Human Factors in Cybersecurity, Vol. 208, 2026, 54–63
https://doi.org/10.54941/ahfe1007412

Calibrating Trust in AI-Driven Cyber Defences: Human Reliance, Resistance, and Decision Dynamics

Vangelis Malamas[1] and Dimitris Koutras[2]

[1]Hellenic Open University, School of Social Sciences, 18 Aristotelous St., Greece
[2]Focal Point sprl, Avenue Prince d'Orange 53, Belgium

ABSTRACT

AI-supported cybersecurity tools are increasingly embedded in operational environments, yet an important question remains underexplored: how do human analysts decide when to trust, doubt, or challenge automated recommendations? While prior research addresses trust in automation broadly, studies grounded in security operations remain limited. In Security Operations Centers (SOCs), analysts process high volumes of alerts under time pressure, while automated outputs vary in reliability. These conditions influence how trust develops, but their combined effects are rarely examined systematically. This paper approaches trust as a dynamic process that evolves during real investigative work. The study adopts a mixed-method research design combining controlled experiments with qualitative analysis. Simulated SOC scenarios allow participants to interact with an AI-based alert triage tool while their behavior and interpretations are observed. Results indicate that small interface design elements—such as explanation phrasing and the frequency of high-confidence alerts—can significantly influence analyst behavior, shaping patterns of over-reliance or persistent skepticism. The findings inform design principles for AI-driven cybersecurity systems that support balanced human–AI collaboration.

Keywords: Human–AI interaction, Trust calibration, Cybersecurity operations, Decision-making, Automation reliance

INTRODUCTION

AI is becoming a standard component of modern cybersecurity operations. In Security Operations Centres (SOCs), AI-driven tools are increasingly used to support alert triage, event correlation, and prioritization in environments characterized by high volumes, time pressure, and uncertain or incomplete information. Although technical progress has largely concentrated on improving detection performance and automation coverage, the operational value of these systems ultimately depends on how effectively human analysts integrate AI recommendations into real decision workflows (Giannopoulos et al., 2025). Even in virtual environments, trust relies mainly on the human factor (Malamas et al., 2024). This raises a human factors problem that has received far less attention than the underlying technology: how analysts decide when to trust, doubt, or override AI outputs. In practice, analysts rarely treat AI recommendations as neutral "answers" (Antoniadis et al.,

Received February 5, 2026; Revised April 3, 2026; Accepted April 20, 2026; Available online July 20, 2026

2024). Instead, reliance is negotiated moment by moment, shaped by task urgency, perceived costs of mistakes, and prior interactions with the tool. When that negotiation becomes miscalibrated, two problematic patterns often emerge: over-reliance, where analysts defer to automated judgements even when cues suggest caution, and persistent resistance, where early failures or negative experiences lead to sustained scepticism even after performance improves (Hagen et al., 2025). A central limitation of many trust treatments—especially when transferred from generic decision support contexts to SOC settings—is the tendency to conceptualize trust as a relatively stable attitude. Recent works in human–AI interaction, such as (Mehrotra et al., 2024) show that "appropriate trust" is not simply higher trust, but trust aligned with the actual capabilities of the system and the situational demands. In this paper, we examine trust calibration in a setting designed to reflect realistic investigative work rather than isolated, one-off decisions. We study how reliance and resistance develop and shift as analysts interact with an AI-based alert triage tool across sequences of clean detections, borderline cases, and intentionally misleading alerts. This focus on temporal dynamics is important because early system behaviour can anchor expectations, shaping later decisions even when the underlying evidence changes. In parallel, interface-level signals that implicitly communicate certainty can accelerate over-reliance or prolong distrust, particularly when analysts operate under cognitive strain (Miedema et al., 2026). Methodologically, the study adopts a mixed approach that combines controlled experimentation—supporting repeatability and behavioural comparison—with qualitative evidence that captures how analysts interpret and justify decisions as events unfold.

From this perspective, the contributions of this paper are threefold. First, we provide empirical insight into how analyst trust in AI-driven cybersecurity tools evolves across realistic sequences of decisions. Second, we characterize behavioural patterns of over-reliance and sustained scepticism, with attention to the role of early errors, workload, and interface cues. Third, we conclude with design implications for AI-driven cyber defences that support adaptive trust calibration, promoting reliable human oversight without drifting into either automatic acceptance or chronic resistance.

RELATED WORK

AI-enabled tools increasingly shape investigative work in SOCs, yet the human factors associated with operating alongside these systems—particularly trust calibration under time pressure—remain less consolidated than the technical detection literature. A key driver of reliance behaviour in SOC environments is the scale and tempo of alerting. In (Tariq et al., 2025) the authors characterise the problem as inherently socio-technical, shaped by the analyst's workload, organisational constraints, and the cumulative effects of interacting with heterogeneous detection tools. Building on this perspective, (Chhetri et al., 2024) explicitly argue for human–AI teaming approaches to mitigate alert fatigue in SOCs. At a broader process level, recent research has examined automation and collaboration in threat detection and response ecosystems. In (Nitz et al., 2025), the authors analyse how

automation interacts with organizational workflows, privacy constraints, and information-sharing mechanisms. Although their focus extends beyond individual analyst behaviour, their findings underscore that trust calibration does not occur in isolation: automated tools are embedded within procedural, legal, and collaborative structures that shape how outputs are interpreted and acted upon. This reinforces the need to study trust within realistic operational contexts rather than abstract evaluation settings. Another active research strand concerns explainable AI (XAI) as a mechanism for fostering trust. For example, (Shoukat et al., 2025) presents an explainable intrusion detection system that integrates feature-attribution techniques to increase transparency for monitoring industrial networks. Even though their work demonstrates how explanations can enhance interpretability they also raise an unresolved question: in high-tempo triage environments, explanations may function not only as sense-making aids but also as persuasive indicators that accelerate acceptance, particularly when paired with high-confidence recommendations. Other recent studies further complicate the assumption that explanation straightforwardly improves trust calibration. In (Zhong and Yayla, 2025), the authors examine the cognitive impacts of XAI in the response to cybersecurity incidents, highlighting that explanations introduce additional interpretive demands and can reshape analyst attention and workload. These works confirm that the effectiveness of explanations depends on situational factors, including time pressure and mental fatigue, supporting the view that trust calibration is closely intertwined with cognitive resource management.

Taken together, recent literature increasingly recognizes the importance of human-AI collaboration and trust in cybersecurity operations. However, empirical studies that capture how trust evolves over time during realistic investigative work remain limited. In particular, there is a lack of behavioural evidence showing how early system performance, interface cues, and workload interact to produce sustained reliance or resistance across alert sequences. Addressing this gap motivates the mixed-method, scenario-based approach adopted in the present study.

RESEARCH APPROACHES AND STUDY DESIGN

In our research, we adopted a mixed-method approach to investigate how trust in AI-driven cybersecurity tools is calibrated during operational decision-making. This choice was motivated by the recognition that trust-related behaviours cannot be fully captured through performance metrics or self-reported attitudes alone, particularly in complex, time-constrained environments such as SOCs. Instead, understanding reliance and resistance requires observing behaviour as it unfolds, while also capturing how participants interpret and rationalize system outputs in context. Methodologically, the approach aligns with established human factors and human–AI interaction research practices that combine controlled experimentation with qualitative inquiry to study cognitive and behavioural dynamics in socio-technical systems. As discussed by the authors in (Malamas et al., 2023), there is a need to move beyond static trust measurements toward methods that reveal

how trust evolves through interaction, workload fluctuations, and exposure to system errors.

The quantitative component of our study is grounded in behavioural experimentation, enabling systematic observation of analyst decisions under comparable task conditions. This strategy allows patterns of reliance, override behaviour, and decision latency to be examined across participants and across sequences of events, supporting analysis of temporal trust dynamics rather than isolated judgements. In addition, qualitative methods are used to capture the' sense making processes of the participants as they interact with the AI system. As argued by (Zhong and Yayla, 2025), qualitative insight is particularly important in AI-assisted decision-making contexts, where explanations, confidence cues, and perceived system intent can subtly shape cognitive workload and reliance strategies.

Together, these methods support triangulation between observed behaviour and articulated reasoning, strengthening the validity of findings related to trust calibration. This frame of reference is consistent with the contemporary socio-technical perspectives on trustworthy AI, which emphasize that reliability emerges from the interaction between human operators, system behaviour, and contextual constraints rather than algorithmic performance alone (Miedema et al., 2026).

EXPERIMENTAL ENVIRONMENT

The experimental environment was designed to approximate key characteristics of the SOC investigative work while maintaining sufficient control to support systematic behavioural analysis. Rather than aiming for full operational fidelity, the environment reflects a common human factors strategy for studying complex domains: selectively reproducing task-relevant pressures, information flows, and decision constraints that are known to influence operator behaviour. Participants interacted with an AI-based alert triage interface within a simulated SOC setting. The environment presented alerts sequentially, reflecting the stream-based nature of operational triage rather than isolated decision prompts. Alerts varied in quality and ambiguity, including clearly benign cases, clearly malicious detections, and borderline or misleading instances. This mixture was intentionally constructed to avoid stable performance expectations and to create conditions under which trust calibration could plausibly shift over time. The AI system provided recommendations regarding alert prioritisation and severity, accompanied by brief explanatory elements and confidence-related cues embedded in the interface. These cues were designed to resemble common design practices in contemporary SOC tools, where confidence is often communicated implicitly through language, visual emphasis, or consistency of recommendations rather than through explicit probability values. Importantly, system behaviour was not uniformly reliable across the scenario, allowing participants to encounter both supportive and problematic recommendations during the same session.

The experimental environment was implemented to ensure consistency across participants while preserving temporal structure. All participants experienced the same alert sequence order and system behaviour, enabling comparison of reliance patterns and decision trajectories across individuals. This design choice follows established guidance for experimental studies in dynamic decision-making environments, where controlling event sequences is critical for analysing temporal effects and cumulative experience (Endsley, 2023). To minimise confounding influences, the environment excluded extraneous SOC functions such as ticket management, collaboration tools, or external intelligence feeds. This abstraction allowed participants to focus on triage and judgement without introducing variability unrelated to the research objectives. At the same time, the environment retained sufficient complexity to require sustained attention, judgement under uncertainty, and adaptive interaction with the AI system across an extended sequence of events.

Participants and Procedure

Participants were recruited to reflect the profile of users who routinely engage in analytic and decision-making tasks involving security-related information. While not all participants were professional SOC analysts, selection criteria emphasised familiarity with technical systems, alert-based interfaces, and structured decision-making under time constraints. This approach supports controlled comparison across participants while avoiding the logistical and ethical constraints associated with live operational environments. Each participant completed the study individually in a single session. At the beginning of the session, participants received a brief introduction to the experimental interface and the general task objective, namely to assess and triage alerts with the support of an AI-based recommendation system. Care was taken to avoid framing the system as highly reliable or unreliable, to minimize expectation bias.

Throughout the session, the interaction data was automatically logged, including the decision results, response times, and instances where participants diverged from the system recommendations. In addition, participants were periodically asked to verbally state their reasoning or provide brief explanations for their decisions. These verbal accounts were used to capture how participants interpreted system outputs, assessed confidence cues, and justified reliance or resistance in situ. Following completion of the alert sequence, participants took part in a short debriefing, during which they were invited to reflect on their interaction with the AI system, including moments of trust, doubt, or reassessment. These reflections were used to contextualize the observed behaviour rather than as primary outcome measures.

RESULTS

Analysis of participant behaviour revealed two dominant and recurring patterns in interaction with the AI-based triage system: progressive reliance on automated recommendations and sustained resistance following perceived system failure. These patterns did not emerge uniformly between

participants, but were shaped by the temporal structure of the interaction and the sequencing of alert outcomes.

A first group of participants exhibited a gradual shift toward increased reliance on the AI system as the task progressed. Participants increasingly accepted system outputs with minimal additional scrutiny, often describing their decisions as "obvious" or "straightforward" in later stages of the scenario. Behaviourally, this pattern was reflected in shorter decision times and a decreasing rate of divergence from system recommendations across alert sequences. In contrast, a second pattern was characterised by persistent resistance to automation. Participants who encountered misleading or erroneous system recommendations early in the task were more likely to maintain a sceptical stance throughout the session. Even when subsequent alerts were handled correctly by the system, these participants continued to manually verify recommendations and frequently overrode automated suggestions. Some participants displayed mixed strategies, alternating between acceptance and intervention depending on perceived alert criticality or ambiguity. However, once a dominant pattern became established—either reliance or resistance—it tended to stabilise, suggesting that trust calibration was path-dependent rather than continuously re-evaluated from scratch at each decision point.

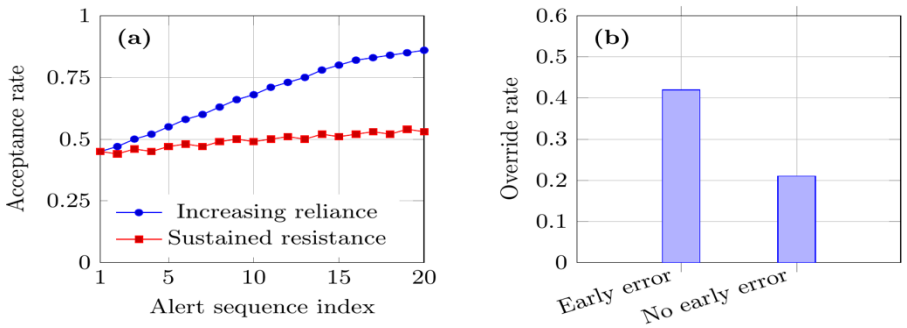

Figure 1: (a) acceptance of AI recommendations across an alert sequence; (b) override rates for participants exposed to an early misleading system recommendation.

Figure 1-a illustrates a representative trend in the acceptance of AI recommendations in alert sequences, highlighting the divergence between participants who shifted towards automation dependency and those who maintained sceptical engagement. Figure 1-b further shows how early exposure to misleading alerts was associated with elevated override rates later in the task, even when the system accuracy improved. This temporal dependency underscores the importance of examining trust calibration as an evolving behavioural process rather than as a static preference or attitude toward automation.

Temporal Dynamics of Trust

Beyond stable patterns of reliance and resistance, the behaviour of the participants exhibited clear temporal dynamics in the way trust was calibrated over the course of the interaction. Trust was not updated continuously or proportionally with each new system outcome. Instead, changes were often abrupt and strongly influenced by early or salient events, after which behaviour tended to stabilize. Participants who experienced a sequence of correct recommendations early in the task frequently transitioned into a mode of near-automatic acceptance. In contrast, participants exposed to early misleading recommendations often showed a sharp decline in trust, followed by a prolonged period of increased scrutiny that was only weakly affected by subsequent correct system behaviour. Similar effects have been observed in recent human--AI interaction research, where initial system performance acts as a reference point that biases future reliance decisions (Schemmer et al., 2023). Rather than converging towards optimal reliance over time, the behaviour of the participants in the present study suggests that the trust trajectories can become ``locked in,'' particularly under sustained cognitive load. In particular, trust recalibration was more likely to occur after high-impact failures than after gradual improvements in system accuracy. This finding aligns with the results of (Zhang et al., 2024), which show that negative trust violations carry more weight than positive confirmations in sequential decision-making contexts.

Influence of Interface Cues

In addition to system performance and temporal experience, participant behaviour was measurably influenced by interface-level cues that implicitly communicated confidence and system authority. These cues shaped reliance decisions even when underlying alert quality remained unchanged, indicating that trust calibration was sensitive to how recommendations were presented rather than solely to their correctness.

One salient factor was the formulation of the system recommendations, where participants were more likely to accept outputs framed in assertive or definitive language than those expressed with hedging or conditional phrasing. In several cases, participants referenced the system's tone when justifying acceptance decisions, suggesting that linguistic confidence acted as a heuristic for perceived reliability. This effect persisted even among participants who had previously expressed scepticism, indicating that the language of the interface could temporarily override established trust trajectories. A second influential signal related to the frequency with which the system presented high-confidence recommendations. Sequences dominated by confident output accelerated transitions toward automated acceptance. In contrast, variability in confidence presentation appeared to slow trust and encouraged more deliberate engagement. Visual emphasis and consistency also played a role. Recommendations that were visually prominent or consistently aligned with prior system outputs were more likely to be followed, even in borderline cases. Participants rarely articulated these visual factors explicitly; however, interaction logs revealed systematic differences in response behaviour corresponding to changes in interface presentation. Similar effects have been reported in recent studies of explainable AI interfaces, where presentation

choices subtly guide user attention and trust independent of explanation content (Lai et al., 2024). Importantly, interface cues interacted with temporal trust dynamics rather than acting in isolation. Participants already inclined toward reliance were especially susceptible to confident framing, while those in a resistant mode used cautious phrasing as confirmation of their scepticism. This suggests that interface cues do not uniformly increase or decrease trust, but instead modulate existing trust states, reinforcing the path-dependent patterns observed earlier.

Overall, these findings indicate that interface design choices play an active role in shaping analyst reliance strategies. Trust calibration in AI-driven cyber defences is therefore influenced not only by what the system recommends, but by how those recommendations are framed, emphasised, and repeated within the user interface.

DISCUSSION AND CONCLUSION

This study examined how trust in AI-driven cybersecurity tools is calibrated during realistic investigative work, with particular attention to reliance, resistance, temporal dynamics, and interface-mediated cues. Taken together, the findings indicate that analyst trust is neither static nor solely performance-driven, but emerges through accumulated interaction, early system experiences, and subtle design signals that shape behaviour over time. A central observation is the path-dependent nature of trust calibration. Early interactions exerted a disproportionate influence on later decision-making, often anchoring participants into sustained reliance or persistent scepticism. In high-tempo cybersecurity settings, this suggests that initial system behaviour may be as consequential as long-term accuracy (Koutras et al., 2024). The results also highlight that trust calibration is mediated by interface design choices that implicitly communicate confidence and authority. Linguistic framing, consistency of confident recommendations, and visual emphasis influenced analyst behaviour independently of alert quality. These effects did not operate uniformly across participants, but instead amplified existing trust states, reinforcing either reliance or resistance. Importantly, deviations between system performance and human decisions should not be interpreted as user error. Instead, they reflect adaptive strategies for managing uncertainty, workload, and perceived risk. In this sense, both over-reliance and sustained scepticism can be viewed as rational responses to the demands of sequential decision-making under cognitive strain. This perspective aligns with socio-technical approaches to trustworthy AI, which emphasise that effective human--AI collaboration depends on how systems support sensemaking and judgement over time, rather than on accuracy metrics alone.

The study relied on a simulated SOC environment, which necessarily abstracts away organizational pressures, collaboration dynamics, and real-world consequences of error. While this abstraction enabled controlled observation of trust dynamics, it may limit direct generalisability to live operational settings. In addition, participants represented a proxy analyst population rather than exclusively experienced SOC professionals, which may influence reliance strategies and risk perception. Future work should extend this approach to longitudinal and field-based studies, examining how

trust trajectories evolve across shifts, teams, and operational contexts, as well as how adaptive interfaces might actively support trust recalibration.

ACKNOWLEDGMENT

The research conducted in this paper was triggered by the authors' involvement in the projects: 'Collaborative, Multi-modal and Agile Professional Cybersecurity Training Program for a Skilled Workforce In the European Digital Single Market and Industries' (CyberSecPro) GA No 101083594 and "Harmonizing People, Processes, and Technology for Robust Cybersecurity" (CyberSynchrony) GA No 101158555. The views expressed in this paper represent only the views of the authors and not of the European Union nor of the partners in the above-mentioned projects.

REFERENCES

Antoniadis, D., Giannopoulos, P. G., Malamas, V., Chountalas, P. T., Pollalis, Y. A. and Dasaklis, T. (2024), The integration of artificial intelligence, big data, and iot in e-recruitment and selection: A systematic review, in 'Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics', pp. 269–274.

Chhetri, M. B., Tariq, S., Singh, R., Nepal, S. and Paris, C. (2024), 'Towards human-ai teaming to mitigate alert fatigue in security operations centres', ACM Transactions on Internet Technology 24(3), 1–22. URL: https://dl.acm.org/doi/10.1145/3670009

Endsley, M. R. (2023), Situation Awareness for Systems Design and Analysis, CRC Press, Boca Raton, FL.

Giannopoulos, P. G., Malamas, V., Verykios, V. and Dasaklis, T. K. (2025), Mitigating covariate shift in managerial decision-making: A tailored data augmentation approach for offline behavioral cloning, in '2025 16th International Conference on Information, Intelligence, Systems & Applications (IISA)', IEEE, pp. 1–8.

Hagen, R. A., Øverlier, L. and Helkala, K. (2025), 'Human factors in ai-driven cybersecurity: Cognitive biases and trust issues', Digital Threats: Research and Practice.

Koutras, D., Kioskli, K. and Kotzanikolaou, P. (2024), 'The human factor impact on a supply chain tracking service through a risk assessment methodology', Human Factors in Cybersecurity p. 198.

Lai, V., Tan, C., Smith-Renner, A. and Heer, J. (2024), 'Understanding the role of interface design in explainable ai systems', IEEE Transactions on Visualization and Computer Graphics 30(1), 612–622.

Malamas, V., Koutras, D., Dasaklis, T. K., Vassilakopouls, V. and Kotzanikolaou,P. (2024), Blockchain revolution in the metaverse: Challenges, applications and future directions, in '2024 International Conference on Artificial Intelligence, Metaverse and Cybersecurity (ICAMAC)', IEEE, pp. 1–6.

Malamas, V., Koutras, D. and Kotzanikolaou, P. (2023), Uninterrupted trust: Continuous authentication in blockchain-enhanced supply chains, in '2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)', IEEE, pp. 1–6.

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. and Tielman, M. (2024), 'A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges', ACM Journal on Responsible Computing.

Miedema, E., Waschull, S. and Emmanouilidis, C. (2026), 'Towards trustworthy artificial intelligence for decision-making: A lifecycle perspective on knowledge- and data-driven arti cial intelligence systems', Computers in Industry, 174, 104409.

Nitz, L., Gurabi, M. A., Cermak, M., Zadnik, M., Karpuk, D., Drichel, A., Schafer, S., Holmes, B. and Mandal, A. (2025), 'On collaboration and automation in the context of threat detection and response with privacy-preserving features', Digital Threats: Research and Practice 6(1), 1–36.

Schemmer, M., Kühl, N. and Goutier, M. (2023), 'The role of initial trust and experience in human-ai collaboration', Information Systems Research 34(4), 1771–1789.

Shoukat, S., Gao, T., Javeed, D., Saeed, M. S. and Adil, M. (2025), 'Trust my ids: An explainable ai integrated deep learning-based transparent threat detection system for industrial networks', Computers & Security 149, 104191. URL: https://doi.org/10.1016/j.cose.2024.104191

Tariq, S., Chhetri, M. B., Nepal, S. and Paris, C. (2025), 'Alert fatigue in security operations centres: Research challenges and opportunities', ACM Computing Surveys 57(9), 1–38. URL: https://dl.acm.org/doi/10.1145/3723158

Zhang, Y., Liao, Q. V. and Bellamy, R. K. E. (2024), 'Trust repair in human–ai interaction: The role of error severity and timing', ACM Transactions on Human-Computer Interaction 31(2), 1–28.

Zhong, C. and Yayla, A. (2025), 'Cognitive impacts of explainable ai in cybersecurity incident response: Challenges and propositions', Information Systems Frontiers . URL: https://link.springer.com/article/10.1007/s10796-025-10609-y