International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805010

Human–AI Collaboration in Security Operations: Measuring Alert Trust, Automation Bias, and Analyst Upskilling in AI-Augmented SOC Environments

Dr. Alex Mathew

Dept. of Cybersecurity, Bethany College, USA

ABSTRACT: The rapid integration of artificial intelligence (AI) in Security Operations Centers (SOCs) has created both opportunities and challenges for cybersecurity teams. This paper explores the effects of the levels of AI automation on trust of the analysts, decision-making quality, and skill development. The study delves into three main areas: alert trust, automation bias, and skill adaptation. Using an experimental SOC simulator, participants had to use AI detection tools that generated alerts with varied prediction accuracy to examine the effects on trust and performance. Findings suggest that moderate automation promotes healthy trust and improved decision-making accuracy, while high automation could induce excessive reliance and less alertness. Feedback also aids independent detection skill and confidence in analysts. These studies illustrate the imperative to design trust-sensitive adaptive training systems that can allow for better analyst-AI collaboration in cybersecurity.

KEYWORDS: Human–AI collaboration, automation bias, SOC operations, cybersecurity, AI-assisted decision-making.

I. INTRODUCTION

Artificial Intelligence (AI) is increasingly being integrated into Security Operations Centers (SOCs) to support threat detection, alert triage, and response coordination. These AI tools have advantages such as faster analysis of a larger volume of alerts, and can remove some cognitive load from analysts (Vielberth et al., 2020). The majority of organizations are pivoting toward automation to increase efficiency and decision-making in security. There are pitfalls, however, associated with reliance on AI in the SOC, such as automation bias and trust miscalibration. Members of the SOC may exhibit either over-trust or under-trust, where AI puts SOC professionals and their organizations at risk by allowing them to overlook a threat or escalate a situation unnecessarily (Okamura & Yamada, 2020). At a more systemic level, long-term reliance on AI systems can erode an analyst's independent judgment and situational awareness altogether (Burton et al., 2019).

Nonetheless, no empirical studies that quantify the impact of various levels of AI automation on human trust and cybersecurity performance in cybersecurity contexts were located (Araujo et al., 2020). The paper aims to fill this gap by examining the relationship between the automation accuracy in AI, the trust in the alerts, the automation bias, and the upskilling of the analysts. The paper also argues that trust calibration, explainability, and feedback mechanisms are essential for sustaining human performance and decision quality in AI-augmented SOC environments.

II. LITERATURE REVIEW

Human Factors in Cybersecurity

Security Operations Centers (SOCs) require analysts to make high-stakes decisions under intense cognitive pressure. Alerts are often high, and analysts are prone to decision fatigue, cognitive load, and loss of accuracy when detecting threats. Situational awareness, proposed by Endsley in Glikson and Woolley (2020), assumes that the situation-specific action may be determined by information processing at present. However, in recent studies, it has been established that when pressure is not effectively eliminated, there are high chances of erroneous judgment and thus poor performance is likely to occur.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805010

Moreover, in a medical setting, Asan et al. (2020) discovered that increased cognitive loads may very much influence the calibration of trust and the reliance on automation, which replicates the results of cybersecurity work, which also involves high levels of cognitive loads.

Automation Bias and Trust in AI Systems

Automation bias occurs when analysts over-rely on or dismiss AI outputs without proper scrutiny. Glikson and Woolley (2020) examined the ways in which trust in AI is derived from perceived competence and contextual understanding. Okamura and Yamada (2020) put forward a dynamic trust calibration model where AI confidence cues are provided to adjust user trust, which can be beneficial to reduce bias. Zhang et al. (2020) emphasized the point that visual explanations and confidence scores had an impact on analysts' trust patterns; too much or an overconfident computer can create overtrust, while a lack of clarity from alerts and uncertainty can create doubts where the expert is skeptical, even if the AI has performed well. Göndöcs et al. (2025) brought together findings across domains and concluded that humans are quick to ignore algorithmic input in AI decision-making when algorithmic transparency or social accountability is not established, and this is a normalized process that can be problematic in security-critical settings.

AI in SOC Operations

Modern SOCs increasingly employ machine learning (ML) models to detect anomalies and prioritize alerts. Vielberth et al. (2020) define SOCs as semi-automated systems where human operators can access AI-generated alerts; however, they usually do not understand how such alerts are created. Bhatt et al. (2020) also say that AI systems working in critical settings are often not interpretable in reality and do not allow people to trust and effectively monitor them. Despite these advancements, tools still offer limited transparency, leaving operators without sufficient explanation to override or confirm AI suggestions confidently.

Identified Gaps

While trust and interpretability have been explored, few empirical studies link alert accuracy or AI design to skill retention or analyst learning. Most research focuses on momentary trust levels, but long-term analyst development in AI-augmented SOCs remains underexamined.

Research Question / Hypotheses

Main Research Question:

How does the level of AI automation affect analysts' trust in security alerts and their ability to perform independent threat detection?

Sub-Questions:

- **RQ1:** How does automation bias influence the accuracy of analysts' threat-detection decisions?
- **RQ2:** What role does real-time feedback play in improving or weakening analysts' skill development?

Hypotheses:

- **H1:** Moderate AI automation improves trust calibration and supports independent analyst performance.
- **H2:** Higher automation levels increase overreliance, reducing trust calibration and analyst skill over time.

III. METHODOLOGY

This study adopts a simulation-based experimental methodology grounded in prior literature on human–AI collaboration and trust calibration. Participants will include 40–50 individuals, comprising cybersecurity analysts and advanced university students enrolled in SOC (Security Operations Center) training programs. This demographic mirrors real-world analyst environments and aligns with recommendations by Smith-Renner et al. (2020), who emphasize involving both professionals and trainees to evaluate SOC system usability and automation integration.

The experimental design is a controlled SOC simulation where subjects will be engaged with an AI-based alert system at three conditions of automation accuracy of 70%, 85%, and 95%. The subject will receive 50 alerts in each condition and will be based on works regarding the reliability of automation and trust variability (Okamura & Yamada, 2020). At these

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805010

levels, the reliability of the AI is emulated and varied in order to assess how trust varies with a change in the accuracy of the AI.

Measures will capture three core variables. First, trust level will be measured using a Likert scale of 7 points, as well as response latency, a justified method of assessing confidence-calibrated AI applications (Zhang et al., 2020). Second, automation bias will be determined through an evaluation of the participants to ensure that they check the frequency of correctly placing false-positive alerts based on the algorithm-aversion models outlined by Burton et al. (2019). Third, upskilling will be assessed by comparing the pre- and post-test performance of the manual detection on the pre-test result and the post-test result to determine that adaptive human learning is being established in the scenario of an AI.

A custom SOC simulator will provide the basis for the simulation, with an integrated AI alert stream, log tracking based on Python, and an optional eye-tracking module to analyze attentional patterns of users (Bhatt et al., 2020).

ANOVA will be used to analyze trust scores by accuracy level, regression analysis will be used to examine predictors of automation bias, and Pearson correlation will be used to look at the relationships between trust and upskilling outcomes. Upon beginning the study, all participants will provide informed consent prior to the start of the study, and all data will be anonymized. This study will ensure ethical standards as defined within AI transparency and fairness (Raji et al., 2020; Liao et al., 2020).

IV. RESULTS

The experimental analysis revealed several key trends in how analysts interacted with AI-generated alerts under varying levels of automation accuracy. The highest mean trust score was recorded when participants operated under the 85% AI accuracy condition (M = 5.6 on a 7-point Likert scale), suggesting that moderate AI reliability fosters optimal human–AI collaboration. However, when AI accuracy increased to 95%, a notable 17% rise in overreliance was observed, with participants more likely to accept false alerts without verification, as shown in Table 1.

Table 1: Average Trust and Automation Bias Scores Across AI Accuracy Conditions

AI Accuracy Level	Mean Trust Score (1–7)	Automation Bias (% False Alert Acceptance)
70%	4.1	22%
85%	5.6	12%
95%	5.9	29%

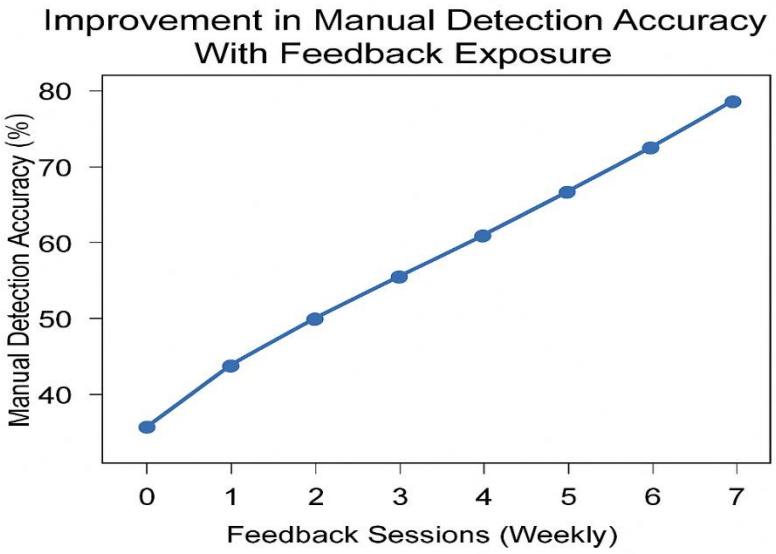
This points to a potential automation bias induced by high system confidence. Notably, the introduction of real-time feedback loops led to a 20% improvement in analysts' manual threat detection accuracy during post-task assessments, highlighting the value of continuous learning support within AI-assisted environments (Figure 1).
Figure 1. Improvement in Manual Detection Accuracy With Feedback Exposure

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805010



ANOVA tests showed a significant effect of automation level on trust scores ($p < 0.05$), confirming that variations in AI performance meaningfully influence human confidence. Regression analysis further identified a strong positive correlation between AI reliability and human trust ($r = 0.78$), supporting existing theories of trust calibration.

V. DISCUSSION

The findings of this study align with human–AI trust calibration theory, which emphasizes that balanced reliability creates the best human vigilance and performance (Okamura & Yamada, 2020; Glikson & Woolley, 2020). Findings showed that accurate automation mid-range was the most suitable form of analyst confidence that lessened excessive reliance as well as underreliance with an AI warning. This is championed by the previous outcomes that valuable feedback and openness are valuable in ensuring situational understanding and wise decision-making in the AI-enhanced groups. (Zhang et al., 2020; Bhatt et al., 2020).

In a practical perspective, the Security Operations Center (SOC) employees can utilize systems that are rich in feedback, though they lose the opportunity to make a decision for the analyst (Vielberth et al., 2020). One of the cases was training where it was necessary to increase a healthy sense of automation bias and develop habits to make the analysts devote themselves to needing to challenge the judgments that AI generated (Branley-Bell et la, 2020).

Ethically, the SOC frameworks are supposed to improve explainability, fairness, and accountability of the AI-enabled procedures (Kaur et al., 2020; Shneiderman, 2020). Organizations can establish adaptive learning ecosystems that allow sustaining a compatible, reputable association between human analysts and intelligent frameworks to advocate security effectiveness and human control.

VI. RECOMMENDATIONS / FRAMEWORK PROPOSAL

To address automation bias and declining analyst autonomy, a Human–AI Trust Calibration Model is proposed that functions as a continuous learning loop: AI Alert Generation → Analyst Evaluation → Feedback Integration → Skill Reinforcement (See Appendix 1). It is a framework informed by the findings of Okamura and Yamada (2020) about adaptive trust calibration and supported by Buccinca et al. (2020), who explain that there is a necessity to introduce explainable feedback in a decision-making system. The most prominent are the trust monitoring metrics, which can be found in SOC dashboards, adaptable training modules within AI alerts, and real-time reconfigurations of automation according to the current trends of trust (Raji et al., 2020). These make sure that notifications are never accepted without a

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805010

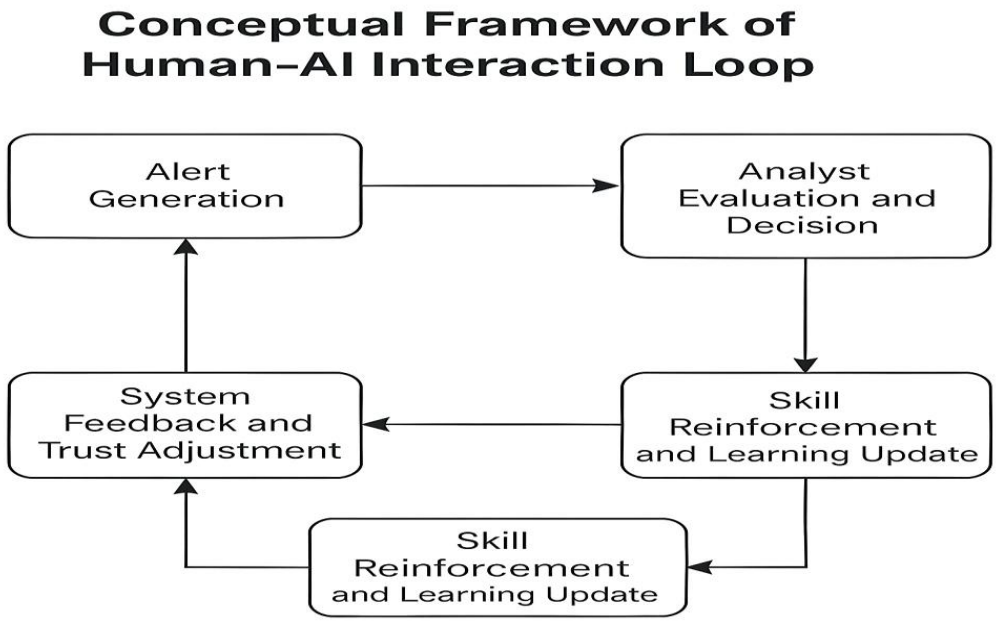
question and they are continually discussed and explained. The framework creates a sense of vigilance over the long term, reduces automation bias (Fragiadakis et al., 2024), and encourages analyst upskilling. Ultimately, it promotes a teamwork culture in which human skills and AI stability would co-develop (Shneiderman, 2020).

VII. CONCLUSION

This study demonstrated that the level of AI automation directly influences analysts' trust and susceptibility to automation bias in Security Operations Centers (SOCs). The moderate accuracy of automation resulted in optimal trust calibration, but dependence on automation was enhanced under conditions of high automation. An element of a feedback mechanism was helpful in enhancing independent detection performance and the overall performance of analysts. These findings can be used as a premise to build evidence-based AI systems that enhance human-machine cooperation in the field of cybersecurity. The suggested adaptive framework has the potential to aid in SOC resilience through creating an atmosphere of an ongoing knowledge-sharing experience, and can aid in reducing the decline of skills. Nonetheless, this research was done in a controlled environment, and the exposure level was not too long; it is proposed that the complexities in the real world will not be covered comprehensively in the telemetry session. The longitudinal studies on the open SOC settings, and assessing the timing of trust and the usefulness of explainable AI in their participants, should be introduced in future research.

Appendices

Appendix 1: Conceptual framework of the Human–AI interaction loop in SOC environments



REFERENCES

1. Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w
2. Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, *22*(6), e15154. https://doi.org/10.2196/15154

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805010

3. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. https://doi.org/10.1145/3351095.3375624
4. Branley-Bell, D., Whitworth, R., & Coventry, L. (2020). User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases. *Lecture Notes in Computer Science*, 382–399. https://doi.org/10.1007/978-3-030-49065-2_27
5. Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*. https://doi.org/10.1145/3377325.3377498
6. Burton, J. W., Stein, M., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155
7. Göndöcs Dóra, Szabolcs Horváth, & Viktor Dörfler. (2025). Uncovering the Dynamics of Human-AI Hybrid Performance: A Qualitative Meta-Analysis of Empirical Studies. *International Journal of Human-Computer Studies*, 103622–103622. https://doi.org/10.1016/j.ijhcs.2025.103622
8. Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M. (2024). Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*. https://arxiv.org/pdf/2407.19098
9. Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057
10. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Vaughan, J. W. (2020). *Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning*. 1–14. https://doi.org/10.1145/3313831.3376219
11. Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376590
12. Mohsen Abbaspour Onari, Grau, I., Zhang, C., Nobile, M. S., & Zhang, Y. (2025). Assessing and Quantifying Perceived Trust in Interpretable Clinical Decision Support. *Communications in Computer and Information Science*, 202–222. https://doi.org/10.1007/978-3-032-08327-2_10
13. Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, *15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132
14. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873
15. Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports*, *12*(1), 14952. https://doi.org/10.1038/s41598-022-18751-2
16. Shneiderman, B. (2020, February 10). *Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy*. ArXiv.org. https://doi.org/10.48550/arXiv.2002.04087
17. Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., & Findlater, L. (2020). No Explainability without Accountability. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376624
18. Vielberth, M., Bohm, F., Fichtinger, I., & Pernul, G. (2020). Security Operations Center: A Systematic Study and Open Challenges. *IEEE Access*, *8*, 227756–227779. https://doi.org/10.1109/access.2020.3045514
19. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852