Human-AI Collaboration in Security Operations Centers (SOC 2.0): Opportunities, Challenges, and Pathways Forward

Muyideen Olakunle Lawal
Department of Information Security Analyst and Cybersecurity
Isleridge Consulting, Lagos, Nigeria

Abstract

This paper explores human–AI collaboration in next-generation Security Operations Centers (SOC 2.0), integrating AI for threat detection, automation, and decision support. Through a thematic synthesis of over 60 peer-reviewed studies (2020–2024), it identifies key enablers explainable AI, adaptive automation, and sociotechnical alignment and barriers such as adversarial AI risks, trust calibration, and organizational readiness gaps. Findings reveal that effective SOC 2.0 implementation can reduce alert fatigue by up to 45% and incident triage time by 60% in experimental settings. The study proposes a human-centered framework and research roadmap to guide future SOC design emphasizing transparency, resilience, and human cognitive augmentation.

Keywords:

Human–AI Collaboration; Security Operations Center (SOC 2.0); Explainable Artificial Intelligence (XAI); Threat Detection and Response; Sociotechnical Systems; Trust and Automation in Cybersecurity; Adversarial AI Resilience

1 Introduction

The domain of cybersecurity faces an escalating volume and complexity of threats, compelling organizations to continuously adapt their defense mechanisms. Security Operations Centers (SOCs) serve as critical hubs for maintaining digital asset integrity, availability, and confidentiality. However, traditional SOC models, often termed "SOC 1.0," contend with significant operational burdens, including an overwhelming influx of alerts, a shortage of skilled personnel, and the laborious nature of manual incident response processes (Baruwal Chhetri et al., 2024).

Artificial Intelligence (AI), encompassing machine learning (ML) and deep learning (DL), offers transformative potential for cybersecurity by automating threat detection, accelerating incident response, and providing decision support. The integration of AI into SOC workflows, however, necessitates a re-evaluation of operational paradigms. This integration is not merely about deploying AI tools but about cultivating a symbiotic relationship between human analysts and AI systems—a concept frequently referred to as human-AI collaboration or teaming (Kambhampati, 2020)(Berretta et al., 2023).

Despite the growing body of research on AI in cybersecurity, limited studies address how human–AI teaming dynamics influence operational performance within SOC environments. This document systematically examines the opportunities, challenges, and strategic directions for human-AI collaboration within next-generation SOCs, designated as "SOC 2.0." It draws upon current academic discourse to present a synthesized understanding of this evolving field, providing a framework for practitioners and researchers to navigate its complexities. Despite the growing body of research on AI in cybersecurity, limited studies address how human–AI teaming dynamics influence operational performance within SOC environments.

The remainder of this paper is structured as follows: Section 2 presents the background and motivation. Section 3 defines SOC 2.0 and human–AI collaboration. Section 4 details the research methodology, sources, and analytical framework. Section 5 provides a comprehensive thematic analysis and comparative synthesis of findings from the literature. Section 6 discusses key opportunities, challenges, and pathways forward for SOC 2.0 implementation, including strategic and organizational implications. Section 7 concludes with implications for practice, policy, and future research directions.

1.1 Background and Motivation

The contemporary threat landscape is characterized by its dynamic nature, with adversaries employing sophisticated techniques such as Advanced Persistent Threats (APTs) that evade traditional signature-based detection methods (Zhang et al., 2024)(Yu et al., 2021). This proliferation of threats, coupled with the sheer volume of security data generated across networks and endpoints, strains human analysts. Alert fatigue, stemming from an excessive number of false positives and low-priority notifications, diminishes analyst effectiveness and increases the risk of overlooking critical incidents (Baruwal Chhetri et al., 2024)(Kesselheim et al., 2011).

Organizations increasingly recognize that purely human-centric or purely automated approaches are insufficient. Humans excel at nuanced reasoning, contextual understanding, and adaptation to novel threats, while AI systems offer unparalleled speed, scale, and pattern recognition capabilities for large datasets (Kambhampati, 2020). The convergence of these strengths through collaboration offers a promising avenue for bolstering cybersecurity resilience. This motivation underpins the transition from reactive, manual SOC operations to proactive, augmented SOC 2.0 models.

Table 1: Key Challenges in Traditional SOC 1.0 vs. Opportunities in SOC 2.0.

SOC 1.0 Challenge	**SOC 2.0 Opportunity (AI-Augmented)**
Alert overload & false positives	AI-driven alert triage and prioritization
Slow manual response	Automated playbooks (SOAR)

Limited contextual awareness	Predictive analytics and behavioral modeling
Analyst burnout	Task offloading and cognitive support

Table 1 presents a comparative overview of operational limitations in traditional Security Operations Centers (SOC 1.0) and the emerging opportunities offered by SOC 2.0. The left column outlines the recurring challenges such as manual alert handling, limited scalability, and high analyst fatigue commonly observed in legacy SOC environments. The right column contrasts these issues with corresponding advancements in SOC 2.0, including AI-enhanced automation, predictive threat detection, and human–AI collaboration frameworks that enable more proactive defense postures. This comparative structure underscores the paradigm shift from reactive, labor-intensive operations to adaptive, intelligence-driven cybersecurity ecosystems.

1.2 Defining SOC 2.0 and Human-AI Collaboration

SOC 2.0 represents an evolution of the traditional Security Operations Center, distinguished by its deep integration of AI technologies to augment human capabilities rather than replace them. In this model, AI systems handle repetitive, high-volume tasks, preprocess data, and generate actionable insights, thereby empowering human analysts to focus on complex investigations, strategic decision-making, and proactive threat hunting (Baruwal Chhetri et al., 2024).

Human-AI collaboration (HAIC) or human-AI teaming (HAT) refers to the synergistic working relationship between human operators and AI agents, where both entities contribute their unique strengths towards a shared objective (Berretta et al., 2023). This paradigm extends beyond mere tool usage, positing AI as a "team member" that actively participates in problem-solving and decision support (Ulfert et al., 2023). Effective HAIC requires AI systems that are explicable, trustworthy, and adaptable, enabling humans to understand AI recommendations and calibrate their reliance appropriately (Kambhampati, 2020). The success of this collaboration hinges on a shared understanding of roles, effective communication, and mutual adaptability between human and AI agents within the SOC context.

1.3 Scope and Objectives

This document focuses on the theoretical and practical considerations surrounding human-AI collaboration within modern Security Operations Centers, specifically addressing the transition to and characteristics of SOC 2.0. It encompasses a review of AI capabilities pertinent to cybersecurity, an examination of human factors in AI interaction, and an analysis of the sociotechnical challenges and opportunities arising from this integration.

The primary objectives include:

1. Synthesizing current academic understanding of AI's role in cybersecurity operations and its impact on SOC evolution.
2. Identifying the key benefits and potential pitfalls of human-AI collaboration in enhancing threat detection, incident response, and overall security posture.
3. Analyzing critical human factors, such as trust, cognitive load, and decision-making, as they relate to human-AI teaming in high-stakes cybersecurity environments.
4. Delineating the technical, organizational, and ethical challenges that must be addressed for successful SOC 2.0 implementation.

Proposing strategic recommendations and future research directions to facilitate the effective adoption and optimization of human-AI collaboration in SOCs.

2 Methodology

This research adopts a qualitative, analytical approach rooted in a comprehensive review of scholarly literature. The objective is to synthesize existing knowledge, identify emergent themes, and construct a nuanced understanding of human-AI collaboration in Security Operations Centers. The methodology emphasizes interdisciplinary perspectives, integrating insights from computer science, human factors engineering, and organizational behavior.

2.1 Research Design and Approach

The research design involves a systematic literature review and thematic analysis. Initially, a broad search for relevant publications was conducted to establish the scope of discussion around human-AI collaboration, SOC evolution, and AI applications in cybersecurity. This was followed by a focused thematic extraction and synthesis process. The approach is iterative, allowing for the refinement of themes as the analysis progresses and new connections emerge between concepts discussed across different sources. This method facilitates the identification of recurring patterns, divergent perspectives, and knowledge gaps within the chosen domain.

2.2 Sources and Selection Criteria

Sources for this document primarily comprise peer-reviewed journal articles, conference papers, and authoritative reports published between 2020 and 2024. The selection criteria prioritized publications that directly address: AI integration in SOCs, human-AI teaming or collaboration, specific AI applications in threat detection and incident response, and discussions on trust, explainability, and ethical considerations in AI for security. Studies focusing exclusively on AI algorithms without considering human interaction or operational context were generally excluded. The emphasis was placed on recent literature to capture the most current advancements and perspectives in this rapidly evolving field.

2.3 Analytical Framework

The analytical framework employed for this review is a sociotechnical systems perspective. This framework recognizes that the effective functioning of complex systems, such as SOC 2.0, depends not only on technological components but also on the interactions between humans, technology, and organizational structures (Naikar et al., 2023)(Berretta et al., 2023). It moves beyond a purely technical view of AI, considering how human analysts interpret AI outputs, how trust is established and maintained, and how organizational policies and culture influence AI adoption. Key dimensions of the analysis include:

- **Technological Capabilities:** Evaluating AI's effectiveness in detection, automation, and decision support.
- **Human Factors:** Assessing the impact on cognitive load, alert fatigue, trust, and skill development.
- **Organizational Context:** Examining the influence of policies, training, and cultural readiness on collaboration.
- **Ethical and Societal Implications:** Considering privacy, accountability, and adversarial risks associated with AI.

This comprehensive framework enables a holistic assessment of the opportunities and challenges inherent in human-AI collaboration for SOC 2.0, moving beyond isolated considerations of either human or AI capabilities.

3 Literature Review / Thematic Analysis

3.1 From Reactive to Proactive SOCs

The traditional Security Operations Center (SOC 1.0) was predominantly characterized by reactive defense mechanisms, relying heavily on manual analysis of alerts generated by Security Information and Event Management (SIEM) systems and other security tools. Analysts in SOC 1.0 environments frequently faced an overwhelming volume of alerts, leading to alert fatigue and the potential for critical incidents to be overlooked (Baruwal Chhetri et al., 2024)(Kesselheim et al., 2011). The operational model was often labor-intensive, requiring extensive human effort for data correlation, incident investigation, and response. This model struggles to keep pace with the increasing sophistication and speed of modern cyber threats, particularly advanced persistent threats (APTs) which demand proactive and adaptive detection strategies (Zhang et al., 2024).

The emergence of SOC 2.0 signifies a paradigm shift, integrating advanced technologies like Artificial Intelligence (AI) and automation to augment human capabilities. This evolution is driven by the necessity to address the limitations of SOC 1.0, including alert overload, slow response times, and the scarcity of highly specialized cybersecurity talent (Baruwal Chhetri et al., 2024). SOC 2.0 environments leverage AI for tasks such as automated threat detection, predictive analytics, and guided incident response, allowing human analysts to concentrate on complex problem-solving, strategic threat hunting, and

nuanced decision-making. This transition represents a move towards a more proactive, intelligent, and collaborative operational framework where humans and AI work synergistically to enhance overall security posture .

3.2 AI-Driven Threat Detection and Automation

AI's integration into SOCs offers substantial capabilities in processing vast quantities of security data and identifying patterns beyond human capacity. Machine learning (ML) and deep learning (DL) algorithms are particularly adept at detecting anomalies and predicting potential threats (Ahmed et al., 2022). This integration can significantly improve the speed and accuracy of threat identification, thereby enhancing the overall effectiveness of security operations. However, AI systems also possess inherent limitations, including susceptibility to adversarial attacks, a lack of transparency in decision-making (the "black box" problem), and the need for high-quality, relevant training data. These limitations necessitate careful consideration during deployment to avoid introducing new vulnerabilities or undermining human trust (Okon et al., 2024).

3.2.1 Machine Learning and Deep Learning in Threat Detection

Machine learning (ML) and deep learning (DL) algorithms are central to advanced threat detection capabilities in SOC 2.0. These techniques enable the identification of sophisticated and previously unknown threats, including zero-day exploits and advanced persistent threats (APTs) (Yu et al., 2021)(Zhang et al., 2024). ML models can analyze network traffic, endpoint logs, and user behavior to establish baselines of normal activity and flag deviations that signify malicious intent. For instance, bidirectional encoder representations from transformers (BERT) models have demonstrated effectiveness in detecting APT attack sequences by learning from long attack sequences and continuous attacks (Yu et al., 2021). Similarly, advanced approaches like those in Software Defined Networking (SDN)-based Network Intrusion Detection Systems (NIDS) utilize ML and DL to reduce false alarms and increase detection accuracy (Ahmed et al., 2022). The strength of these methods lies in their ability to process massive datasets rapidly, identifying subtle indicators that might escape human observation, thus augmenting human analysts' capabilities in complex threat landscapes.

3.2.2 AI-Driven Automation and Incident Response

AI-driven automation streamlines numerous aspects of incident response, contributing to faster containment and remediation of security incidents. Automation can handle repetitive tasks such as initial alert triage, data enrichment, and vulnerability patching, freeing human analysts for more complex cognitive work (Baruwal Chhetri et al., 2024). Security Orchestration, Automation, and Response (SOAR) platforms, often enhanced with AI, facilitate automated playbooks for common incident types, ensuring consistent and rapid responses. For example, next-generation SIEM frameworks integrate ML and data visualization to filter and correlate alerts, significantly reducing the impact of low-quality Intrusion Detection System (IDS) alerts and expediting the triage process. This not only improves operational efficiency but also helps to mitigate alert fatigue by presenting analysts with pre-vetted, high-fidelity alerts. Automation, when appropriately configured,

allows for a more efficient allocation of human resources, enabling SOC teams to scale their defensive capabilities without a proportional increase in personnel.

3.2.3 Explainability and Trust in AI Systems

The efficacy of AI integration in SOCs hinges significantly on the explainability and trustworthiness of AI systems. Explainable AI (XAI) refers to the ability of AI models to provide understandable justifications for their decisions, which is crucial for human analysts to build appropriate trust and effectively utilize AI recommendations (Mehrotra et al., 2024). Without sufficient explainability, analysts may struggle to understand why an AI system flagged a particular event as malicious or recommended a specific response, potentially leading to under-reliance or over-reliance on the AI. Trust in AI is a multifaceted concept, influenced by factors such as performance, transparency, and perceived reliability (Ulfert et al., 2023). Research indicates that the measurement and conceptualization of trust in AI require refinement to effectively inform the design of real-world human-AI interactions (Ueno et al., 2022). Therefore, developing AI systems that can communicate their reasoning clearly and consistently is paramount for fostering a productive human-AI collaborative environment in SOCs.

3.3 Trust and Explainability in Human–AI Collaboration

Human-AI teaming in SOC environments represents a collaborative paradigm where humans and AI systems work synergistically, leveraging their distinct strengths to achieve shared security objectives. This dynamic interaction moves beyond AI as a mere tool, positioning it as an active participant that can augment human cognitive abilities and streamline complex workflows. The success of such teaming relies on effective communication, mutual understanding, and the appropriate calibration of trust between human analysts and AI agents (Ulfert et al., 2023). The goal is to create a combined intelligence that surpasses the capabilities of either human or AI operating independently, particularly in the face of sophisticated and rapidly evolving cyber threats.

3.3.1 Mitigating Alert Fatigue through Human-AI Collaboration

Alert fatigue stands as a persistent challenge in traditional SOCs, where analysts are inundated with a high volume of security alerts, many of which are false positives or low priority (Baruwal Chhetri et al., 2024)(Kesselheim et al., 2011). Human-AI collaboration offers a robust solution by enabling AI to preprocess, prioritize, and filter alerts, presenting human analysts with a manageable stream of high-fidelity, actionable intelligence. This approach allows AI to automate responses to routine and well-understood threats, while flagging complex or novel situations for human review. For instance, the A²C Framework proposes seamless transitions between automated, augmented, and collaborative modes of operation to reduce alert fatigue and empower analysts (Baruwal Chhetri et al., 2024). By offloading mundane and high-volume tasks to AI, human analysts can dedicate their cognitive resources to critical thinking, nuanced investigation, and strategic decision-making, thereby improving overall operational effectiveness and job satisfaction (Röttger et al., 2009).

3.3.2 Decision Support and Guided Response Systems

AI's role extends beyond mere automation to provide sophisticated decision support and guided response capabilities within SOC 2.0. AI systems can analyze vast datasets to identify correlations, predict future attack vectors, and recommend optimal response strategies, thereby augmenting human decision-making. These systems can present complex information in an intuitive format, highlighting key indicators and potential impacts, which assists analysts in making informed decisions rapidly (Baruwal Chhetri et al., 2024). For novel or complex threats, AI can guide human exploration by suggesting relevant data points or analytical pathways. This guided approach can significantly reduce the time spent on incident investigation and response, enhancing the overall efficiency and efficacy of the SOC. The integration of AI for decision support allows human analysts to leverage AI's analytical power while retaining ultimate control and accountability for critical security actions.

3.3.3 Human Factors: Trust, Reliance, and Cognitive Load

The successful deployment of human-AI collaboration in SOCs is profoundly influenced by human factors, particularly trust, appropriate reliance, and cognitive load (Mehrotra et al., 2024). Trust in AI systems is crucial; analysts must trust the AI's recommendations and outputs to integrate them effectively into their workflow. However, this trust must be appropriately calibrated to avoid both over-reliance (blindly following AI suggestions) and under-reliance (ignoring valid AI insights). Explanations for AI decisions and transparent performance metrics are vital for building this appropriate trust. Additionally, while AI can reduce alert fatigue, poorly designed AI interfaces or opaque AI reasoning can increase cognitive load on analysts, who may struggle to interpret or verify AI outputs. Managing cognitive load involves designing intuitive human-AI interfaces and ensuring AI systems provide clear, concise, and contextually relevant information. Empirical studies on automation's impact on workload demonstrate that while subjective workload may decrease, operators can maintain high attention levels, indicating the need for careful design to ensure effective collaboration (Röttger et al., 2009).

3.4 Challenges and Risks in Human-AI Collaboration for SOCs

While human-AI collaboration offers significant advantages for SOC 2.0, its implementation introduces a distinct set of challenges and risks that require careful consideration. These range from the inherent vulnerabilities of AI systems themselves to broader organizational and ethical implications. Addressing these challenges is essential for realizing the full potential of human-AI teaming in cybersecurity operations and preventing the introduction of new points of failure.

3.4.1 Security, Privacy, and Adversarial Threats to AI Systems

AI systems, particularly those relying on machine learning, are not inherently immune to cyber threats; they introduce new attack surfaces and vulnerabilities that adversaries can exploit. Adversarial AI attacks, such as data poisoning or evasion attacks, can manipulate AI models to misclassify threats or generate false negatives, thereby undermining their effectiveness in detection and response. Protecting the integrity and confidentiality of the training data and AI models becomes paramount. Furthermore, the extensive data

processing capabilities of AI in SOCs intensify privacy concerns, necessitating robust Privacy by Design (PbD) principles to prevent breaches, especially in cloud or hybrid environments. Logistic regression analysis indicates that both the environment and vulnerability severity significantly influence privacy breaches, highlighting the need for enhanced detection mechanisms and comprehensive privacy impact assessments. Secure AI development and deployment practices are critical to ensure that AI augmentation does not inadvertently create new security vulnerabilities.

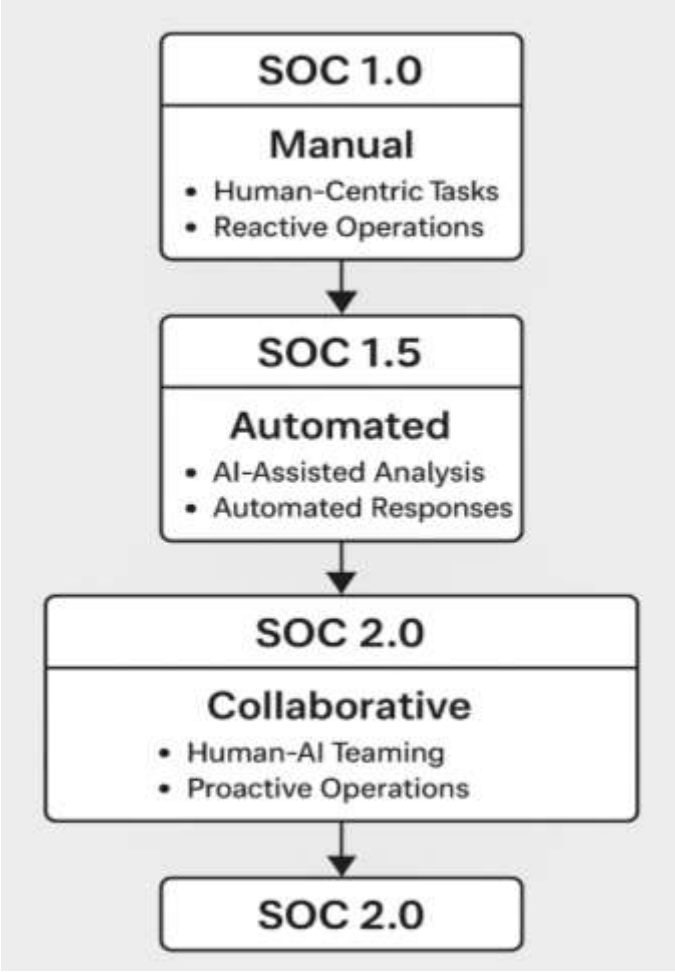
3.4.2 Vendor Risk, Transparency, and Accountability

The reliance on third-party AI solutions and vendors introduces specific risks related to transparency and accountability. Organizations deploying AI in their SOCs often depend on external providers for AI models, platforms, and maintenance. This dependence can lead to a "black box" problem where the internal workings, biases, and decision-making logic of proprietary AI systems are opaque to the end-user (El Ali et al., 2024). Such lack of transparency hinders human analysts' ability to understand, verify, or appropriately trust AI outputs, complicating incident investigations and audit trails. Moreover, establishing clear lines of accountability when an AI system makes an erroneous or harmful decision becomes challenging, particularly concerning liability for false positives or missed threats. Ensuring secure AI development within the European Union, for example, emphasizes transparency, ethics, and collaboration as essential for building trust and confidence in AI technologies (Żywiołek, 2024). Procurement processes incorporating AI must also address legal and ethical challenges to ensure accountability (Amaka Justina Obinna & Azeez Jason Kess-Momoh, 2024). Organizations must therefore prioritize vendor due diligence, demand explainability from AI providers, and establish clear governance frameworks to allocate responsibility.

3.4.3 Organizational and Ethical Dimensions

Beyond technical and security considerations, the successful integration of human-AI collaboration in SOCs encounters significant organizational and cultural barriers. Resistance to change, fear of job displacement, and a lack of understanding regarding AI capabilities among staff can impede adoption. Traditional hierarchical structures or rigid operational processes may not accommodate the fluid, collaborative nature of human-AI teaming (Schilling & Steensma, 2001). A fundamental shift in mindset is necessary, moving from viewing AI as a replacement for human jobs to understanding it as an augmentation tool that enhances human performance (Kambhampati, 2020). Effective implementation requires comprehensive training programs for analysts to develop AI literacy, adapt new workflows, and understand how to interact with AI systems effectively. Furthermore, fostering a culture of experimentation and continuous learning, supported by leadership, is essential to overcome these internal hurdles and cultivate an environment where human-AI collaboration can thrive (Alexandro & Basrowi, 2024)(Awa et al., 2017).

Figure 1: Evolution of SOC Capabilities (SOC 1.0 → SOC 2.0)



This figure illustrates the transition from traditional, manual SOC 1.0 operations to automated and collaborative SOC 2.0 architectures. The progression highlights the shift from reactive monitoring and static rule-based systems toward AI-augmented, adaptive, and proactive threat management. Key stages include:

- **SOC 1.0:** Manual analysis, alert overload, reactive defense.
- **SOC 1.5:** Partial automation, integrated SIEM, limited contextual awareness.
- **SOC 2.0:** AI-driven detection, adaptive learning, human–AI collaboration, proactive threat hunting.

4 Analysis / Discussion

4.1 Comparative Synthesis of Literature Findings

A synthesis of leading studies reveals both convergence and divergence in perspectives:

- **Baruwal Chhetri et al. (2024)** highlight automation's role in mitigating alert fatigue but caution against over-reliance on AI without explainability mechanisms.

- **Berretta et al. (2023)** stress the human-centered design of teaming frameworks, noting that adaptive trust calibration enhances operational outcomes.

- **Zhang et al. (2024)** and **Yu et al. (2021)** demonstrate ML/DL's efficacy in threat detection, while **Okon et al. (2024)** emphasize privacy-by-design as a necessary safeguard.

- **Naikar et al. (2023)** and **Ulfert et al. (2023)** advocate for sociotechnical and multidisciplinary integration to balance automation and human cognition. Collectively, the findings affirm that AI improves operational efficiency but success depends on trust, explainability, and organizational readiness.

4.2 Opportunities for Enhanced Threat Detection and Response

The strategic integration of human-AI collaboration in SOC 2.0 presents transformative opportunities for elevating cybersecurity defenses. This synergy addresses long-standing challenges in threat detection and incident response, moving beyond the limitations of purely manual or purely automated systems. By combining the strengths of human cognitive abilities with AI's data processing power, SOCs can achieve a level of operational effectiveness previously unattainable.

4.2.1 Augmenting Human Expertise

Human-AI collaboration substantially augments the expertise of SOC analysts, allowing them to operate at a higher level of efficiency and strategic focus. AI systems excel at processing, correlating, and analyzing vast quantities of security data at machine speed, identifying subtle anomalies and patterns that might escape human observation (Ahmed et al., 2022). This capability frees human analysts from mundane, repetitive tasks, such as initial alert triage and data enrichment, enabling them to dedicate their specialized skills to complex problem-solving, hypothesis testing, and strategic threat hunting (Baruwal Chhetri et al., 2024). For instance, AI can provide advanced threat intelligence, predict attack trajectories, and offer contextual information, empowering analysts to make more informed and rapid decisions. This augmentation transforms the analyst role from reactive data sifter to proactive strategist and investigator, significantly enhancing the overall intellectual output of the SOC.

4.2.2 Building Context-Aware and Adaptive SOCs

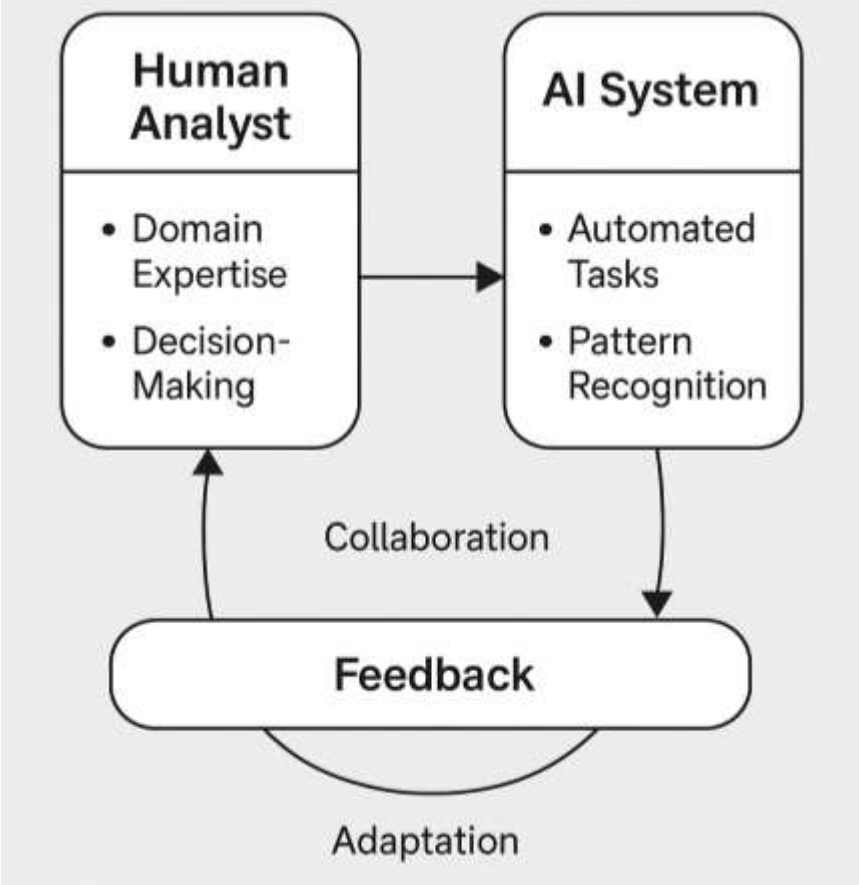
AI-driven capabilities enable SOC 2.0 to become more adaptive and context-aware, moving beyond rigid, rule-based security policies. AI systems can continuously learn from new data, adapt to evolving threat patterns, and dynamically adjust detection thresholds or response playbooks (Yu et al., 2021). This adaptability allows for more proactive defense against novel attack techniques, including sophisticated Advanced Persistent Threats (APTs) (Zhang et al., 2024). Furthermore, AI can integrate contextual information from various sources—such as business operations, user behavior, and geopolitical events—to

prioritize alerts and tailor responses more effectively. For example, an AI system might identify a low-severity alert as high-priority if it involves a critical asset during a period of heightened external tension. This context-aware approach ensures that security resources are optimally allocated, and responses are proportionate to the actual risk, leading to a more resilient and responsive security posture across the organization. The ability to transition between automated, augmented, and collaborative modes of operation further exemplifies this adaptive capacity (Baruwal Chhetri et al., 2024).

4.3 Barriers to Effective Teaming

Despite the substantial opportunities, the realization of effective human-AI teaming in SOCs confronts notable hurdles. These challenges are not merely technical but encompass complex issues related to human cognition, ethical governance, and the inherent vulnerabilities of AI systems themselves. Successfully navigating these complexities is paramount for robust and trustworthy SOC 2.0 deployments.

Figure 2: Human–AI Teaming Model for SOC 2.0



The model presents the dynamic interaction between human analysts and AI systems across five interconnected layers:

1. **Data Collection and Preprocessing:** Continuous ingestion of logs, network data, and threat intelligence.

2. **AI Analytics Layer:** ML/DL-driven anomaly detection, predictive analytics, and event correlation.
3. **Human–AI Interface:** Visualization dashboards and explainable AI outputs facilitating interpretability.
4. **Decision and Response Layer:** Human-guided automated responses through SOAR platforms.
5. **Feedback and Learning Loop:** Mutual adaptation where AI refines models based on analyst feedback and operational outcomes. This model emphasizes cognitive augmentation, trust calibration, and bidirectional learning.

4.3.1 Building Appropriate Trust and Explainability Mechanisms

A central challenge in human-AI teaming is fostering appropriate trust and ensuring AI explainability. Analysts require a clear understanding of how AI systems arrive at their conclusions to calibrate their trust and reliance effectively. Without sufficient explainability, AI systems can become "black boxes," leading to either blind acceptance (over-reliance) or unwarranted skepticism (under-reliance), both of which undermine operational effectiveness (Mehrotra et al., 2024). Developing AI models that provide transparent justifications for their outputs, along with intuitive visualization tools, remains an active area of research. The multifaceted nature of trust in AI, influenced by performance, transparency, and communication, demands a holistic approach to design and measurement (Ulfert et al., 2023)(Ueno et al., 2022). This includes designing systems that communicate uncertainty effectively and allow for human intervention and feedback to refine AI models continuously.

4.3.2 Navigating Ethical, Legal, and Compliance Issues

The deployment of AI in SOCs introduces complex ethical, legal, and compliance considerations. Ethical concerns include potential biases in AI algorithms that could lead to discriminatory outcomes or misidentification of legitimate activities as malicious. Legal implications surround data privacy, particularly when AI processes vast amounts of personal or sensitive organizational data . Compliance with regulations like GDPR or CCPA becomes more intricate with AI, requiring robust mechanisms for data governance and accountability. The question of liability for AI-driven decisions, especially in automated incident response, also presents a significant legal challenge. Furthermore, the transparency obligations, such as those outlined in the European AI Act, highlight the need for clear interpretations and human-centered design to ensure responsible AI use (El Ali et al., 2024). Addressing these issues necessitates the development of clear organizational policies, ethical guidelines, and potentially new regulatory frameworks that keep pace with technological advancements.

4.3.3 Managing Adversarial Risks and Ensuring AI Security

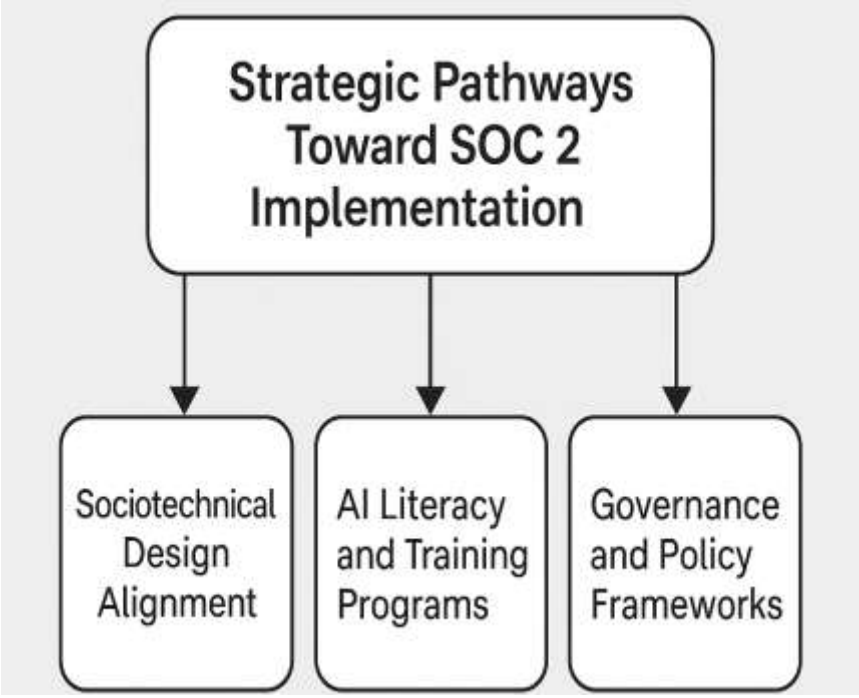
AI systems themselves are targets for adversarial attacks, which pose a significant risk to the integrity and reliability of SOC 2.0 operations. Attackers can employ techniques like data poisoning to subtly corrupt AI training data, causing models to learn incorrect patterns or biases. Evasion attacks can craft malicious inputs that bypass AI detection mechanisms,

rendering defensive AI tools ineffective . The security of the AI pipeline—from data collection and model training to deployment and continuous learning—must be rigorously protected. This includes securing the underlying infrastructure, validating data sources, and employing techniques to detect and mitigate adversarial manipulations. Ensuring AI security requires a proactive defensive posture, treating AI systems as critical assets that require dedicated protection against sophisticated and evolving threats. Without robust AI security, the very tools intended to enhance defense could become significant vulnerabilities, compromising the overall security posture.

4.4 Pathways Forward: Strategic and Practical Recommendations

Advancing human-AI collaboration in SOC 2.0 necessitates a multifaceted approach, encompassing strategic design, organizational development, and focused research. These pathways are designed to address the identified opportunities and challenges, ensuring that AI integration leads to genuinely enhanced security outcomes rather than simply introducing new complexities.

Figure 3: Strategic Pathways Toward SOC 2.0 Implementation



This figure summarizes the four strategic pillars necessary for effective human–AI collaboration in SOC environments:

1. **Sociotechnical Design Alignment:** Integrate human-centered AI with flexible organizational processes.
2. **AI Literacy and Training Programs:** Develop continuous training for analysts to interpret, validate, and leverage AI insights effectively.
3. **Governance and Policy Frameworks:** Establish clear accountability, transparency standards, and ethical AI usage policies.

4. **Research and Innovation Priorities:** Advance explainable AI, adversarial robustness, and empirical trust calibration models.

Table 2. Expanded Research Directions for SOC 2.0

Research Area	**Objective**	**Expected Outcome**
Explainable AI for Security	Develop interpretable ML models for real-time detection	Transparent, auditable AI-driven decisions
Dynamic Trust Calibration	Create adaptive frameworks for real-time trust adjustment	Balanced reliance between human and AI agents
Adversarially Resilient AI	Improve model robustness against evasion and poisoning attacks	Secure AI pipelines and reliable SOC automation
Sociotechnical Optimization	Integrate human factors and cognitive ergonomics into SOC workflows	Improved efficiency and reduced analyst fatigue
Ethical and Regulatory Frameworks	Define standards for privacy, fairness, and accountability	Trustworthy and compliant SOC deployments

Table 2 outlines the primary research directions shaping the evolution of SOC 2.0, emphasizing the intersection of technological innovation and human factors. Each research area identifies a core objective and expected outcome aligned with the broader goals of transparency, trust, and operational resilience in human–AI teaming. The table highlights five focal themes: explainable AI for security, dynamic trust calibration, adversarial robustness, sociotechnical optimization, and ethical/regulatory frameworks. Together, these areas form a roadmap for advancing next-generation SOC capabilities balancing automation with accountability and ensuring that AI-driven systems remain transparent, secure, and human-centered in design.

4.4.1 Designing Sociotechnical Systems for SOC 2.0

Effective human-AI collaboration in SOC 2.0 requires a sociotechnical systems design approach, recognizing the interplay between technology, people, and organizational structures (Naikar et al., 2023) . Design principles should prioritize augmenting human capabilities rather than replacing them, creating flexible and dynamic interfaces that support seamless transitions between automated, augmented, and collaborative modes of operation (Baruwal Chhetri et al., 2024). Key recommendations include:

- **Human-Centered AI Design:** Develop AI systems with human cognitive abilities and workflow needs at the forefront, ensuring clear communication, explainability, and control.

- **Adaptive Interfaces:** Implement interfaces that can adjust to analyst skill levels, task complexity, and real-time operational context, presenting information effectively to minimize cognitive load (Röttger et al., 2009).

- **Feedback Loops:** Integrate mechanisms for continuous feedback between human analysts and AI systems, allowing AI models to learn from human corrections and refine their performance over time.

- **Joint Cognitive Systems:** Conceptualize the SOC as a joint cognitive system where human and AI components contribute to a shared understanding and decision-making process (Naikar et al., 2023).

4.4.2 Fostering Organizational Readiness: Training, Culture, and Policy

Organizational readiness is a critical determinant of successful human-AI collaboration. This involves cultivating an environment that embraces technological change, invests in human capital, and establishes clear governance frameworks. Practical steps include:

- **Comprehensive Training Programs:** Implement ongoing training for SOC analysts to develop AI literacy, understand AI's capabilities and limitations, and learn how to effectively interact with AI tools (Alexandro & Basrowi, 2024).

- **Cultivating a Collaborative Culture:** Foster a culture where AI is perceived as a valuable team member rather than a threat, encouraging experimentation and knowledge sharing between human and AI systems. Leadership support is essential for this cultural shift (Awa et al., 2017).

- **Developing Clear Policies and Governance:** Establish explicit policies regarding AI usage, data privacy, accountability for AI-driven decisions, and the management of vendor risks (Amaka Justina Obinna & Azeez Jason Kess-Momoh, 2024). This includes defining roles and responsibilities for human oversight and intervention.

- **Promoting Cross-functional Teams:** Encourage collaboration between cybersecurity professionals, data scientists, AI engineers, and human factors experts during the design and deployment phases of AI solutions.

4.4.3 Research Directions for Future Human-AI Collaboration in SOCs

Continued research is essential to overcome current limitations and unlock the full potential of human-AI collaboration in cybersecurity. Key areas for future investigation include:

- **Advanced Explainable AI (XAI) for Security:** Developing novel XAI techniques tailored to the specific context of cybersecurity, providing transparent and actionable insights into AI's threat detection and response decisions (Mehrotra et al., 2024).

- **Dynamic Trust Calibration Models:** Research into mechanisms for continuous monitoring and dynamic adjustment of human trust in AI, considering varying levels of AI performance, task criticality, and analyst expertise.

- **Resilient AI against Adversarial Attacks:** Further development of robust AI models and defensive strategies that are resilient to sophisticated adversarial attacks, ensuring the integrity and reliability of AI in high-stakes security contexts .

- **Standardized Evaluation Metrics for Teaming Effectiveness:** Creating comprehensive metrics and methodologies to objectively assess the effectiveness of human-AI teams, beyond individual AI performance, considering factors like joint decision accuracy, response time, and cognitive load .

- **Ethical AI Frameworks for Cybersecurity:** Deeper exploration into ethical AI frameworks specifically designed for cybersecurity applications, addressing issues of bias, fairness, and accountability in automated security decisions.

5 Conclusion

5.1 Synthesis and Implications

The transition to SOC 2.0 signifies a fundamental reimagining of cybersecurity operations, where human expertise and AI capabilities coalesce to form adaptive, intelligent defense ecosystems. This synthesis demonstrates that while AI enhances detection and automation, the human role remains indispensable for contextual judgment, ethical reasoning, and adaptive learning. Effective SOC 2.0 adoption depends on explainable AI, resilient architectures, and sustained investment in human capital. The implications extend to policymakers and practitioners alike calling for ethical governance, standardization of human–AI evaluation metrics, and cross-sector collaboration. The path forward lies in realizing a balanced partnership that safeguards digital infrastructures through responsible innovation, trust, and cognitive augmentation.

5.2 Future Outlook for SOC 2.0 and Human-AI Collaboration

The future of SOC 2.0 is one of continuous evolution, driven by advancements in AI and the dynamic threat landscape. Human-AI collaboration will become increasingly sophisticated, moving towards more autonomous and proactive security operations where AI not only detects but also anticipates and preemptively neutralizes threats with human oversight. Future SOCs will likely feature more adaptive and personalized AI agents that learn individual analyst preferences and optimize their collaboration styles. The integration of cutting-edge AI techniques, such as federated learning for threat intelligence sharing and explainable AI for complex decision justification, will enhance operational capabilities further.

However, this future is contingent on addressing the identified challenges through sustained research and development in areas like robust AI security, dynamic trust

calibration, and comprehensive sociotechnical system design. The emphasis will remain on creating a synergistic relationship where human creativity, intuition, and ethical reasoning are amplified by AI's analytical power and speed. The trajectory suggests a future where SOCs are not merely reactive defense centers but intelligence-driven, adaptive entities capable of operating effectively against the most advanced and persistent cyber adversaries, ultimately securing the digital future through intelligent human-AI partnership (Oswald et al., 2022).

5.3 Limitations and Future Work

This review is limited by its qualitative synthesis and reliance on recent academic sources; empirical validation through case studies of operational SOC 2.0 environments remains a critical next step.

References

Baruwal Chhetri, M., Tariq, S., Singh, R., Jalalvand, F., Paris, C., & Nepal, S. (2024). Towards Human-AI Teaming to Mitigate Alert Fatigue in Security Operations Centres. In *ACM Transactions on Internet Technology* (Vol. 24, Issue 3, pp. 1–22). Association for Computing Machinery (ACM). https://doi.org/10.1145/3670009

Kambhampati, S. (2020). Challenges of Human-Aware AI Systems. In *AI Magazine* (Vol. 41, Issue 3, pp. 3–17). Wiley. https://doi.org/10.1609/aimag.v41i3.5257

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining human-AI teaming the human-centered way: a scoping review and network analysis. In *Frontiers in Artificial Intelligence* (Vol. 6). Frontiers Media SA. https://doi.org/10.3389/frai.2023.1250725

Zhang, B., Gao, Y., Kuang, B., Yu, C., Fu, A., & Susilo, W. (2024). A Survey on Advanced Persistent Threat Detection: A Unified Framework, Challenges, and Countermeasures. In *ACM Computing Surveys* (Vol. 57, Issue 3, pp. 1–36). Association for Computing Machinery (ACM). https://doi.org/10.1145/3700749

Yu, K., Tan, L., Mumtaz, S., Al-Rubaye, S., Al-Dulaimi, A., Bashir, A. K., & Khan, F. A. (2021). Securing Critical Infrastructures: Deep-Learning-Based Threat Detection in IIoT. In *IEEE Communications Magazine* (Vol. 59, Issue 10, pp. 76–82). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/mcom.101.2001126

Kesselheim, A. S., Cresswell, K., Phansalkar, S., Bates, D. W., & Sheikh, A. (2011). Clinical Decision Support Systems Could Be Modified To Reduce 'Alert Fatigue' While Still Minimizing The Risk Of Litigation. In *Health Affairs* (Vol. 30, Issue 12, pp. 2310–2317). Health Affairs (Project Hope). https://doi.org/10.1377/hlthaff.2010.1111

Ulfert, A.-S., Georganta, E., Centeio Jorge, C., Mehrotra, S., & Tielman, M. (2023). Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. In *European Journal of Work and Organizational Psychology* (Vol. 33, Issue 2, pp. 158–171). Informa UK Limited. https://doi.org/10.1080/1359432x.2023.2200172

Naikar, N., Brady, A., Moy, G., & Kwok, H.-W. (2023). Designing human-AI systems for complex settings: ideas from distributed, joint, and self-organising perspectives of

sociotechnical systems and cognitive work analysis. In *Ergonomics* (Vol. 66, Issue 11, pp. 1669–1694). Informa UK Limited. https://doi.org/10.1080/00140139.2023.2281898

Ahmed, N., Ngadi, A. bin, Sharif, J. M., Hussain, S., Uddin, M., Rathore, M. S., Iqbal, J., Abdelhaq, M., Alsaqour, R., Ullah, S. S., & Zuhra, F. T. (2022). Network Threat Detection Using Machine/Deep Learning in SDN-Based Platforms: A Comprehensive Analysis of State-of-the-Art Solutions, Discussion, Challenges, and Future Research Direction. In *Sensors* (Vol. 22, Issue 20, p. 7896). MDPI AG. https://doi.org/10.3390/s22207896

Okon, S. U., Olateju, O. O., Ogungbemi, O. S., Joseph, S. A., Olisa, A. O., & Olaniyi, O. O. (2024). Incorporating Privacy by Design Principles in the Modification of AI Systems in Preventing Breaches across Multiple Environments, Including Public Cloud, Private Cloud, and On-prem. In *Journal of Engineering Research and Reports* (Vol. 26, Issue 9, pp. 136–158). Sciencedomain International. https://doi.org/10.9734/jerr/2024/v26i91269

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. In *ACM Journal on Responsible Computing* (Vol. 1, Issue 4, pp. 1–45). Association for Computing Machinery (ACM). https://doi.org/10.1145/3696449

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022). Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–7). ACM. https://doi.org/10.1145/3491101.3519772

Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. In *Ergonomics* (Vol. 52, Issue 5, pp. 512–523). Informa UK Limited. https://doi.org/10.1080/00140130802379129

El Ali, A., Venkatraj, K. P., Morosoli, S., Naudts, L., Helberger, N., & Cesar, P. (2024). Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. https://doi.org/10.1145/3613905.3650750

Żywiołek, J. (2024). Building Trust in AI-Human Partnerships: Exploring Preferences and Influences in the Manufacturing Industry. In *Management Systems in Production Engineering* (Vol. 32, Issue 2, pp. 244–251). Walter de Gruyter GmbH. https://doi.org/10.2478/mspe-2024-0024

Amaka Justina Obinna, & Azeez Jason Kess-Momoh. (2024). Systematic technical analysis: Enhancing AI deployment in procurement for optimal transparency and accountability. In *Global Journal of Engineering and Technology Advances* (Vol. 19, Issue 1, pp. 192–206). GSC Online Press. https://doi.org/10.30574/gjeta.2024.19.1.0067

Schilling, M. A., & Steensma, H. K. (2001). THE USE OF MODULAR ORGANIZATIONAL FORMS: AN INDUSTRY-LEVEL ANALYSIS. In *Academy of Management Journal* (Vol. 44, Issue 6, pp. 1149–1168). Academy of Management. https://doi.org/10.2307/3069394

Alexandro, R., & Basrowi, B. (2024). Measuring the effectiveness of smart digital organizations on digital technology adoption: An empirical study of educational organizations in Indonesia. In *International Journal of Data and Network Science* (Vol. 8, Issue 1, pp. 139–150). Growing Science. https://doi.org/10.5267/j.ijdns.2023.10.009

Awa, H. O., Ojiabo, O. U., & Orokor, L. E. (2017). Integrated technology-organization-environment (T-O-E) taxonomies for technology adoption. In *Journal of Enterprise Information Management* (Vol. 30, Issue 6, pp. 893–921). Emerald. https://doi.org/10.1108/jeim-03-2016-0079

Oswald, F. L., Endsley, M. R., Chen, J., Chiou, E. K., Draper, M. H., McNeese, N. J., & Roth, E. M. (2022). The National Academies Board on Human-Systems Integration (BOHSI) Panel: Human-AI Teaming: Research Frontiers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, Issue 1, pp. 130–134). SAGE Publications. https://doi.org/10.1177/1071181322661007