the need for situational awareness because it shifts the analysis from an isolated alert to a broader operational narrative.

Figure 1 summarizes the proposed conceptual framework, showing how AI-SOAR triage can be connected to three layers of explainability and analyst decision support. As shown in Figure 1, local, contrastive, and temporal explanations do not replace human decision-making; they organize evidence so that the analyst can understand the recommendation, review or challenge the escalation, and maintain meaningful oversight over the automated workflow.

Revista Científica Sistemática
ISSN: 2675-521

Figura 1

Three-Layer XAI Framework for Trust in AI-SOAR Escalation Workflows.



Source: Created by author.

3 LARGE LANGUAGE MODELS AS AN EXPLANATORY LAYER

As an extension, large language models can function as an explanatory translation layer. Their role, however, must be carefully bounded. The LLM should not perform triage or decide escalation. Instead, it may transform structured XAI outputs, such as weighted attributes, correlated events, timelines, and indicators of compromise, into a clear textual rationale for the analyst [10,11]. This function may reduce the cognitive cost of interpreting feature weights and fragmented logs, provided that the explanation remains anchored in verifiable evidence.

This restriction is necessary because LLMs may produce fluent but incorrect or untraceable responses. In cybersecurity, a linguistically convincing explanation that is not aligned with the underlying data can induce operational error. Therefore, any use of LLMs in AI-SOAR should preserve auditability, traceability, and separation between explanation and decision. The ideal output should allow the analyst to inspect the original data, verify the inference chain, and identify which elements supported the escalation.

4 CONCLUSION

The maturity of SOC automation should therefore not be measured only by alert reduction or response speed. A mature AI-SOAR workflow must also explain its

Revista Científica Sistemática
ISSN: 2675-521

recommendations in a contextual, comparable, and temporally situated manner. By integrating local, contrastive, and temporal explanations, with optional support from LLMs as language interfaces, explainability can function as operational trust infrastructure. It does not eliminate the need for human judgment, but it strengthens the analyst's capacity to act safely, challenge weak recommendations, and meaningfully supervise automated systems.

It is also necessary to recognize the ceiling of automation itself. AI triage models improve with data, but their accuracy plateaus under novel attack patterns, adversarial manipulation, and behavioral drift that no training set fully anticipates [12]. The analyst is not the bottleneck to be eliminated from the workflow — the analyst is the adaptive layer that covers what the model cannot. Research has demonstrated that role-aware, context-rich XAI designs substantially improve analyst triage efficiency and confidence [13]. Explainability is what keeps that judgment calibrated. An AI-SOAR system without explainability does not augment the analyst — it removes them from the loop without acknowledging it. The goal is not to build systems that work without analysts, but systems that allow analysts to work at a scale and quality that was previously impossible.

REFERENCES

Ali T, Kostakos P. HuntGPT: integrating machine learning-based anomaly detection and explainable AI with large language models. arXiv. 2023. doi:10.48550/arXiv.2309.16021.

Chhetri MB, Tariq S, Singh R, Jalalvand F, Paris C, Nepal S. Towards Human-AI teaming to mitigate alert fatigue in Security Operations Centres. ACM Trans Internet Technol. 2024;24(3):1-22. doi:10.1145/3670009.

Habibzadeh A, Feyzi F, Atani RE. Large language models for Security Operations Centers: a comprehensive survey. arXiv. 2025. doi:10.48550/arXiv.2509.10858.

Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors. 2015;57(3):407-434. doi:10.1177/0018720814547570.

Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50-80. doi:10.1518/hfes.46.1.50_30392.

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. Red Hook: Curran Associates; 2017. p. 4765-4774.

Nadeem A, Verwer S, Moskal S, Yang SJ. Alert-driven attack graph generation using S-PDFA. IEEE Trans Dependable Secure Comput. 2022;19(2):731-746. doi:10.1109/TDSC.2021.3117348.

Rastogi N, *et al.* Too much to trust? Measuring the security and cognitive impacts of explainability in AI-driven SOCs. arXiv. 2025. doi:10.48550/arXiv.2503.02065.

Revista Científica Sistemática
ISSN: 2675-521

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

Sadlek L, Yamin MM, Celeda P, Katt B. Severity-based triage of cybersecurity incidents using kill chain attack graphs. J Inf Secur Appl. 2025;89:103956. doi:10.1016/j.jisa.2024.103956.

Sommer R, Paxson V. Outside the closed world: on using machine learning for network intrusion detection. In: 2010 IEEE Symposium on Security and Privacy. Oakland: IEEE; 2010. p. 305-316. doi:10.1109/SP.2010.25.

Trend Micro. Overworked and under-resourced: why 70% of SOC teams feel overwhelmed. Trend Micro Newsroom; 2021. Available from: https://newsroom.trendmicro.com/2021-05-25-70-Of-SOC-Teams-Emotionally-Overwhelmed-By-Security-Alert-Volume

Vectra AI. State of threat detection: SOC analyst survey. Vectra AI Research; 2023. Available from: https://www.devx.com/daily-news/soc-teams-overwhelmed-ignore-most-alerts/