International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 3
DOI: https://doi.org/10.15157/IJITIS.2026.9.3.2064-2089
Received: 18.03.2026; Revised: 13.08.2026; Accepted: 18.09.2026

Research Article

Human–AI Interaction in Cybersecurity: A Theoretical Study on Cognitive Bias and Decision Reliability

Dolantina Hyka[1]*, Elva Leka[2], Luis Lamani[2], Marija Milošević Stokić[3], Ali Hassan Sodhro[4]

[1]Department of Information Technology, Mediterranean University of Albania, Tirana, Albania
[2]Department of Applied Geology & Geoinformatic, Polytechnic University of Tirana, Tirana, Albania
[3]Faculty of Electronic Engineering, University of Niš, Niš, Serbia
[4]Department of Computer Science, Kristianstad University, Kristianstad, Sweden
***dolantina.hyka@umsh.edu.al**

Abstract

Security Operations Centers (SOCs) increasingly employ artificial intelligence (AI) to detect, filter, prioritize, and interpret security alerts. However, the effectiveness of AI-assisted cybersecurity depends not only on detection accuracy but also on how analysts interpret and act upon AI-generated recommendations. This paper examines human–AI decision-making in SOCs as a socio-technical process shaped by cognitive biases and AI transparency. Particular attention is given to automation bias, algorithm aversion, and confirmation bias, as well as the roles of explanations, supporting evidence, and uncertainty communication in AI-assisted alert management. A layered conceptual framework is proposed, comprising the human layer, the AI detection layer, and the human–AI interaction layer. The framework conceptualizes decision reliability as a function of cognitive bias, explanation quality, uncertainty communication, workload, and calibrated trust. Three testable hypotheses are formulated and operationalized through measurable indicators, including unverified acceptance of AI recommendations, false dismissal rates, hypothesis revision behavior, and trust alignment. The framework is developed through a focused semi-systematic synthesis of the literature on human–AI decision-making, trust calibration, explainable AI, cognitive biases, and SOC operations. To support future empirical evaluation, a reproducible validation approach is outlined, incorporating hypothesis operationalization, planned statistical analyses, and agent-based simulation. A structured comparison with existing human–AI decision-making models and SOC-related studies is conducted to clarify the framework's contributions and limitations. The study contributes a theoretically grounded and empirically testable framework for investigating decision reliability in AI-assisted SOC environments, highlighting the interplay among cognitive bias, transparency, workload, and trust while reserving the assessment of predictive validity for future empirical research.

Keywords: Cybersecurity; Human–AI Interaction; Explainable AI; Cognitive Bias; Trust Calibration; Security Operations Centre; Decision Reliability.

INTRODUCTION

Contemporary Security Operations Centers (SOCs) increasingly rely on collaboration between machine-learning-based detection systems and human analysts. Security alerts are generated, prioritized, investigated, and interpreted through a combination of algorithmic analysis and human judgment, creating a socio-technical environment in which cybersecurity outcomes depend on both technological performance and human decision-making [1]. While AI-driven systems enhance the speed and scale of threat detection, effective cybersecurity operations require analysts to critically evaluate AI-generated recommendations rather than accept them uncritically. Consequently, the effectiveness of human–AI collaboration depends not only on the accuracy of detection algorithms but also on the quality of the decision-making processes that govern their use.

AI-based cybersecurity solutions are widely applied in intrusion detection, distributed denial-of-service (DDoS) mitigation, web application protection, and Internet of Things (IoT) security monitoring [2, 3]. Previous studies have demonstrated the effectiveness of machine learning and blockchain-enabled approaches in identifying and mitigating cybersecurity threats, including DDoS attacks and IoT-based intrusions [4–6]. However, the practical value of these technologies extends beyond their detection capabilities. Their effectiveness ultimately depends on how analysts interpret system outputs and whether they choose to trust, verify, or override AI-generated recommendations during operational decision-making [3].

Human decision-making in SOC environments is often performed under conditions characterized by high workloads, time pressure, and information uncertainty [7–9]. Under such circumstances, analysts may be influenced by cognitive biases, prior beliefs, fatigue, and limited attentional resources. Although AI systems can process large volumes of data consistently and efficiently, their outputs may contain uncertainty, limited contextual information, or explanations that are difficult to interpret. As cybersecurity technologies continue to evolve through advances in detection methods, encryption mechanisms, and automated defense strategies [10], a critical challenge remains the development of security systems that account for human cognitive processes rather than treating them as external factors. While the human factor in cybersecurity is frequently discussed in relation to user awareness and education [11], within SOC environments it is represented by professional analysts whose decisions directly influence operational outcomes.

The interaction between human cognition and AI-assisted automation constitutes the central focus of this study. In particular, the study examines how overreliance on automation may encourage analysts to accept AI-generated recommendations without sufficient verification, while excessive distrust or previous experiences with inaccurate alerts may lead to the rejection of valid system recommendations [8, 12]. These tendencies are further influenced by confirmation bias, which may encourage individuals to seek information that supports existing assumptions while discounting contradictory evidence [7]. The limited transparency of many AI systems can further amplify these

effects by making it more difficult for analysts to evaluate the reasoning underlying algorithmic outputs.

From a practical perspective, effective SOC performance requires security systems that incorporate principles of human psychology alongside technological capabilities. Successful collaboration between analysts and AI systems depends on appropriately calibrated trust, whereby system recommendations are neither accepted uncritically nor dismissed without justification. In this context, explainability, uncertainty communication, feedback mechanisms, and meaningful human oversight become essential components of reliable cybersecurity decision-making rather than optional system features.

A review of the existing literature reveals three important research gaps. First, although explanation design and trust calibration have been extensively investigated in general human–AI decision-making contexts, the relationship between transparency and workload within SOC triage processes remains insufficiently understood. Second, existing SOC-oriented frameworks addressing autonomy, oversight, and human-in-the-loop operations rarely incorporate cognitive mechanisms such as automation bias, algorithm aversion, and confirmation bias in a systematic manner. Third, these cognitive mechanisms have seldom been integrated into a unified and empirically testable framework tailored to SOC environments. To address these gaps, this study develops a conceptual framework that integrates cognitive biases with explanation design, uncertainty communication, workload, and dynamic trust. The proposed framework aims to support the systematic investigation of decision reliability in AI-assisted SOC operations and to provide a foundation for future empirical validation.

RELATED WORK

One aspect that is considered to be inherent in explainable AI (XAI) is its role as a baseline for using AI in security because the analyst has to validate the alert and provide arguments for decisions such as containment or escalation [2, 13]. However, producing an explanation is only one side of the coin. Evaluating the quality of an explanation with regard to its accuracy, usefulness, and usability is another challenge in the area of research [14-16].

When studying the interaction between a human and an AI in the process of decision making, confidence communication and explanation design influence the degree of trust people put into the system [17-19]. It has been shown experimentally that confidence and explanation design impact decision accuracy and trust calibration in AI-supported decisions [9, 19-21] and that the right design of these aspects helps avoid over-reliance on the AI [21, 22]. Trust is also shaped by broader perceptions of system trustworthiness, including reliability, transparency, accountability, and the socio-technical context in which the AI system is deployed [15, 23].

Cognitive theories of human-AI interaction provide evidence of how biases like automation bias, confirmation bias, and algorithm aversion affect the interpretation of

evidence and decision-making under uncertainty [7, 24]. In addition, studies of human-in-the-loop machine learning draw a similar conclusion: in critical settings, ML algorithms should be designed to encourage interactive verification and learning instead of being passive predictors [25].

From trust calibration literature, it is found that the reliance can be observed and adjusted using behavioral cues adaptively [26], whereas from meta-analyses, it is reported that XAI may not inherently improve the decision-making beyond mere AI-based support [27]. Later research gets more closer to operationalized Human-AI teaming. Autonomy levels and roles of humans in the loop for SOC tasks are defined by [28], while [29] experimentally demonstrate that automation transparency and decision suggestions impact accuracy, workload, trust, and usability in a non-cybersecurity task. Also, from a qualitative meta-analysis of empirical Human-AI systems, it is demonstrated that hybrid performance is determined by human-AI-task-interaction variables [30, 31]. These literature reviews highlight the significance of trust, transparency, and the importance of task context, but none of these consider workload, cognitive biases, and decision robustness together for SOC triage.

Applied cybersecurity research in turn has shown the potential of machine learning, blockchain and intelligent control mechanisms for cyber defense, such as DDoS detection and mitigation, IoT intrusion detection and resilient next-generation network infrastructures [4-6, 32, 33]. These works solidify the technical basis of AI-assisted cybersecurity, but they do not directly address how analysts cognitively engage with AI outputs under workload, uncertainty, and alert fatigue. Thus, this paper contributes to the technical literature by highlighting the human–AI decision layer.

Research Gaps and Contributions

Table 1 contrasts the current approach with the most relevant streams of previous research. It reveals a continuing dilemma regarding the balance between empiricism and SOC-specificity. While experiments in [20, 22, 26, 29] offer concrete findings on confidence, explanations, reliance, transparency, and trust calibration, these are based on activities that fall outside SOC triage and typically focus on just one element at a time. On the other hand, while the SOC-centered framework by [28] defines autonomy levels, human monitoring, and trust levels in an operational way, cognitive biases of the analysts are left out of its formal framework. What the current approach brings to the table is not necessarily that it already beats these existing models, but rather that it links them together.

The uniqueness of this research comes from the way it combines and operationalizes distinct components of human cognition, artificial intelligence, and situational factors into a coherent decision-making process in the field of cybersecurity. It does not attempt to contribute to an already large pool of XAI algorithms and trust models but rather proposes a relationship between cognitive biases, explanation quality, uncertainty cues, workload, and trust calibration and its impact on decision reliability during SOC triage. An interesting feature of the proposed framework is that it introduces the interaction terms E×W and U×W, which imply that the value of explanations and uncertainty cues might

differ depending on the workload. In other words, it goes beyond models where trust-related variables are only considered in their isolated form or as simple add-ons and makes it possible to explore their context-dependent effects. Finally, the connections proposed by this framework can be associated with concrete outcomes, such as automation bias, false dismissal, triage accuracy, and proper escalation decisions. In conclusion, the research contribution here is more conceptual and operational in nature than empirical.

Table 1. Comparison of the Proposed Model with Existing State-Of-The-Art Approaches

Study	Focus	Key features	Evaluation / quantitative basis	Biases addressed	Limitation relative to this work
[13]	XAI survey	Taxonomy of explanation methods	Narrative survey	—	Broad XAI taxonomy; no analyst-bias, workload, or SOC decision model.
[20]	Confidence and explanation in AI-assisted decisions	Case-specific confidence and local explanations	Two controlled human experiments; generic decision task	Over-/under-reliance	Shows that confidence can improve trust calibration, but the task is not SOC-specific and does not model workload or multiple biases.
[22]	Explanations and over-reliance	Cognitive-effort account of reliance	Five controlled studies; N = 731	Automation-related over-reliance	Demonstrates that verification cost matters, but does not integrate SOC workload, trust dynamics, and multiple bias mechanisms.
[25]	Human-in-the-loop ML	Taxonomy of interactive learning	State-of-the-art review	—	Human participation is structured broadly; trust calibration and SOC decision reliability are not formalized.
[26]	Adaptive trust calibration	Behavior-based detection of miscalibrated trust and calibration cues	Online drone-simulator experiment; 116 analyzed participants	Over-trust / under-trust	Provides a measurable trust-calibration mechanism, but not SOC-specific and does not combine workload with multiple cognitive biases.

[27]	Utility of XAI in human-AI decision-making	Cross-study evidence on XAI-assisted performance	Structured review identified 33 relevant articles; statistical meta-analysis	Reliance in general	Finds heterogeneous XAI effects; no SOC-specific model linking workload, bias type, and decision reliability.
[28]	SOC human-AI autonomy framework	Five autonomy levels, HITL roles, task-specific trust thresholds	2025 preprint; conceptual framework with simulated cyber-range illustration	Trust / oversight	Closest SOC-oriented comparator, but autonomy is the organizing construct; analyst cognitive biases and a joint reliability equation are not formalized.
[29]	Automation transparency and decision recommendations	Additional information, recommendations, trust, workload, usability	Controlled experiment; N = 142; 120 mission trials per participant	Misuse / disuse of automated advice	Strong human-factors evidence, but in an uninhabited-vehicle task rather than SOC triage and without a multi-bias model.
[4-6, 32, 33]	ML/blockchain-based cybersecurity and IoT/6G security	Detection, mitigation, trustworthiness, resilience	Applied cybersecurity studies / systematic review	—	Technical security mechanisms are addressed, but analyst trust, cognitive bias, and collaborative decision reliability are outside their evaluation scope.
This work	Joint bias-transparency-trust model for SOC triage	Layered model; trust surface; calibration cycle; D = f(B, E, U, W, T); H1-H3	Theoretical; explicit operationalization, planned mixed-effects tests, and agent-based validation protocol	Automation bias, algorithm aversion, confirmation bias	Integrates the constructs in one testable SOC structure, but empirical superiority is not claimed until the validation protocol is executed.

The comparison shows that the current approaches make a strong but mostly isolated contribution to human–AI cybersecurity decision-making. XAI-focused studies show that it is necessary to focus on the qualities of explanations, their transparency, and evidences which allow humans to verify them, but there is no clear link between these elements and workload and cognitive biases in SOC triage [13-15]. Studies related to human–AI decision-

making reveal more about the role of trust, confidence, reliance, and explanations, but most of them are conducted in a general environment of decision-making and not in the environment of security operations where alerts are numerous, deadlines are tight, and false alarms influence the analysts' behavior [19-22]. Thus, they describe the mechanisms of human–AI reliance but do not form a complete framework for SOC.

Another avenue of research is related to adaptive calibration of trust and human-in-the-loop learning. Authors in [26] show the possibility of trust calibration based on behavior of users, while in [25] discuss the ways to consider the role of human in interactive machine learning process. Both approaches can be considered essential because they make it possible to avoid static assumptions about the level of trust. However, neither of them considers the impact of workload, multiple cognitive biases, explanation quality, and uncertainty communication on decision reliability. Also, according to the meta-analysis conducted by [27], the effect of XAI is heterogeneous. Therefore, explanations are not effective in improving decisions in general cases. This information is especially relevant in the case of SOC, where the use of explanations is time- and cognition-dependent.

In the more recent research on SOC, the conversation gets a lot closer to the practicalities of cybersecurity operations. The study in [28] develop autonomy categories for human–AI interactions, the role of humans in the loop, and trust issues related to the specific tasks. The focus, however, is again on autonomy, whereas automation bias, algorithm aversion, and confirmation bias do not get a place in the model of decision-making process. Authors in [29] conduct an experiment that proves that transparency and automatic recommendations can affect trust, workload, usability, and decision-making process. But this study does not happen within the realm of cybersecurity, which means that the experiment does not take into account the specific impact of alert fatigue, false positives, escalation decisions, and accuracy of security triage. Applied studies in cybersecurity [4–6, 32, 33] prove the importance of intelligent detection, mitigation and resilient security mechanisms, but the target of their evaluation is system-level security outcomes and not human-AI cognitive interaction.

As a result, the state of the art shows the existence of structural gap in relation to the issue at hand rather than the absence of relevant research. Current research describes separate aspects of the problem, such as explanation quality, trust, workload, cognitive biases, human oversight, and cybersecurity threat detection, yet there are few studies that analyze these elements in the context of their being part of one decision-making system of SOC triage. Contribution of the proposed framework is thus integrative in nature and practical rather than an attempt to outperform existing models empirically. In particular, the framework enables to examine and then test if the performance of explanations and uncertainty cues is affected by workload and if this affects over-reliance, false dismissals, hypothesis revision, and escalation.

The following Figure 1 highlights the conceptual positioning of the framework through the integration of the key dimensions found in current approaches into a single socio-technical cybersecurity decision-making framework.

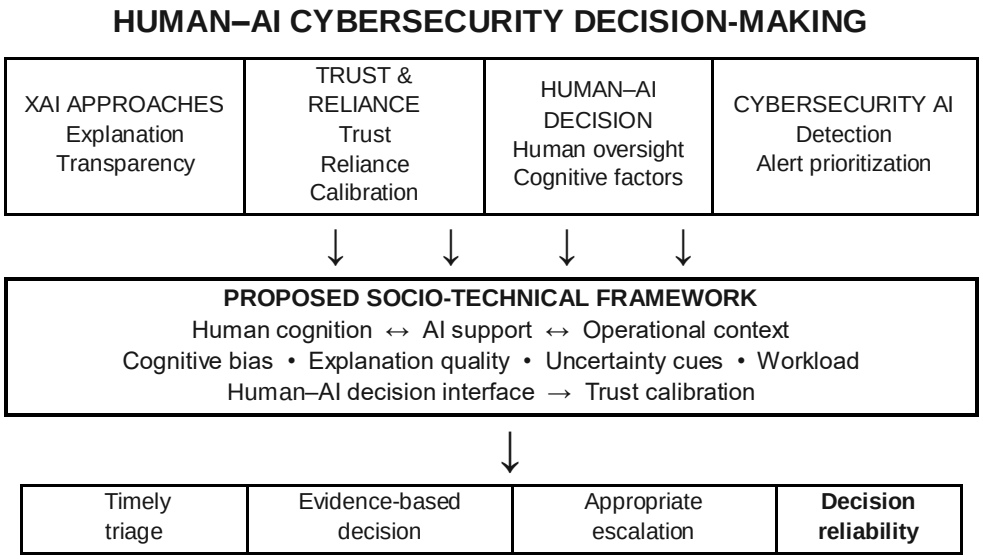

Figure 1. Conceptual Comparison of Existing Human–AI Approaches and the Proposed Socio-Technical Cybersecurity Framework.

The reason why this context-dependent approach is important is that the same AI help can yield different results depending on the circumstances under which it is employed. An explanation that would work well in case an analyst has enough time to analyze the evidence can work poorly in case of high workload and time pressure. The same thing can happen with the uncertainty information as well it might lead the analyst to show proper caution in certain situations and be ignored in others.

Additionally, this positioning highlights the capabilities that this proposed approach offers beyond those provided by the reviewed techniques when taken together. Instead of viewing trust calibration as a stand-alone characteristic of the AI system and explanation as a general good user interface design principle, this framework considers reliance as a situational decision process. The workload becomes the situational condition that impacts the effectiveness of transparency, whereas cognitive bias serves as the behavioural channel for this impact on decision-making. This framework thus enables future empirical research not only to determine if an analyst trusts the AI system, but if such trust is calibrated to the reliability of the AI system and leads to effective security decisions.

Furthermore, the quantitative data casts doubt on another typical assumption that underlies most XAI research: more explanation is not necessarily better. Specifically, the study at [22] demonstrate in five different experiments (N = 731) that over-reliance alters the effort needed to verify an AI's recommendation, while in [27] meta-analysis of numerous studies finds heterogeneity of the benefit and absence of any clear superiority of explanations over AI prediction accuracy. This point is critical for SOC analysis as an explanation may be very informative from the theoretical perspective but practically irrelevant due to the lack of time for its analysis by the analyst. Thus, in the proposed

framework, E and U are modeled as the elements whose impact might be contingent on workload W.

Although the literature extends the scope of the problem, the problem itself is still not solved yet. Namely, surveys of human-centric XAI suggest to focus on linking explanation evaluation and cognitive/social factors [34], self-confidence calibration was shown to affect appropriate reliance on AI [35], and a systematic qualitative analysis of 206 empirical studies on human/AI interaction identified interacting human, AI, task, and interaction factors [30]. Although these are significant findings, they have a general nature. In terms of SOC triage, the question remains open about the role of workload, explanation, uncertainty, trust, and analyst biases in making observable security decisions such as acceptance, false dismissal, and escalation.

METHODOLOGY

The present study adopts a theoretical and conceptual research design where therefore no empirical data are collected or analyzed. The proposed framework is developed through a focused synthesis of the relevant literature, followed by construct mapping and the formulation of a prospective validation strategy. The literature synthesis serves primarily to identify, critically examine, and integrate established concepts related to human–AI decision-making, cognitive bias, trust calibration, explainable artificial intelligence, uncertainty communication, workload, and SOC operations.

The objective of the literature synthesis is theoretical development and critical integration rather than the production of an exhaustive systematic review. Accordingly, the reviewed literature is used to establish the conceptual foundations of the proposed framework, identify relationships among its principal constructs, and derive testable hypotheses and corresponding indicators. The resulting framework provides a structured basis for future empirical investigation and validation in operational SOC environments.

Literature Selection and Synthesis

The evidentiary basis is the scholarly literature corpus cited in the manuscript and has been structured around five interrelated themes: (i) trust calibration and appropriate reliance; (ii) explainability and explanation quality; (iii) confidence and uncertainty communication; (iv) cognitive bias in human-AI decision making; and (v) cybersecurity, SOC, and human-in-the-loop operation. Only those works have been retained in the conceptual synthesis which directly contributed to a model construct (B, E, U, W, T, or D), a relationship between two constructs or the context of SOC decision-making. Empirical studies, systematic reviews, and meta-analyses were preferred when making a claim regarding human behavior. SOC specific conceptual contributions have been retained only for the purpose of defining domain, while non-peer-reviewed publications are cited as such and not treated as empirical evidence. Technical detection methods [4-6, 32] have been utilized to establish the cyber security context but not cognitive mechanisms.

As part of the present study, the selected literature was systematically compared using a common analytical framework comprising the research domain, study design, quantitative or qualitative orientation, human–AI construct of interest, cognitive bias examined, and limitations relevant to SOC triage. This analytical framework is summarized in Table 1. The comparison was conducted to assess the theoretical basis of the constructs incorporated into the proposed framework and to identify relationships and combinations that remain insufficiently specified in the existing literature. Particular attention was given to studies addressing explanations and confidence, which were examined separately from the literature concerning SOC autonomy, oversight, and human-in-the-loop processes. This distinction was maintained to ensure that the proposed contribution was not derived solely from overlapping research domains, but instead emerged from the integration of complementary strands of literature that have been insufficiently connected in the context of SOC decision-making.

Framework Derivation

The proposed framework was derived through three stages. First, recurrent human–AI failure mechanisms identified in the literature were mapped across three analytical levels: the analyst level, encompassing automation bias, algorithm aversion, confirmation bias, workload, and fatigue; the AI/interface level, encompassing explanation quality and uncertainty communication; and the relational level, encompassing trust calibration. Second, these concepts were operationally organized into the principal constructs B, E, U, W, and T, with decision reliability D defined as the principal outcome construct. Third, the directional relationships identified in the literature were formalized as propositions (P1)–(P3) and subsequently translated into testable hypotheses (H1)–(H3), together with corresponding observable indicators.

The directional signs specified in the first-order model represent theoretically grounded expectations derived from the reviewed literature rather than empirically estimated coefficients. Accordingly, the mathematical formulation serves to specify the relationships and assumptions that can be subjected to future empirical testing; it should not be interpreted as evidence of empirical validation. The proposed framework therefore provides a formal theoretical structure for subsequent validation rather than claiming empirical confirmation of the hypothesized relationships.

The Reproducibility of the Conceptual Synthesis

To enable reproducible updating of the evidence base, future searches should include various combinations of the key terms "human-AI" or "human-automation" along with "trust calibration," "appropriate reliance," "explainable AI," "automation bias," "algorithm aversion," "confirmation bias," "workload," "security operations center/SOC," "alert triage," and "human-in-the-loop." Updated papers should be added to the extraction frame utilized in Table 1 before the frame itself is updated. Such an approach allows tracing updates while recognizing that the current semi-systematic synthesis cannot encompass all relevant literature.

PROBLEM FORMULATION

This problem stems from the partnership of two suboptimal decision-makers: an analyst operating within certain limitations of the cognitive and operational capacity and an artificial intelligence system that may be correct but unsure or vague or hard to understand. The former, under conditions of high alert volume and tight deadlines, may fall victim to automation bias and follow the recommendations of the latter without proper justification. Frequent mistakes will cause the reverse reaction, that is, algorithm aversion and dismissal of future alerts due to them. Confirmation bias will cause limited consideration of data in favor of a hypothesis already suggested. In case the system gives a forecast without providing any proof or information about its uncertainty, there will be nothing for an analyst to go on.

It leads to the following research question:

Research Question: How do cognitive biases and AI transparency (evidence, explanations, and cues of confidence/uncertainty) interact to influence the analysts' trust and decision-making, measured through triage accuracy, false dismissal rate, and escalation appropriateness in SOC triage and incident response?

Figure 2 summarizes the main biases that shape how analysts rely on AI outputs in these workflows.

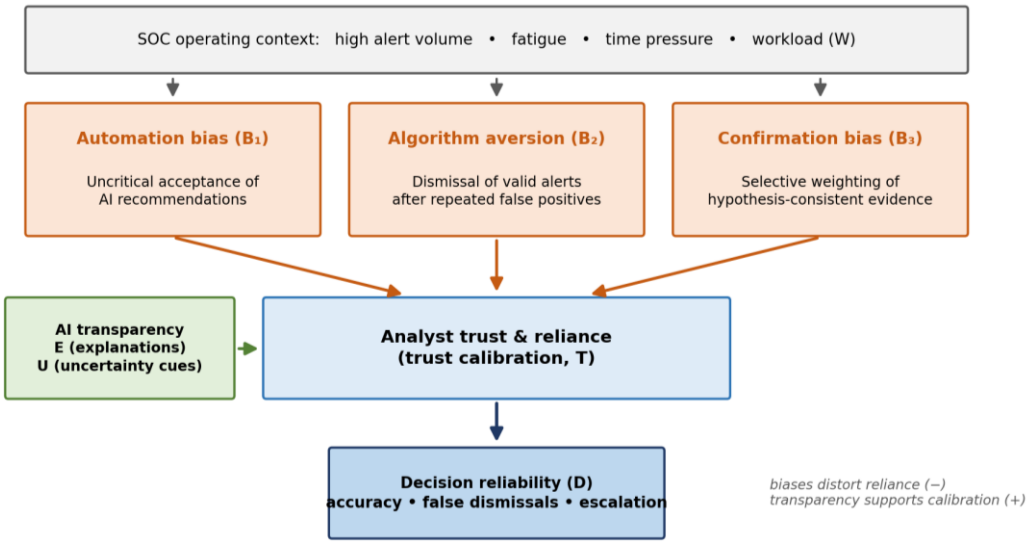

Figure 2. Cognitive Biases Influencing Analyst Trust and Decision Reliability in SOC Environments

THEORETICAL FRAMEWORK

From this perspective, SOC decision-making can be understood as a socio-technical process in which outcomes emerge from the interaction between human cognition and machine-based threat detection rather than from either component operating independently. Accordingly, cybersecurity effectiveness depends not only on the

performance of the AI-based detection system or the expertise of the analyst, but also on the quality of coordination between them through the human–AI interface.

At the human level, analyst decision-making is informed by domain expertise and situational awareness, while also being influenced by operational factors such as workload, fatigue, and time pressure. These conditions may contribute to established cognitive biases and affect how information is interpreted, evaluated, and acted upon, including decisions concerning alert escalation [24]. At the machine level, AI-based detection systems aggregate telemetry from multiple sources, prioritize potential threats, and estimate the likelihood that detected events represent genuine security incidents. Where supported by the system design, these outputs may also include explanations and confidence estimates intended to facilitate analyst verification and interpretation [13, 14].

At the human–AI interface, machine-generated outputs are translated into information that supports operational decision-making. The interface determines how alerts, findings, confidence estimates, and recommendations are presented to analysts, as well as the extent to which analysts can challenge, reject, annotate, or provide feedback on system outputs. Because trust in the AI system is formed and adjusted through this interaction, the interface is conceptualized in this study as the trust surface of the human–AI system. Figure 3 illustrates these components and their bidirectional interactions.

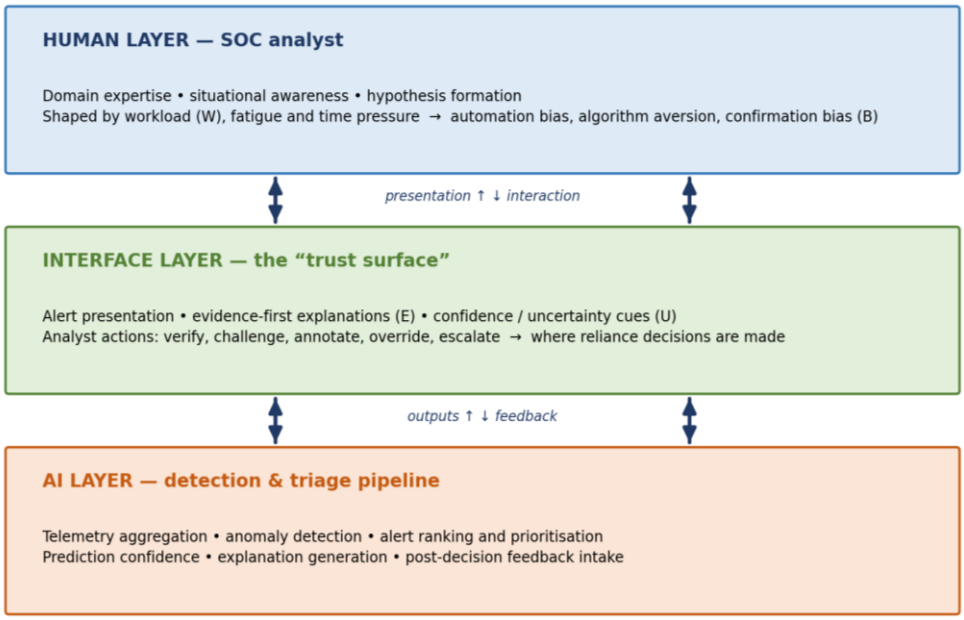

Figure 3. Socio-Technical Layers of Human–AI Decision-Making in Cybersecurity Operations

Trust calibration refers to the ongoing process in which the analyst modifies his level of trust for the system based on its performance, explanation, uncertainty, and situational requirements [19, 20, 22, 26]. The trust calibration process will be effective if the system makes strong evidence available to the user easily, conveys uncertainty clearly [8, 19], considers the alternative hypothesis, and gives feedback to the decision [1, 3, 17, 25]. This cycle has been illustrated in Figure 4. Here the system creates an alert signal, the analyst

tries to interpret it in the context, and the feedback from his process of verification, overriding, escalation, and outcome will be used for future reliance.

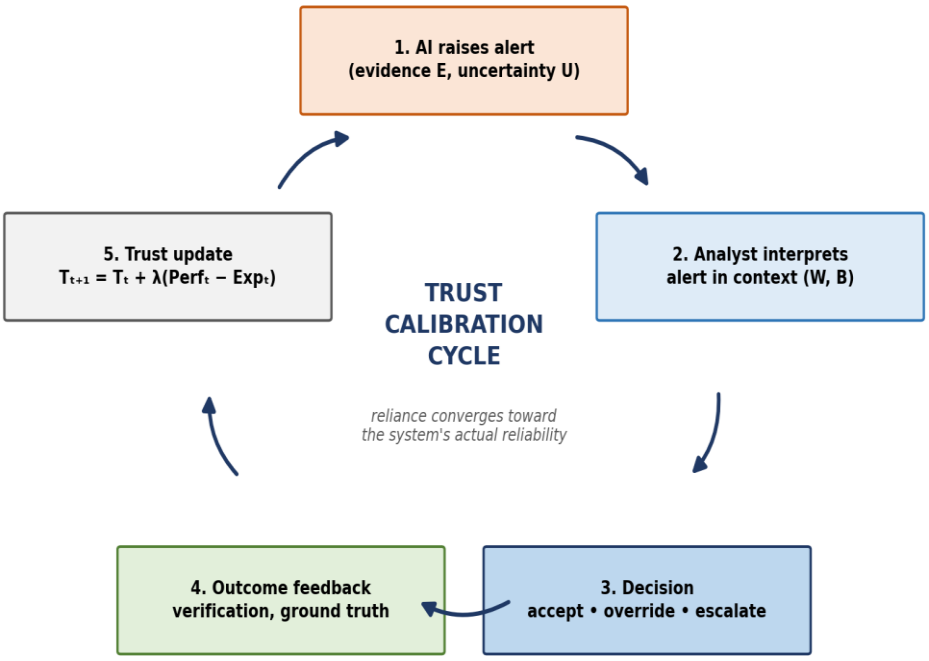

Figure 4. Trust Calibration Cycle Between Human Analysts & AI Systems in SOC Decision-Making

These considerations lead to the following three propositions:

- P1: When workloads are high and time is tight, automation bias increases unless explanations enable verification for the analyst.
- P2: False positives in the absence of an explanation rationale decrease trust and increase the likelihood of overlooking valid alarms.
- P3: Evidence-based explanations, combined with cues and alternative hypotheses about uncertainty, reduce confirmation bias.

To make these propositions falsifiable, we restate each one as a hypothesis with a concrete measurement:

- H1 (from P1): Under high workload, the rate of unverified acceptance of AI recommendations (automation-bias index) is significantly higher than under low workload, and evidence-first explanations with uncertainty cues significantly reduce this rate relative to standard alerts.
- H2 (from P2): A chain of false alarms without a clear explanation greatly reduces trust and increases the rate of false dismissals for future real alarms; presenting the distinguishing evidence solves both problems.
- H3 (from P3): Explanation followed by evidence, uncertainty indicators, and an alternative hypothesis result in high hypothesis revision rates (low confirmation bias), with trust being more in line with the reliability of the system.

Formalization of the Trust-Reliability Model

This construct can be reduced to six variables that include decision reliability D (the quality of the triage decisions made by the analyst compared with reality), cognitive bias B, explanation E, uncertainty U, workload W, and trust calibration T. The most general model of this construct is

$$D = f(B, E, U, W, T), \tag{1}$$

that is, as a first-order approximation, written in the following linear format:

$$D = \alpha(1 - B) + \beta E + \gamma U - \delta W + \varepsilon T, \tag{2}$$

where α, β, γ, δ, and ε are non-negative coefficients to be estimated empirically. In order to be comparable, B, E, U, W, T, and a continuous form of D can be scaled to {0,1}. Bias is assumed to be context-dependent: B = f(W, fatigue, mistakes made before). Similarly, trust depends on interface characteristics and experience, T = f(E, U, past performance), and is updated via a simple learning rule:

$$T_{t+1} = T_t + \lambda(Performance_t - Expectation_t), \tag{3}$$

where λ denotes the learning rate in trust calibration. Equation (2) is intentionally kept as a first-order, hypothesis-generating equation. Signs encode the direction of expected influence, as predicted by the literature: higher bias and workload are expected to decrease reliability, while better explanations, uncertainty communication, and calibrated trust are expected to increase reliability. Since H1 predicts that the influence of transparency depends on workload, interaction terms need to be considered in the empirical specification instead of assuming additivity of influences. One possible extension is

$$D = \alpha_0 + \alpha_1(1 - B) + \beta E + \gamma U - \delta W + \varepsilon T + \theta_1(E \times W) + \theta_2(U \times W) + \zeta, \tag{4}$$

where θ_1 and θ_2 reflect whether explanation quality and uncertainty communication have differing influences as a function of workload, and ζ is the error term. Equation (3) is used to account for the history dependence assumed in P2/H2. Equation (4) must not be estimated as a single composite model where the outcome variable is binary. Instead, generalized linear mixed models need to be estimated for the case of unverified acceptance or false dismissal.

Beyond current formulations of trust calibration, where explanations, uncertainty communication, and workload are considered as distinct antecedents of trust or reliance [19, 20, 22, 26], the inclusion of E×W and U×W in the proposed model makes it dependent on operational workload. The aim is to see if the value of either explaining or communicating uncertainty persists, diminishes, or increases under conditions of increasing alert rate and time pressure. This differentiates the current model from an additive perspective where transparency would always yield similar effects regardless of context, and is consistent with a SOC perspective where interface assistance should still be usable under high workload.

Figure 5 illustrates the formal structure of the proposed trust-reliability framework that relates the input variables and their impact on the reliability of decisions to the trust updating process and the resulting propositions/hypotheses.

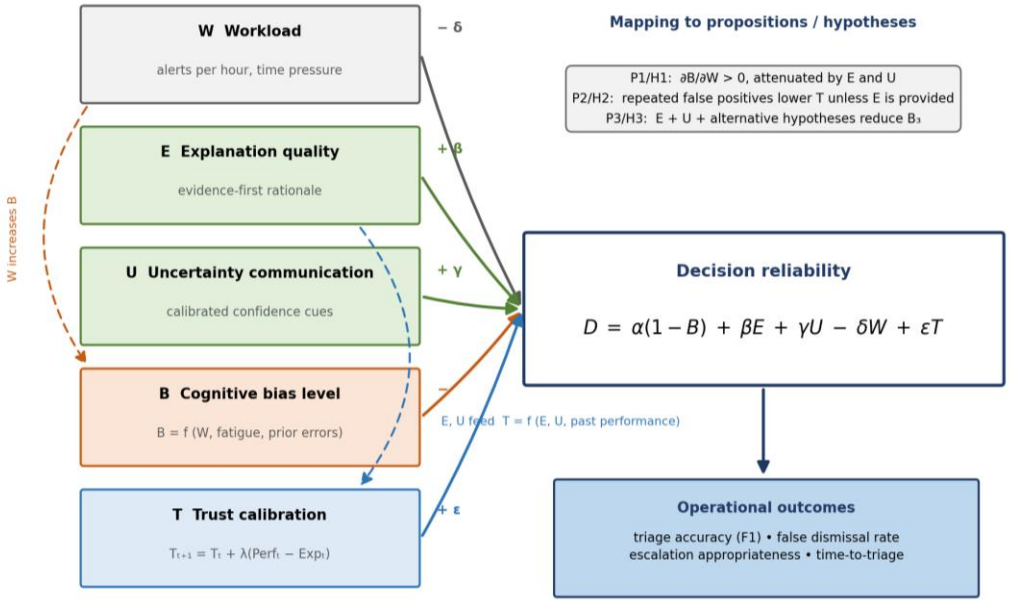

Figure 5. Formal Structure of the Proposed Trust–Reliability Model: Input Variables, Signed Effects on Decision Reliability D, the Trust-Update Rule, and The Mapping to Propositions P1–P3 / Hypotheses H1–H3.

Overall, the model assumes that interaction between humans and machines in decision-making does not represent the process of transferring the decision from one agent to another but rather an ongoing cycle in which transparency, interfaces, and feedback determine whether trust will be an asset or a liability.

OPERATIONALIZATION AND VALIDATION DESIGN

These hypotheses should be falsifiable, and not merely rhetorical in nature. Table 2 provides an indication of the variables, observable indicators, null hypotheses, and analysis that would be used in a SOC-like task in which behavior is separated from trust, as the two do not have to be correlated. In order to test the triaging and escalation decision, it is necessary to use ground truth tags for every alert.

Table 2. Operationalization of H1-H3 and Planned Statistical Tests

Hypothesis / null	Predictors / manipulation	Primary outcomes	Operational measure	Planned analysis
H1: Under high workload, unverified acceptance increases, while	Workload (low/high); evidence-first explanation (absent/present);	Unverified acceptance; triage accuracy; time-to-triage.	Acceptance without opening supporting evidence; correct triage / total;	Mixed-effects logistic models for acceptance/accuracy; mixed-effects linear or survival model for

evidence-first explanations and uncertainty cues reduce it. H01: workload, transparency, and their interaction have no effect on unverified acceptance.	uncertainty cue (absent/present).		decision time in seconds.	time; test W×E and W×U interactions.
H2: Repeated false alarms increase later false dismissal and trust-calibration error; distinguishing evidence reduces both. H02: false-alarm history and distinguishing evidence do not change later false dismissal or trust calibration.	False-alarm exposure; distinguishing evidence on subsequent true alert.	False-dismissal rate; trust-calibration error.	True alerts incorrectly dismissed / true alerts; $\lvert T - RAI \rvert$ where RAI is observed AI reliability.	Mixed-effects logistic model for false dismissal; repeated-measures/mixed-effects model for calibration error; sequence and evidence effects.
H3: Evidence-first explanations, uncertainty cues, and an alternative hypothesis increase hypothesis revision and appropriate escalation. H03: these interventions do not affect revision or escalation.	Evidence-first explanation; uncertainty cue; alternative-hypothesis prompt.	Hypothesis revision; confirmation-bias index; escalation appropriateness.	Revision after disconfirming evidence; proportion of disconfirming evidence inspected; decision scored against scenario ground truth.	Mixed-effects logistic/ordinal models; compare baseline with full debiasing condition and test analyst-level random effects.
H1: Under high workload, unverified acceptance increases, while evidence-first explanations and	Workload (low/high); evidence-first explanation (absent/present); uncertainty cue (absent/present).	Unverified acceptance; triage accuracy; time-to-triage.	Acceptance without opening supporting evidence; correct triage / total; decision time in seconds.	Mixed-effects logistic models for acceptance/accuracy; mixed-effects linear or survival model for time; test W×E and W×U interactions.

uncertainty cues reduce it. H01: workload, transparency, and their interaction have no effect on unverified acceptance.				
H2: Repeated false alarms increase later false dismissal and trust-calibration error; distinguishing evidence reduces both. H02: false-alarm history and distinguishing evidence do not change later false dismissal or trust calibration.	False-alarm exposure; distinguishing evidence on subsequent true alert.	False-dismissal rate; trust-calibration error.	True alerts incorrectly dismissed / true alerts; $\lvert T - RAI \rvert$ where RAI is observed AI reliability.	Mixed-effects logistic model for false dismissal; repeated-measures/mixed-effects model for calibration error; sequence and evidence effects.
H3: Evidence-first explanations, uncertainty cues, and an alternative hypothesis increase hypothesis revision and appropriate escalation. H03: these interventions do not affect revision or escalation.	Evidence-first explanation; uncertainty cue; alternative-hypothesis prompt.	Hypothesis revision; confirmation-bias index; escalation appropriateness.	Revision after disconfirming evidence; proportion of disconfirming evidence inspected; decision scored against scenario ground truth.	Mixed-effects logistic/ordinal models; compare baseline with full debiasing condition and test analyst-level random effects.

Agent-Based Simulation Design

The first validation stage can be done without deriving any results from human participants. Analyst agents are endowed with initial expertise, fatigue, current trust state T, and bias propensity B. Every alert contains ground truth, AI recommendation, AI reliability level RAI, explanation condition E, and uncertainty cue U. Workload W is manipulated by varying the rate of alerts and decision-time constraints. In every triage, the

analyst either accepts, verifies, overrides, discards, or escalates the recommendation; the probabilities of these decisions are controlled by directional structure (Equations (2)-(4)), and trust updates are done based on outcome feedback (Equation (3)). H1-H3 are tested through running the alert sequence under controlled changes in workload, explanation, uncertainty, false alarm history, and alternative hypothesis promptings.

Simulation validation should rely on repeating experiments with different random seeds until confidence intervals of primary outcomes stabilize. Outputs should include triage accuracy, unverified acceptance, false dismissal rate, escalation appropriateness, time-to-triage, and trust-calibration error. Reliability validation should involve reproducing the same experiment across random seeds, while sensitivity validation should cover different settings for AI reliability, workload intensity, λ, and model weights across plausible ranges. The point of validation is not to tweak parameters until H1-H3 are proven true; parameter ranges and decision rules should be predefined prior to running the confirmatory experiments, including null outcomes.

Human and Qualitative Validation Pathway

The next step would be a replication study involving human participants who are exposed to the same factorial conditions in a simulated SOC interface. Prior to recruiting participants, the number of participants required to reach a pre-defined power should be calculated. Besides the behavioral measures listed in Table 2, workload measure and short trust-calibration test should be included in the protocol. If escalation appropriateness needs to be judged by experts, the coding scheme should be created beforehand and used to independently code the collected data, with the level of agreement reported. A qualitative validation may consist of interviewing participants regarding reasons for accepting, challenging, or ignoring AI recommendation. With a pre-specified coding scheme and independent coding, it is possible to differentiate explanation usefulness, workload effects, distrust due to false alarms, and confirmation-seeking behavior. The qualitative findings will describe underlying mechanisms but will not replace the validation of hypotheses.

Reproducibility Roadmap

To make the current paper replicable, the validation package should contain alert scenarios and their ground truth labels, AI reliability schedule, explanation and uncertainty templates, variable definitions, random seeds, parameter files, exclusion rules, analysis scripts, and exact models of H1-H3. If the large language model was used to generate explanations, the model version, system prompt, generation settings, and frozen explanations that were used in the experiment should be saved as part of the replication package. Human study hypotheses and analysis plan should be preregistered, and anonymized data and analysis code should be released after ethics and security approval, if any.

IMPLICATIONS

The key practical implication from this approach is that making predictions more accurate alone is not enough. In order to make reliable decisions, one also needs to look at how predictions are used to inform human decision making, as well as how the use of models is adjusted following feedback.

Design and operational principles:

- Evidence-driven explanations: provide the recommendation along with relevant evidence, such as logs, process trees, and network traces, plus the short explanation [2, 17, 21].
- Intuitive confidence estimates: express the confidence level in clear language, and if possible, compare it to the system's false positive rate.
- Multiple hypotheses: generate at least one additional hypothesis, as well as evidence that will help differentiate them from each other.
- Progressive disclosure: provide a concise first view under high workload and allow deeper inspection when needed.
- Post-decision feedback: show the outcome, whether the decision was correct, and which evidence was most diagnostic.

The following three scenarios show how everything fits into practice.

Scenario 1: a rise in phishing attacks and automation bias. With many alerts generated within an hour of a large-scale phishing attack, there is a situation in which the system generates a command to suspend the account with high confidence level. The overworked security analyst tends to agree with the suggestion and finds out that the suspended accounts were partner mail. The implementation of a two-click validation window regarding sender reputation, URL detonation, and header oddities, confidence presented in the context of what is really happening with the campaign, and a safer alternative such as quarantining prior to suspension will allow the analyst to pause.

Scenario 2: False positive fatigue and missing lateral movement. After a streak of noise from EDR alerts, the analyst gets conditioned to ignore that type of alert as just background noise. Now the EDR solution raises an alert of credential dumping along with anomalous SMB authentication, with medium confidence. The bias developed by the analyst because of previous noise results in ignoring this alert too early. It becomes useful here to highlight what's unique about this situation compared to the previous false positives, such as privileged logons, unusual host relationships, and unusual sequence.

Scenario 3: Confirmation bias when formulating a hypothesis. In the case of suspicious outbound traffic, the analyst might be inclined to conclude quickly that an external attacker is responsible, whereas the system suggests that there is an insider who transfers information out because of the time of access and user behaviour. Once one version of the story becomes fixed in the mind of the analyst, he or she will find it increasingly difficult

to consider any other hypothesis due to confirmation bias. The alternative hypothesis card, the counterfactual question, and the review step make this possible.

The relationship here in all instances is the thesis behind the framework, which states that reliable protection cannot be afforded simply by automating. Reliable protection comes from the fact that the human-AI relationship has been honed in order to counteract the cognitive biases and guarantee constant learning from the analyst.

CHALLENGES AND PROPOSED RESEARCH DIRECTION

Challenge 1: Dynamic Trust Calibration – The current trust models are static and do not account for dynamic variations in the analyst's level of trust as operations in the SOC develop [8, 19, 34].

Proposed Research Direction 1: Creation of adaptive trust calibration systems that can update AI explanations and trust indicators based on the behavior of the analyst, his workload, and system performance [34].

Challenge 2: Personalized Cognitive Bias Management – The current XAI systems deliver explanations in the same way regardless of the different types of expertise and cognitive biases of the analysts [12, 15, 35].

Proposed Research Direction 2: Development of personalized explanation methods that consider the expertise of the analysts.

Challenge 3: Reliability Assessment of Real-Time Decisions: Heterogeneous measures for trust, comprehension, usability, and performance have been proposed within human-centered XAI, which prevents a common assessment of reliability at SOC level [35].

Proposed Research Direction 3: Assess decision reliability through a concise set of measures related to AI performance, behavioral dependence on AI, trust calibration, workload, and explanation utilization; self-confidence should be accounted for since it may influence proper reliance [36].

Challenge 4: Resilience of Human-AI Interaction from Cognitive Point of View: Decision support may go wrong due to confirmation of cognitive biases by AI, so the interaction itself should ensure verification rather than blind adoption [2, 24, 37].

Proposed Research Direction 4: Provide uncertainty-aware explanations, competing pieces of evidence, and explicit steps for verification of the results so that the analyst can identify and dispute human and model-side biases.

LIMITATIONS AND BOUNDARY CONDITIONS

First, there is a methodological limitation of the literature synthesis that is semi-systematic and theoretical. It explicates the selection process and comparison criteria but does not provide PRISMA-like counting of studies as well as does not claim full coverage. Also, one of the SOC-specific comparators [28] is preprint and so will be used for

conceptual positioning purposes only. Further updates will use the protocol described in the Methodology section on a more extensive indexed database.

Second, there is the limitation related to the formal model itself. Equations (2)-(4) are intentionally concise and should be considered hypotheses-generating structures, but not validated prediction equations. B and T are latent constructs; some of the predictors could be correlated; and the true relationship may be either nonlinear or threshold-type. The addition of interaction terms to Equation (4) solves the most immediate problem of workload and transparency but any additional interactions should be introduced only where the theory allows it and data supports it, and not because the model became too complex.

Third, the study has no new empirical observations. Operation and implementation of the framework in terms of its operationalization, statistical testing, agent-based simulation, expert review and qualitative interviews pathway show how the framework could be validated; they do not represent the results of such an investigation. Consequently, the claims in this paper are limited to conceptual integration, testability, and positioning within the domain; reliability, sensitivity, effect sizes, and predictive performance can only be claimed once the validation procedure will be implemented.

Finally, the framework is focused on AI-assisted SOC triage and incident response decisions only. Any transfer to another cybersecurity decision-making task, to another organization, to different experience of the analysts, to another architecture of the AI system, is out of scope. Live SOC includes the elements like team communication, organizational procedures, legal constraints, and asymmetric costs of the false positives and false negatives that are excluded from the single-analyst model.

CONCLUSION

The collaboration between artificial intelligence (AI) systems and human analysts in anomaly assessment, alert prioritization, evidence evaluation, and response decision-making is becoming an increasingly important component of Security Operations Center (SOC) operations. The reliability of such collaboration cannot be assessed solely on the basis of AI detection performance. How analysts interpret, verify, accept, challenge, or reject AI-generated recommendations under conditions of workload, uncertainty, time pressure, and cognitive bias can substantially influence the resulting security decisions.

This study proposes an integrated socio-technical framework that brings together cognitive biases, explanation quality, uncertainty communication, workload, and trust calibration within the context of AI-assisted SOC decision-making. The framework provides a structured basis for examining why analysts may over-rely on AI recommendations, reject valid system outputs, or fail to revise initial hypotheses when confronted with contradictory evidence. The proposed hypotheses and operational indicators further establish a basis for empirically examining these relationships.

The primary contribution of the study is conceptual and methodological. The proposed framework integrates human, AI, and contextual factors and links them to the measurable construct of decision reliability. It also provides a basis for examining how workload may influence the effectiveness of explanations and uncertainty communication, rather than treating these factors as independent determinants of human–AI decision-making.

The empirical dimension remains an open area for future investigation. Although the proposed operational definitions, statistical procedures, and agent-based validation methodology provide a structured basis for testing the hypothesized relationships, they do not constitute empirical evidence. The effects of cognitive biases, explanation quality, uncertainty communication, workload, and trust on decision reliability therefore require direct empirical assessment.

Future research should initially evaluate the proposed relationships through controlled simulation and sensitivity analysis, followed by empirical studies conducted in SOC-like environments involving human analysts and, where feasible, experienced SOC professionals. Such investigations should jointly consider multiple outcome measures, including triage accuracy, false-dismissal rates, escalation appropriateness, time to triage, hypothesis revision, workload, and trust calibration. This multi-dimensional evaluation would provide a stronger basis for assessing the reliability and practical applicability of the framework.

The contribution of this paper is therefore not to position AI as a substitute for human expertise in cybersecurity, but to provide a framework for examining how human judgment and AI assistance can be integrated within SOC decision-making. The framework emphasizes three central principles: trust should be dynamically calibrated, explanations should be interpreted in relation to the decision context, and workload should be treated as an important component of human–AI interaction. Together, these principles provide a foundation for future empirical research on reliable and cognitively informed AI-assisted cybersecurity decision-making.

NOMENCLATURE

SOC	Security Operations Center
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
ML	Machine Learning
HITL	Human-in-the-Loop
EDR	Endpoint Detection and Response
SMB	Server Message Block (network file-sharing protocol)
URL	Uniform Resource Locator
Automation bias	Tendency to over-rely on automated recommendations, especially under high workload

Algorithm aversion	Tendency to reject or under-rely on automated recommendations, often after observed errors
Confirmation bias	Tendency to seek evidence that supports an existing hypothesis and to discount evidence against it
Trust calibration	The ongoing adjustment of reliance so that it matches the system's actual reliability
Trust surface	The human–AI interface, where reliance decisions are made
P1–P3	The three propositions stated in the Theoretical Framework
H1–H3	Testable hypotheses derived from P1–P3
D	Decision reliability (quality of triage decisions against ground truth)
B	Cognitive bias level; B = f (W, fatigue, prior errors)
E	Explanation quality (evidence-first rationale)
U	Uncertainty communication (calibrated confidence cues)
W	Workload (e.g., alerts per hour, time pressure)
T	Trust calibration level; updated as $T_{t+1} = T_t + \lambda(\text{Performance}_t - \text{Expectation}_t)$
$\alpha, \beta, \gamma, \delta, \varepsilon$	Non-negative model weights in $D = \alpha(1 - B) + \beta E + \gamma U - \delta W + \varepsilon T$
λ	Trust calibration (learning) rate
RAI	Observed reliability of the AI system used to compute trust-calibration error $\lvert T - RAI \rvert$.
θ_1, θ_2	Interaction weights for E × W and U × W in Equation (4).
ζ	Residual or unexplained variation in the empirical extension of the reliability model.

AUTHOR CONTRIBUTIONS

Conceptualization, D.H.; Methodology, D.H., and E.L.; Validation, E.L., and L.L; Investigation, D.H., and E.L.; Resources, E.L., and M.M.S.; Data Curation, D.H., and E.L.; Writing – Original D.H.; writing—review and editing, D.H., and A.H.S.; Visualization, L.L., and M.M.S.; Supervision, D.H., and A.H.S.; Project Administration, D.H., E.L., M.M.S. and A.H.S.

ACKNOWLEDGMENT

The authors acknowledge the institutional support of the Mediterranean University of Albania and the scientific networking context provided by COST Action CA22104 "Behavioural Next Generation in Wireless Networks for Cyber Security" (BEiNG-WISE).

CONFLICT OF INTERESTS

The authors confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W.S., & Horvitz, E. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, **2019**, *33*(01), 2429–2437.
2. Charmet, F., Tanuwidjaja, H.C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. Explainable artificial intelligence for cybersecurity: A literature survey. *Annals of Telecommunications*, **2022**, 77, 789–812.
3. Chen, Y., Zahedi, F.M., Abbasi, A., & Dobolyi, D.G. Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools. *Information & Management*, **2021**, *58*(1), 103394.
4. Leka, E., Lamani, L., Aliti, A., & Hoxha, E. Web Application Firewall for Detecting and Mitigation of Based DDoS Attacks Using Machine Learning and Blockchain. *Technology, Education, Management, Informatics*, **2024,** *13*(4), 2802–2811.
5. Işın, L. İ., Dalveren, Y., Leka, E., & Kara, A. Securing the Internet of Things: Challenges and Complementary Overview of Machine Learning-Based Intrusion Detection. *In 2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Ankara, Turkey, **2024**, pp. 1–4.
6. Leka, E., Lamani, L., & Hoxha, E. Securing the Foundations of 6G: Innovative Intelligent Controls at the Physical Layer for Trustworthiness and Resilience. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, **2024**, pp. 1526–1531.
7. Moruzzi, C., & El-Assady, M. Which biases and reasoning pitfalls do explanations trigger? Decomposing communication processes in human–AI interaction. *IEEE Computer Graphics and Applications*, **2022**, *42*(6), 11–23.
8. Zerilli, J., Bhatt, U., & Weller, A. How transparency modulates trust in artificial intelligence. *Patterns*, **2022**, *3*(4), 100455.
9. Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, **2023**, *139*, 107539.
10. Hyka, D., Vuka, E., & Milošević, S.M. The Future of Web Security: Advanced Testing and Cryptographic Defense. *Wseas Transactions on Information Science and Applications*, **2026**, 23, 253–261.
11. Hyka, D., Meçaj, J., Kodra, F., Milošević, M., & Ćirić, V. Cybersecurity awareness and education for young internet users: A comprehensive analysis of prevention strategies. In *2025 IEEE Conference on Communications and Network Security (CNS)*, **2025**, pp. 1–5.
12. Dikmen, M., & Burns, C. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, **2022**, *162*, 102792.
13. Adadi, A., & Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, **2018,** *6*, 52138–52160.

14. Zhou, J., Gandomi, A.H., Chen, F., & Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, **2021**, *10*(5), 593.
15. Atf, Z., & Lewis, P.R. Human centricity in the relationship between explainability and trust in AI. *IEEE Technology and Society Magazine*, **2023**, *42*(4), 66–76.
16. De Bruijn, H., Warnier, M., & Janssen, M. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, **2022**, *39*(2), 101666.
17. Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. Explainable recommendation: When design meets trust calibration. *World Wide Web*, **2021**, *24*, 1857–1884.
18. Papagni, G., de Pagter, J., Zafari, S., Filzmoser, M., Tscheligi, M., & Koeszegi, S. T. Artificial agents' explainability to support trust: Considerations on timing and context. *AI & Society*, **2023**, *38*, 947–960.
19. Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, **2020**, *1*(4), 100049.
20. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. 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*, **2020**, pp. 295–305.
21. Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, **2023**, *169*, 102941.
22. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1)*, **2023**, 1–38.
23. Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, **2020**, 272–283.
24. Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, **2022**, pp. 78–91.
25. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, **2023**, 56(4), 3005–3054.
26. Okamura, K., & Yamada, S. Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, **2020**, *15*(2), e0229132.
27. Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, **2022**, pp. 617–626.
28. Mohsin, A., Janicke, H., Ibrahim, A., Sarker, I. H., & Camtepe, S. A Unified Framework for Human–AI Collaboration in Security Operations Centers with Trusted Autonomy. *ACM Trans. Internet Technol.* **2026**, [accepted for publication].
29. Tatasciore, M., & Loft, S. Calibrating Reliance on Automated Advice: Transparency and Trust Calibration Feedback. *International Journal of Human-Computer Interaction*, **2025**, *41*(23), 14723–14733.

30. Göndöcs, D., Horváth, S., & Dörfler, V. Uncovering the dynamics of human-AI hybrid performance: A qualitative meta-analysis of empirical studies. *International Journal of Human-Computer Studies*, **2025**, *205*, 103622.
31. Cheng, L., Varshney, K.R., & Liu, H. Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, **2021**, *71*, 1137–1181.
32. Leka, E., Hoxha, E., & Rexha, G. Security and Privacy Concerns Associated with the Internet of Things (IoT) and the Role of Adapting Blockchain and Machine Learning – A Systematic Literature Review. *In 2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, **2023**, pp. 1690–1696.
33. Lamani, L., Leka, E., Rexha, G., & Aliti, A. A secure and resilient framework for wildfire management via UAVs and blockchain federated learning. In *2026 49th MIPRO ICT and Electronics Convention (MIPRO),* **2026**, pp. 2078–2083.
34. Lucas, G.M., Becerik-Gerber, B., & Roll, S.C. Calibrating Workers' Trust in Intelligent Automated Systems. *Patterns*, **2024**, *5*(9), 101045.
35. Rong, Y., Leemann, T., Nguyen, T.-T., *et al.* Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2024**, *46*(4), 2104–2122.
36. Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. Are You Really Sure? Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. *In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, **2024**, pp. 1–20.
37. Echterhoff, J.M., Liu, Y., Alessa, A., *et al.* Cognitive Bias in Decision-Making with LLMs. *In Findings of the Association for Computational Linguistics: EMNLP 2024; Association for Computational Linguistics:* **2024**, pp. 12640–12653.