

BEYOND AUTOMATION: GENERATIVE AI'S INFLUENCE ON HUMAN DECISION-MAKING, TRUST, AND ACCOUNTABILITY IN INTELLIGENT CYBERSECURITY SYSTEMS

Abdul Mussawar^{1*}, Muhammad Younis Maitlo², Mansoor Ali³, Qalab Hassnain Agha⁴,
Akhter Hussain^{5*}, Sabahat Ahmad Khan⁶, Amna Munawar⁷

¹Department of Computer Science, University of Peshawar, 25120, Khyber Pakhtunkhwa, Pakistan

²Department of Public Administration, Shah Abdul Latif University, Khairpur, Sindh

³Lecturer, Department of Computer System Engineering, Dawood University of Engineering and Technology, Karachi, Sindh, Pakistan

⁴Department of Computer and Software Engineering, NUST College of Electrical and Mechanical Engineering, Islamabad, Pakistan. (0009-0006-6242-2761)

⁵Department of Management & Commerce, University of Swat, Khyber Pakhtunkhwa, Pakistan

⁶Computer Science, National University of Modern Languages (NUML), Islamabad, Pakistan

⁷lahore college for women university (lcwu), punjab, pakistan

^{1*}mussawairabdul@gmail.com, ²younismetlo@hotmail.com, ³mansoor_ali@duet.edu.pk,

⁴aghaqalabhassnain@gmail.com, ⁵241163@uswat.edu.pk, ⁶sabahatkhan198@gmail.com,

⁷amna4989@yahoo.com

DOI: <https://doi.org/10.5281/zenodo.23435109>

Keywords

Generative intelligence; human-AI decision-making; trust; intelligent security operations centers; artificial intelligence; cybersecurity; collaboration; calibrated bias; security systems; security operations centers; AI governance.

Article History

Received: 25 June 2026

Accepted: 16 September 2026

Published: 02 October 2026

Copyright @Author

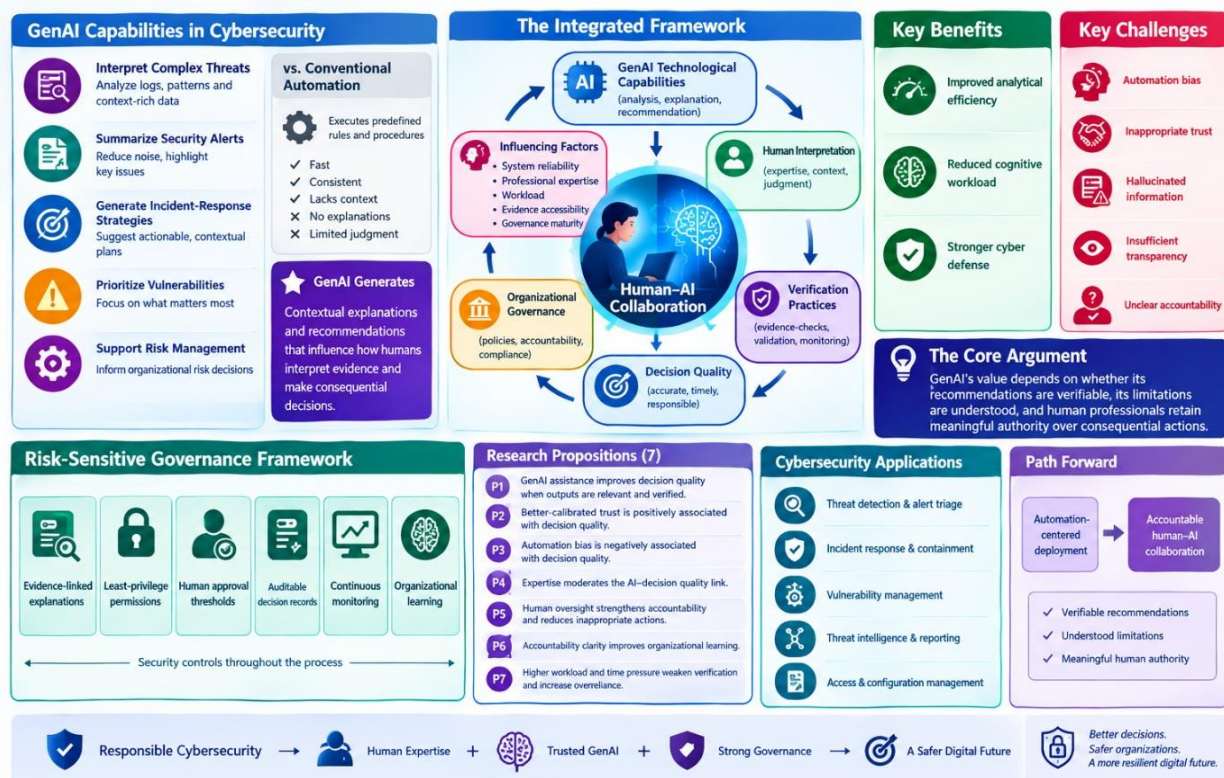
Corresponding Author: *

Abstract

Generative artificial intelligence (GenAI) is transforming cybersecurity by enabling security professionals to interpret complex threat information, summarize security alerts, generate incident-response strategies, prioritize vulnerabilities, and support organizational risk management. Unlike conventional automation, which primarily executes predefined rules and procedures, GenAI generates contextual explanations and recommendations that can influence how humans interpret evidence and make consequential decisions. These capabilities offer opportunities to improve analytical efficiency, reduce cognitive workload, and strengthen cyber defense. However, they also introduce challenges involving automation bias, inappropriate trust, hallucinated information, insufficient transparency, and unclear accountability. This conceptual research article examines the relationship between GenAI, human decision-making, trust calibration, and accountability in intelligent cybersecurity systems. Using a theory-driven narrative synthesis, the article develops an integrated framework connecting technological capabilities, human interpretation, verification practices, decision quality, and organizational governance. The framework identifies system reliability, professional expertise, workload, evidence

Akhter Hussain,
Abdul Mussawar

accessibility, and governance maturity as factors influencing human–AI collaboration. Eight analytical tables summarize technological distinctions, conceptual constructs, research questions, cybersecurity applications, governance mechanisms, research propositions, and implementation priorities. Seven propositions are advanced for future empirical testing. The article argues that GenAI's value depends not simply on its ability to automate tasks but on whether its recommendations remain verifiable, its limitations are understood, and human professionals retain meaningful authority over consequential actions. A risk-sensitive governance framework is proposed, incorporating evidence-linked explanations, least-privilege permissions, human approval thresholds, auditable decision records, continuous monitoring, and organizational learning. The article concludes that responsible cybersecurity requires a transition from automation-centered deployment toward accountable human–AI collaboration, in which generative systems strengthen human analytical capacity without weakening professional judgment or institutional responsibility.



Graphical abstract of the whole study

1. INTRODUCTION

1.1 Background and research context

Cybersecurity has become a critical organizational priority as cloud computing, digital infrastructure, connected devices, online services, and distributed information systems continue to expand. Organizations must monitor networks, applications, endpoints, identities, and data repositories to detect malicious activity and prevent unauthorized access. This responsibility is complicated by the growing volume of security telemetry, increasingly sophisticated cyberattacks, interconnected digital services, and shortages of experienced security professionals. Security operations centers (SOCs) are particularly affected by these challenges. Analysts must interpret alerts, investigate suspicious behavior, correlate information from different sources, distinguish malicious activity from legitimate operations, and recommend appropriate responses. These decisions frequently occur under time pressure and uncertainty. An analyst may need to determine whether an unusual login indicates credential theft, whether a suspicious file is malicious, or whether a potentially compromised endpoint should be isolated.

Conventional security technologies have improved the speed and consistency of many operational tasks. Rule-based systems identify events that satisfy predefined conditions, machine-learning models classify potentially malicious behavior, and security orchestration platforms execute established response procedures. Nevertheless, these technologies depend on the quality of detection logic, data, configurations, and operational assumptions. The development of generative artificial intelligence introduces a different form of computational support. Large language models and related systems can summarize technical reports, interpret security logs, generate investigation queries, compare threat hypotheses, and produce recommendations in natural language. Integrated with security information and event management platforms, endpoint detection systems, threat intelligence repositories, and incident-response tools, these

capabilities can help analysts navigate complex information more efficiently.

However, automation can change human behavior in ways that extend beyond its immediate technical function. Research on human-automation interaction distinguishes appropriate use from misuse, disuse, and poorly designed implementation, demonstrating that the availability of automated assistance does not guarantee appropriate reliance (Parasuraman & Riley, 1997). GenAI therefore represents more than a technical improvement in automation. It can influence how problems are framed, which evidence receives attention, which hypotheses appear plausible, and which responses seem justified. The central question is not simply whether GenAI can perform a security task faster than a human. It is whether the combined human-AI system can make better decisions while preserving adequate oversight and accountability.

1.2 Beyond conventional automation

Conventional automation generally executes predefined operations according to established rules, workflows, or model outputs. For example, a security platform may quarantine an email when its characteristics satisfy a configured detection policy. GenAI can operate at a more interpretive level. It may explain why an event appears suspicious, summarize relationships among several alerts, propose investigative steps, or compare potential containment strategies. This flexibility can reduce the effort required to translate technical evidence into actionable assessments.

Nevertheless, generated output is not equivalent to verified knowledge. A model may produce an inaccurate explanation, overlook relevant context, invent a technical detail, or recommend a procedure that is unsuitable for the target environment. Its language may appear authoritative even when its conclusion is unsupported. The distinction becomes particularly important when a recommendation affects operational continuity, confidentiality, or system integrity. A mistaken summary may have limited consequences when used for preliminary understanding. The same error can be

substantially more damaging if it triggers account suspension, endpoint isolation, firewall modification, or automated changes to production infrastructure. Research on trust in automation emphasizes that trust should support appropriate reliance rather than simply maximize users' confidence in a system (Lee & See, 2004). Applied to GenAI, this principle suggests that organizations must assess the relationship between demonstrated system capability, the operational context, and the consequences of error.

1.3 The human dimension of cybersecurity intelligence

Cybersecurity decisions depend on professional judgment as well as technical information. Analysts must evaluate evidence credibility, consider alternative explanations, estimate the consequences of competing actions, and determine when escalation is necessary. GenAI may improve these processes by reducing information overload and providing structured summaries. However, it may also introduce cognitive risks. An analyst who receives a confident AI-generated explanation may become anchored to its initial interpretation and overlook evidence that contradicts it.

The design of human-AI interaction is therefore important. Research on human-AI interaction guidelines highlights the importance of communicating system capabilities, supporting appropriate user control, and enabling users to recover when system behavior is unexpected (Amershi et al., 2019). These considerations are especially relevant in security operations, where decisions must frequently be made under uncertainty. Automation can support detection and response, but overreliance may cause analysts to neglect contradictory information or fail to monitor automated systems adequately. Such concerns have been examined specifically in the context of SOC automation and analyst behavior (Tilbury & Flowerday, 2024).

1.4 Trust, accountability, and decision quality

Trust determines how professionals respond to AI-generated recommendations. Excessive trust can

encourage uncritical acceptance, whereas insufficient trust can prevent the effective use of reliable assistance. Appropriate trust requires alignment between perceived reliability and demonstrated capability. This distinction is important because the linguistic fluency of generative systems can create a misleading impression of certainty. Users may interpret detailed explanations as evidence of technical correctness even when the underlying claim is unsupported.

Trust also develops through experience. Users' reliance may change after observing successful recommendations, encountering errors, receiving feedback, or working under different operational conditions. A comprehensive review of empirical research on trust in automation identifies dispositional, situational, and learned influences on trust formation (Hoff & Bashir, 2015). Accountability presents a related challenge. In an AI-supported security workflow, the model may generate a recommendation, an analyst may approve it, and an integrated system may execute the action. If the outcome is harmful, responsibility cannot be understood solely by identifying the person who clicked an approval button. Investigators must also consider the model's limitations, available evidence, authorization process, system permissions, and organizational governance. Accountability consequently requires identifiable decision rights, accessible evidence, appropriate human authority, and reliable records of decisions and actions.

1.5 Research problem and significance

GenAI deployment creates a potential gap between technical capability and organizational readiness. A system may generate useful recommendations without ensuring that users can verify them. An analyst may express confidence in a recommendation without correctly identifying its errors. An organization may formally assign responsibility to human operators without providing sufficient information, time, or authority for meaningful review. The problem is therefore socio-technical: technical performance, human behavior, and institutional arrangements

interact to determine cybersecurity outcomes. Generative systems also create a particular challenge because their outputs are produced through statistical language generation rather than guaranteed factual verification. Research examining the limitations of large language models has highlighted how apparently coherent language can coexist with unreliable factual representations (Bender et al., 2021).

Similarly, the broader foundation-model literature emphasizes that the risks and capabilities of a model depend on its training data, deployment context, downstream applications, and surrounding technical infrastructure (Bommasani et al., 2021). The central problem addressed in this article is how GenAI can improve cybersecurity decision-making while preserving calibrated trust, professional judgment, and institutional accountability.

1.6 Research objectives

The article has five objectives:

- i. To examine how GenAI changes human decision-making in intelligent cybersecurity systems.
- ii. To analyze the relationship between system reliability, trust calibration, and automation bias.
- iii. To identify organizational and technical factors affecting meaningful human oversight.
- iv. To propose an integrated framework for accountability in GenAI-assisted cybersecurity.
- v. To establish research propositions and methodological directions for future empirical studies.

1.7 Research questions

Four research questions guide the analysis:

- I.RQ1: How does GenAI influence human judgment and decision quality in intelligent cybersecurity systems?
- II.RQ2: Which technological, cognitive, and organizational factors shape cybersecurity professionals' trust in AI-generated recommendations?
- III.RQ3: How can organizations maintain meaningful human oversight and accountability when GenAI contributes to consequential security decisions?

- IV.RQ4: Which governance mechanisms can support effective and responsible human-AI collaboration in cybersecurity?

2. Literature Review and Theoretical Foundations

2.1 Intelligent cybersecurity and the evolution of automation

Cybersecurity technologies have developed from static rules and signature-based detection toward machine-learning classification, behavioral analytics, automated orchestration, and generative assistance. Rule-based systems are generally inspectable because their triggering conditions can be examined. Their limitations arise when predefined rules fail to recognize unfamiliar patterns or generate excessive alerts when legitimate behavior resembles malicious activity.

Machine-learning systems can identify patterns in large datasets and detect relationships that are difficult to express through manually developed rules. Nevertheless, their predictions depend on data quality, feature selection, model assumptions, and deployment conditions. GenAI introduces the ability to produce contextual narratives and recommendations. A generative assistant may summarize several security events into an incident timeline, explain a vulnerability to a nontechnical manager, or propose investigative queries.

These capabilities can improve accessibility and reduce the time needed to interpret technical information. However, a generated explanation is not necessarily an independent verification of the underlying detection result. If the initial alert is incorrect or the evidence is incomplete, the generated narrative may reinforce the original error. The growing use of AI also increases the importance of monitoring dependencies and maintaining operational discipline. Research on technical debt in machine-learning systems demonstrates that the long-term reliability of AI applications depends on their surrounding infrastructure, data dependencies, monitoring, and maintenance practices (Sculley et al., 2015). Accordingly, GenAI should be treated as a component of a larger cybersecurity architecture rather than as a replacement for conventional

detection, deterministic controls, human expertise, and established incident-response procedures.

2.2 Automation bias and analyst behavior

Automation bias occurs when people rely excessively on automated recommendations or fail to identify errors that could be detected through independent assessment. In cybersecurity, this may involve accepting an incorrect threat classification, overlooking a suspicious event because an AI assistant labels it benign, or executing an inappropriate response without sufficient review. The concept is not unique to generative systems. Earlier research on automated decision aids demonstrated that computer-generated recommendations can influence human judgment even when those recommendations are incorrect.

GenAI may introduce additional dimensions to this problem. A traditional alert may display a score or classification, whereas a generative assistant can provide a coherent explanation and an apparently complete investigative narrative. The narrative may reduce cognitive effort, but it may also discourage further investigation if the user assumes that the system has considered all relevant evidence. The relevant concern is not only whether the human remains formally involved. It is whether the human continues to exercise independent judgment.

2.3 Trust calibration and reliance

Trust calibration refers to the degree to which reliance on a system corresponds to its demonstrated capabilities and limitations. Calibrated trust does not mean trusting the system more. It means relying on it when appropriate, verifying recommendations when necessary, and rejecting outputs when evidence or task requirements justify doing so.

GenAI makes calibration challenging because perceived reliability can be influenced by language quality, response speed, familiarity, and apparent completeness. These characteristics may affect confidence independently of actual correctness. For example, an AI-generated incident summary

may confidently identify a compromised account but omit an important contextual fact. An analyst who equates confidence with accuracy may escalate the incident without checking the original logs. Conversely, an analyst who rejects all AI-generated summaries may spend excessive time on routine tasks and fail to benefit from accurate information synthesis. Organizations should therefore assess trust behaviorally. Relevant measures include whether users accept correct recommendations, reject incorrect ones, verify uncertain outputs, and adjust their reliance after receiving evidence of system performance.

2.4 Human oversight and cognitive support

Human oversight is often treated as a binary property: either a human participates in a decision, or the decision is automated. This distinction is insufficient for GenAI-supported cybersecurity. A human may be formally involved but lack the information, expertise, authority, or time needed to evaluate a recommendation. Conversely, a well-designed AI assistant may perform many low-risk analytical tasks while leaving consequential decisions subject to meaningful human control.

Effective oversight should therefore be evaluated through the quality of the review process rather than the mere presence of an approval step. Cognitive support should reduce unnecessary effort without removing the analyst's ability to reason independently. This requires access to supporting evidence, opportunities to compare alternative interpretations, and clear procedures for rejecting or modifying recommendations.

2.5 Accountability and organizational governance

Accountability concerns the ability to explain decisions, identify responsible actors, assess compliance with established procedures, and implement corrective action when failures occur. AI governance frameworks emphasize that risks must be considered throughout the system lifecycle, from initial design and development to deployment, monitoring, and retirement. NIST's AI Risk Management Framework provides a

general structure for identifying, assessing, and managing these risks (Tabassi, 2023).

Its generative AI profile extends this approach to the characteristics and risks of generative systems, including challenges that require attention during implementation and operational use (Autio et al., 2024). Cybersecurity governance must also consider the risks introduced by AI components themselves. These may include insecure integrations, excessive permissions, malicious inputs, model-related weaknesses, and dependencies on external services. ENISA's work on AI cybersecurity and standardization provides additional context for addressing these concerns across the system lifecycle (ENISA, 2023).

A robust accountability model must therefore distinguish technical responsibility, operational responsibility, managerial responsibility, and governance responsibility. Developers and vendors may be responsible for aspects of system design and documentation; security teams must validate outputs and operate systems within defined boundaries; managers must establish authorization and escalation procedures; and organizational leaders must ensure appropriate governance and resources.

2.6 Research gap

The literature establishes the importance of trust, automation bias, human oversight, and AI risk management. However, these elements should not be treated as independent dimensions. Technical accuracy does not necessarily produce appropriate reliance. High user confidence does not necessarily produce correct decisions. Formal human approval does not necessarily ensure meaningful oversight. Complete audit records do not automatically prevent errors, although they can support investigation and learning. The research gap addressed by this article concerns the interaction among four central dimensions:

- i. GenAI capability and output reliability.
- ii. Human interpretation and trust calibration.
- iii. Decision quality and operational consequences.
- iv. Accountability and organizational governance.

The proposed framework integrates these dimensions to explain why the effectiveness of

GenAI in cybersecurity depends on the design of the entire human-AI decision process.

3. Research Methodology

3.1 Research design

This study adopts a conceptual research design based on theory development and narrative literature synthesis. It does not report original experiments, interviews, surveys, or quantitative measurements. The purpose is to integrate established concepts from human-automation research, cybersecurity operations, and AI governance into a coherent framework explaining how GenAI influences decision-making, trust, and accountability. A conceptual approach is appropriate because the article seeks to define relationships among constructs and develop propositions for future empirical testing. It provides a foundation for subsequent empirical studies without claiming causal relationships that have not been investigated directly. The design follows the principle that an integrative literature review should build new conceptual understanding rather than merely summarize individual publications (Torraco, 2005).

3.2 Literature synthesis approach

The analysis uses a narrative synthesis to identify concepts and relationships relevant to the research questions. A narrative approach is useful for integrating literature from multiple disciplines when the evidence includes different theoretical perspectives and research designs. The synthesis is structured around three broad source categories:

- ❖ Human factors, including trust, automation bias, and human-AI interaction.
- ❖ Cybersecurity operations, including alert triage, incident response, and security decision-making.
- ❖ AI governance, including accountability, risk management, and secure system deployment.

The approach is informed by established guidance for conducting literature reviews that integrate and critically organize research across related areas (Webster & Watson, 2002).

3.3 Conceptual development procedure

The conceptual development process involves four stages.

First, relevant constructs are identified, including output reliability, trust calibration, automation bias, expertise, workload, oversight, accountability, and decision quality.

Second, the constructs are grouped into technological, human, organizational, and outcome-related categories.

Third, relationships among the constructs are developed into a conceptual framework. For example, system reliability may influence perceived reliability, which in turn influences human reliance. Verification and accountability mechanisms may shape whether that reliance produces appropriate outcomes.

Fourth, the relationships are expressed as research propositions that can be examined through controlled experiments, surveys, organizational case studies, or longitudinal research.

This theory-building approach is consistent with methodological work emphasizing the importance of clearly defined constructs, explanatory relationships, and propositions that extend existing understanding (Jaakkola, 2020).

3.4 Analytical framework

The analysis distinguishes between the system's actual performance and the user's perception of that performance. It also separates human oversight from nominal approval, and accountability from the mere existence of technical logs. These distinctions are important because the constructs may be related without

being interchangeable. A system can be technically reliable but poorly understood by users; an analyst can trust a system without verifying its outputs; and an organization can maintain records without assigning clear decision authority. The framework is therefore designed to explain how these dimensions interact rather than treating any single factor as a sufficient condition for responsible deployment. The theory-development process follows the broader principle that useful conceptual contributions clarify relationships among constructs and explain why those relationships matter in a particular context (MacInnis, 2011).

3.5 Scope and methodological limitations

The article focuses on organizational cybersecurity applications, particularly threat triage, incident response, vulnerability management, and security intelligence (Table 1). It considers both human-facing assistants and systems integrated with operational tools. Because the study is conceptual, it does not estimate the strength of relationships among variables or establish the effectiveness of particular interventions. The propositions should be interpreted as theoretically grounded expectations requiring empirical evaluation. The narrative synthesis also does not claim exhaustive database coverage or systematic screening of every potentially relevant publication. Future research could use a systematic review protocol to assess the scope, quality, and consistency of empirical findings.

Table 1. Conceptual constructs and proposed indicators

Construct	Definition	Illustrative indicators
GenAI assistance	Use of generative systems to interpret evidence or generate recommendations	Frequency of use, task coverage, contextual relevance
Output reliability	Accuracy and appropriateness of generated outputs	Correctness, consistency, evidence support
Perceived reliability	User assessment of system capability	Perceived accuracy, confidence, consistency
Trust calibration	Alignment between reliance and demonstrated capability	Appropriate acceptance, rejection, and verification
Automation bias	Undue reliance on AI recommendations	Acceptance of incorrect outputs, missed contradictory evidence

Human oversight	Capacity to examine, challenge, approve, or reject recommendations	Review quality, authority, verification
Accountability clarity	Clarity concerning responsibility for decisions	Defined decision rights, escalation, audit completeness
Decision quality	Accuracy and defensibility of security decisions	Correct classification, justified action, procedural compliance
Cybersecurity resilience	Capacity to detect, respond to, recover from, and learn from incidents	Response effectiveness, recovery, corrective learning

4. Proposed Conceptual Framework

4.1 Framework overview

The proposed framework explains how GenAI can influence cybersecurity outcomes through the interaction of system reliability, human interpretation, trust calibration, verification, and accountability (Figure 1). The process begins when a generative system receives security-related information and produces an interpretation or recommendation. A professional evaluates the output in light of expertise, workload, perceived reliability, available evidence, and organizational

expectations. Trust influences the degree of reliance. The professional may accept the recommendation, verify it, modify it, or reject it. Governance mechanisms establish the boundaries within which those responses occur and determine whether consequential actions require additional authorization. The resulting decision affects operational outcomes, while audit records and post-incident reviews can provide feedback for improving the model, workflow, and governance process.

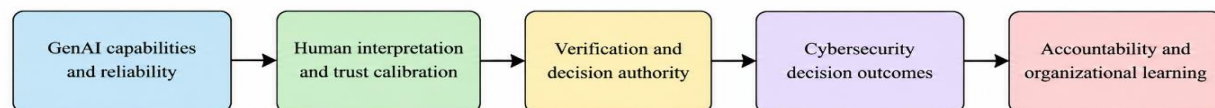


Figure 1. Conceptual framework of GenAI-assisted cybersecurity decision-making

4.2 Effective and ineffective collaboration pathways

In an effective collaboration pathway, the system provides useful information, the professional evaluates the recommendation, and the resulting action reflects both AI-generated analysis and independent evidence. Human review remains meaningful, and the decision can be reconstructed through appropriate records. In a failure pathway, the model produces a plausible but inaccurate recommendation, the analyst accepts it without sufficient verification, and the organization executes an inappropriate action. Weak authorization procedures or inadequate logging

may make the failure difficult to detect or investigate. The same technology can contribute to either pathway. The difference depends on system capability, human behavior, and organizational controls.

4.3 Moderating and mediating factors

Professional expertise may improve the ability to identify unsupported conclusions, although experience alone cannot eliminate automation bias. Workload and time pressure may reduce verification, particularly when analysts must process many alerts or operate during major incidents. Evidence accessibility may improve

decision quality by reducing the effort required to inspect underlying information. Governance maturity may influence whether users have sufficient authority, training, and time to challenge AI recommendations. Task consequences may determine the appropriate level of review. A low-risk summary requires different controls from an action that can interrupt critical infrastructure. Trust is proposed as a potential mediator between system characteristics and reliance, while expertise, workload, and governance maturity may moderate the relationship between GenAI assistance and decision outcomes.

5. GenAI Applications in Cybersecurity

5.1 Threat detection and alert triage

Alert triage involves determining which security events warrant immediate investigation, escalation, or closure. GenAI can summarize related alerts, identify common indicators, organize event timelines, and explain why a pattern may be suspicious. Such support may reduce the time analysts spend interpreting fragmented technical information. It may also help less experienced personnel understand unfamiliar terminology and investigative procedures. However, generated summaries can omit important details or incorrectly associate events that are not causally related. An apparently coherent narrative may encourage analysts to accept an incorrect classification or overlook a benign explanation. Organizations should ensure that generated summaries link to the underlying telemetry and that analysts can inspect relevant evidence. Decisions involving significant operational consequences should not rely solely on an unverified generated narrative.

5.2 Incident response and containment

Incident response involves coordinating actions to identify, contain, investigate, eradicate, and recover from security incidents. GenAI can assist by drafting response plans, summarizing affected assets, identifying relevant procedures, and preparing communications. These capabilities may improve coordination, particularly when

incidents involve multiple teams and complex technical information. Nevertheless, a response recommendation may be inappropriate if it ignores dependencies, business continuity requirements, or the possibility of false detection. Isolating a compromised workstation may be justified, whereas disabling a critical identity service could disrupt an entire organization. Risk-sensitive authorization should determine which actions can proceed automatically, which require analyst approval, and which require incident commander or managerial authorization. Systems should also support rollback or recovery when actions have unintended effects.

5.3 Vulnerability management

GenAI can summarize vulnerability advisories, explain technical weaknesses, compare remediation options, and help prioritize corrective work. Its recommendations should be evaluated against asset criticality, exposure, exploitability, available compensating controls, and the potential consequences of configuration changes. A technically plausible remediation may still be unsuitable for a specific application or infrastructure environment. Generated scripts and configuration changes should be tested in controlled environments before production deployment. Access to execution tools should be limited to the permissions required for the task, and consequential modifications should follow established change-management procedures.

5.4 Threat intelligence and security reporting

Threat intelligence teams collect information from technical indicators, vendor reports, incident investigations, public advisories, and other sources (Table 2). GenAI can synthesize these materials into concise summaries and identify potential relationships among observed events. The principal risk is that generated summaries may combine verified facts with inference or unsupported attribution. Reports should distinguish direct observations, source-reported claims, analyst assessments, and model-generated hypotheses. Generative systems can also help prepare executive briefings and incident

summaries. However, simplification should not remove uncertainty or distort incident severity. Decision-makers may be misled if a tentative hypothesis is presented as a confirmed finding.

Table 2. Cybersecurity applications, benefits, and risks

Application	Potential contribution	Principal risk	Recommended control
Alert triage	Summarizes and prioritizes events	Incorrect prioritization	Verify source telemetry
Threat hunting	Suggests hypotheses and queries	Confirmation bias	Require alternative explanations
Incident response	Drafts response plans	Inappropriate containment	Risk-based approval
Vulnerability management	Explains weaknesses and remediation	Unsuitable configuration changes	Validate and test recommendations
Threat intelligence	Synthesizes reports and indicators	Unsupported attribution	Preserve provenance and corroborate claims
Security reporting	Produces technical and executive summaries	Misrepresentation of uncertainty	Analyst review and evidence checks
Access management	Helps assess access anomalies	Unjustified suspension or privilege changes	Explicit authorization
Configuration management	Drafts policies and scripts	Misconfiguration or exposure	Controlled testing and rollback

5.5 Decision quality as a multidimensional outcome

Decision quality should not be reduced to the speed of completing a task. A fast decision may be inappropriate if it is based on inaccurate information or creates unnecessary operational disruption. A comprehensive evaluation should consider accuracy, timeliness, consistency, evidence quality, procedural compliance, reversibility, and consequences for organizational security. It should also examine the quality of the decision process. A correct outcome reached through unsupported reasoning may not be reliable when applied to future incidents. Conversely, a reasonable decision based on available evidence may still produce an unfavorable outcome because cybersecurity investigations involve uncertainty. Evaluation should therefore distinguish between outcome correctness and decision-process quality. Both are relevant to responsible GenAI deployment.

6. Research Propositions

The conceptual framework supports seven propositions for future empirical investigation (Table 3). These propositions are theoretical expectations rather than findings from original data.

Proposition 1: GenAI assistance and decision quality. GenAI assistance is expected to improve cybersecurity decision quality when outputs are relevant, evidence-supported, and appropriately verified.

Proposition 2: Trust calibration and decision quality. Better-calibrated trust is expected to be positively associated with decision quality because it supports appropriate acceptance and rejection of recommendations.

Proposition 3: Automation bias and decision quality. Greater automation bias is expected to be negatively associated with decision quality, particularly when the system produces inaccurate recommendations and analysts fail to inspect contradictory evidence.

Proposition 4: Expertise as a moderator. Cybersecurity expertise is expected to moderate the relationship between GenAI assistance and decision quality, particularly when experienced analysts can recognize unsupported conclusions and independently evaluate evidence.

Proposition 5: Human oversight and accountability. Meaningful human oversight,

supported by clear authority and adequate review time, is expected to strengthen accountability and reduce inappropriate high-impact actions.

Proposition 6: Accountability clarity and organizational learning. Clear responsibility allocation and traceable decision records are expected to improve post-incident analysis and corrective action.

Proposition 7: Workload and verification. Higher workload and time pressure are expected to weaken verification practices and increase the likelihood of overreliance on AI-generated recommendations.

Table 3. Research propositions and empirical evaluation

Proposition	Predictor	Outcome	Suggested method
P1	GenAI assistance quality	Decision quality	Controlled experiment
P2	Trust calibration	Decision quality	Regression or structural equation modeling
P3	Automation bias	Decision quality	Scenario-based experiment
P4	Expertise	Relationship between AI assistance and decision quality	Moderation analysis
P5	Meaningful human oversight	Accountability and inappropriate-action rates	Comparative study
P6	Accountability clarity	Organizational learning	Interviews and incident-record analysis
P7	Workload and time pressure	Verification behavior and automation bias	Factorial experiment or field study

7. Discussion

7.1 From automation efficiency to decision effectiveness

The central implication of the framework is that cybersecurity organizations should evaluate GenAI according to its contribution to decision effectiveness rather than speed alone. Automation can reduce routine workload, but a reduction in time per task does not necessarily indicate an improvement in security outcomes. A system that rapidly generates incorrect recommendations may increase unnecessary investigations, encourage inappropriate containment, or delay attention to genuine threats. The discussion of the findings is

conceptual; the following implications are derived from the framework and relevant literature rather than from original empirical results.

A useful evaluation should consider accuracy, response time, false-positive and false-negative rates, workload, procedural compliance, and the consequences of inappropriate action. It should also examine whether analysts can explain why an action was justified, which evidence supported it, and what uncertainty remained. Decision-process quality is essential when decisions must withstand internal review, regulatory scrutiny, or post-incident investigation.

7.2 Explainability must support verification

Explainability is often proposed as a mechanism for improving user understanding of AI-generated recommendations. However, an explanation is useful only if it helps users assess the underlying decision rather than simply making the output appear convincing. Research on interpretable machine learning distinguishes between the ability to explain a model's behavior and the practical usefulness of an explanation for a particular audience or decision (Ribeiro et al., 2016).

For cybersecurity, the practical objective should be evidence-supported explanations that enable analysts to inspect source records, compare alternative hypotheses, and identify missing information. The distinction between explanation and interpretability also matters because a post hoc explanation may not faithfully represent the internal process that generated a prediction. A framework for evaluating explainable AI emphasizes the need to define interpretability goals and assess them through appropriate evaluation methods (Doshi-Velez & Kim, 2017). Organizations should therefore avoid treating explanatory detail as proof of reliability. Recommendations should be validated against source evidence and task-specific performance.

7.3 Trust must be grounded in demonstrated capability

Trustworthy GenAI deployment requires separating perceived quality from actual reliability. Fluent language, detailed explanations, and rapid responses can influence confidence without guaranteeing correctness. Organizations should establish task-specific performance evaluations and provide users with access to supporting evidence. Trust should be reinforced by demonstrated capability, not persuasive presentation alone. The appropriate objective is not maximum trust but appropriate reliance. A well-designed system should help users recognize when it is useful, when it is uncertain, and when its recommendation requires independent confirmation.

7.4 Human oversight must be substantive

Human oversight should be understood as a combination of information, competence, authority, and time. A reviewer who cannot access underlying evidence may be unable to assess a recommendation. A reviewer who lacks technical knowledge may not recognize a flawed conclusion. A reviewer who is prohibited from rejecting the system's recommendation lacks meaningful authority. A reviewer facing unrealistic workloads may approve actions without sufficient examination. These limitations demonstrate why human presence alone is not enough. Organizations must design workflows that make review practical and effective. The framework also supports a distinction between oversight and accountability. Oversight concerns how a decision is examined before or during execution, whereas accountability includes the ability to explain the decision, identify responsibilities, and implement corrective action afterward.

7.5 Transparency and the limits of model explanations

Transparency should be designed around the needs of the people responsible for evaluating a recommendation. A system may not need to expose every internal computational detail, but it should make relevant evidence, limitations, and uncertainties accessible. Research on interpretable machine learning has emphasized that high-stakes applications require careful consideration of whether explanations are sufficiently faithful and useful for the intended decision (Rudin, 2019). Evaluation should also account for the context in which the model is used. A model's documentation, known limitations, intended applications, and evaluation conditions should be available to people responsible for deployment and governance. Model documentation provides one way to communicate these characteristics systematically (Mitchell et al., 2019). In cybersecurity, these practices can support better decisions by helping analysts distinguish verified evidence from generated interpretations.

7.6 Hallucination and evidence integrity

Generative systems can produce plausible but inaccurate information, including invented technical details, unsupported causal explanations, or incorrect references to evidence. These errors can be particularly consequential when the output is integrated into incident-response workflows. A broad survey of hallucination in natural-language generation identifies multiple forms and causes of generated factual errors and emphasizes the continuing challenge of detecting and mitigating them (Ji et al., 2023). For cybersecurity, the implication is that generated statements should be treated as candidate interpretations rather than automatically verified facts. Evidence retrieval, source validation, and independent checks are particularly important when recommendations affect access controls, system configurations, or incident containment. Organizations should also distinguish uncertainty about the model's answer from uncertainty inherent in the security incident itself. Even a correctly functioning model may be unable to resolve an ambiguous event when the available evidence is incomplete.

7.7 AI-related risks and secure system design

Generative systems introduce risks through their inputs, outputs, integrations, permissions, and dependencies. Malicious instructions embedded in retrieved documents or external content may attempt to influence model behavior. Excessive permissions can allow an AI-supported workflow to perform actions beyond its intended scope. Research on ethical and operational risks associated with language models highlights the need to examine risks across the model lifecycle rather than treating the model as an isolated component (Weidinger et al., 2021). Security design should therefore include least-privilege access, input isolation, controlled tool execution, monitoring, and clear authorization boundaries. Where possible, informational tasks should use read-only access, while consequential actions should be mediated through explicit approval procedures.

7.8 Human control and the design of autonomous systems

The appropriate level of automation depends on the task, the consequences of error, and the ability to detect and reverse mistakes. Low-risk tasks may be suitable for greater automation, while ambiguous or high-impact decisions require more deliberate human involvement. Research on human-centered AI emphasizes the importance of combining automation with reliable human control, effective feedback, and systems that support people rather than merely replacing their activities (Shneiderman, 2020). For cybersecurity, this means that organizations should grant autonomy incrementally. A system should demonstrate task-specific reliability before receiving broader operational permissions, and the organization should retain the ability to intervene or suspend its operation.

7.9 Explanations should be evaluated through user performance

An explanation is not successful simply because users report that it is clear or satisfying. Its value depends on whether it helps users identify errors, understand limitations, and make better decisions. Research on explainable AI in practice emphasizes the importance of evaluating explanations in relation to actual users, tasks, and organizational settings (Kaur et al., 2020). This principle is relevant to SOC interfaces. A dashboard that displays a generated narrative should be evaluated by examining whether analysts detect incorrect recommendations more effectively, consider alternative hypotheses, and make defensible decisions. Interfaces should not encourage passive acceptance through visual emphasis or authoritative wording. They should make it straightforward to inspect evidence and challenge the recommendation.

7.10 Human-AI collaboration requires accurate mental models

Professionals need an accurate understanding of what a system can and cannot do. If users assume that an AI assistant has independently verified every claim, they may rely on its recommendations

beyond appropriate limits. Research on human-AI collaboration indicates that effective teaming depends on how people understand AI capabilities and on whether collaboration mechanisms improve performance beyond the capabilities of either participant alone (Bansal et al., 2021). Organizations should therefore train staff to understand the distinction between generated explanations, verified observations, model predictions, and operationally authorized actions. This training should be accompanied by realistic exercises in which systems produce both correct and incorrect recommendations. The purpose is to develop appropriate reliance rather than unconditional trust or generalized distrust.

7.11 Accountability as a lifecycle responsibility

Accountability should extend across system development, procurement, deployment, use, monitoring, and retirement. A failure may originate in an inaccurate model output, but the organizational causes can include inadequate testing, inappropriate permissions, insufficient training, or poorly designed approval procedures. Research proposing internal algorithmic auditing frameworks emphasizes the importance of structured processes for evaluating systems and identifying accountability gaps throughout development and deployment (Raji et al., 2020). Organizations should establish identifiable owners for each AI-supported workflow, document its intended use, maintain appropriate records, and specify who may authorize or reverse consequential actions. Post-incident review should investigate both immediate technical failures and the wider organizational conditions that allowed them to affect operations.

7.12 Accountability must consider the broader sociotechnical system

AI decisions do not occur independently of organizational practices, professional incentives, technical infrastructure, or institutional constraints. A model may perform adequately under controlled conditions but produce poor outcomes when integrated into a workflow with unrealistic time limits or insufficient review.

Research on fairness and abstraction in sociotechnical systems demonstrates the risks of evaluating technical systems without accounting for the context in which they operate (Selbst et al., 2019). Although the specific subject of that work is fairness, the broader methodological lesson is relevant to cybersecurity governance: technical performance should not be separated from the environment in which a system is deployed. An accountable cybersecurity organization must therefore assess not only the model but also the human and organizational processes surrounding it.

7.13 Integrating AI governance with cybersecurity risk management

GenAI governance should be integrated into established cybersecurity management rather than treated as an isolated compliance activity. The NIST Cybersecurity Framework 2.0 provides a structure for organizing cybersecurity risk management and governance across organizations (Pascoe et al., 2024). Integrating AI-specific requirements into existing risk processes can help organizations coordinate security monitoring, incident response, access management, and AI oversight. Similarly, an AI management system can support organizational responsibilities by defining policies, roles, risk assessments, and continual improvement processes. The international AI management-system standard provides a framework for establishing and improving organizational AI governance (International Organization for Standardization [ISO], 2023a). Risk assessment should also be continuous because models, integrations, and threat environments change. The international guidance on AI risk management offers a basis for integrating risk identification and treatment into AI-related activities (ISO, 2023b).

7.14 Operational security controls for language-model applications

GenAI systems can be exposed to prompt injection, sensitive-information disclosure, insecure output handling, excessive agency, and other application-level risks. The OWASP

guidance for large language model applications provides a security-oriented classification of risks and associated mitigation considerations (OWASP Foundation, 2025). In cybersecurity deployments, these risks should be addressed through system architecture, permission boundaries, secure input handling, output validation, and monitoring. A generative assistant should not be able to bypass established security controls simply because it has been given access to operational tools. Organizations should also test adversarial scenarios before deployment and repeat testing when the model, tools, or permissions change.

7.15 Practical governance implications

The combined framework suggests that organizations should evaluate GenAI across four dimensions:

- **Technical reliability:** whether outputs are accurate, relevant, and supported by evidence.
 - **Human performance:** whether analysts interpret and challenge recommendations appropriately.
 - **Operational safety:** whether permissions, approvals, and rollback mechanisms limit the consequences of errors.
 - **Organizational accountability:** whether decisions can be reconstructed, responsibilities identified, and failures corrected.
- These dimensions should be evaluated together. Strong performance in one dimension cannot fully compensate for serious weaknesses in another. All these result are depicted in Table 4.

Table 4. Governance controls and evaluation measures

Governance area	Recommended control	Responsible function	Evaluation measure
Model validation	Test representative tasks and failure scenarios	AI engineering and security teams	Task-specific error rates
Evidence quality	Link recommendations to verifiable sources	Developers and analysts	Evidence completeness
Human oversight	Define approval thresholds	SOC leadership	Compliance with review requirements
Trust calibration	Train staff in verification and error recognition	Security management	Appropriate reliance behavior
Access control	Apply least privilege and bounded execution	Security architecture teams	Unauthorized-action rates
Auditability	Maintain secure decision records	Governance and operations teams	Record completeness
Privacy	Minimize sensitive inputs and retention	Data protection teams	Privacy incidents and compliance
Vendor governance	Review dependencies and model changes	Procurement and security governance	Supplier assessment completion
Incident learning	Investigate failures and near misses	Incident response teams	Corrective actions completed
Continuous evaluation	Monitor system and human performance	AI owners and SOC leadership	Control effectiveness over time

7.16 Limitations and future research

This article develops a conceptual framework rather than testing relationships with original data. Its propositions should therefore be interpreted as

hypotheses for future research rather than established causal findings. The narrative synthesis does not constitute a systematic literature review. A future review could employ predefined

databases, search strings, inclusion criteria, screening procedures, and quality assessments to evaluate the breadth and strength of available evidence. GenAI systems also differ in architecture, deployment arrangements, access to external tools, and operational permissions. A read-only assistant that summarizes reports presents different risks from an agent authorized to execute scripts or change network configurations. Future research should therefore distinguish among informational assistants, decision-support systems, and action-capable agents. A useful experimental design would compare conventional security tools alone, GenAI assistance without structured verification, and GenAI assistance with evidence-linked explanations and mandatory review for high-risk actions. Participants could evaluate simulated incidents containing correct and incorrect AI recommendations.

Outcomes could include decision accuracy, time to decision, false-positive and false-negative rates, appropriate rejection of incorrect recommendations, confidence calibration, and perceived accountability. Longitudinal studies could investigate whether sustained GenAI use improves professional expertise or gradually weakens independent verification. Organizational research could examine how responsibility is negotiated among security teams, managers, technology providers, and governance functions following AI-related incidents.

8. Conclusion

Generative AI is changing cybersecurity by influencing how professionals interpret evidence, frame problems, evaluate risk, and authorize responses. Its ability to summarize complex information and generate contextual recommendations can improve analytical efficiency and support more informed security decisions. However, these benefits depend on system reliability, calibrated trust, meaningful human oversight, and clear organizational accountability. This article has developed a conceptual framework connecting GenAI capabilities with human interpretation, trust

calibration, verification, decision quality, and governance. The framework identifies automation bias, workload pressure, insufficient evidence, and unclear responsibility as factors that may undermine the effectiveness of AI-supported security operations. The analysis indicates that responsible deployment requires more than accurate models or formal human approval. Organizations must provide evidence-linked explanations, appropriate authorization boundaries, least-privilege permissions, auditable decision records, realistic review requirements, and continuous evaluation of technical and human performance. The proposed research propositions offer a foundation for empirical studies investigating the effects of GenAI on cybersecurity decisions, analyst behavior, and organizational learning. Controlled experiments, longitudinal studies, and organizational case studies could establish when generative assistance improves outcomes and which safeguards are most effective. Ultimately, intelligent cybersecurity should not be judged by how many decisions AI can make without human involvement. It should be judged by whether the combined system produces accurate, defensible, secure, and accountable outcomes. GenAI can strengthen cybersecurity when it expands human analytical capacity while preserving the ability to question, verify, challenge, and govern its recommendations.

Acknowledgments: We are thankful to all authors for contributing to this research work.

Funding: Not Applicable

Author contribution statement

Abdul Mussawar: Conceptualization, methodology, investigation, data curation, formal analysis, and writing—original draft. Muhammad Younis Maitlo, Sabahat Ahmad Khan, Amna Munawar: Methodology, validation, and writing—review and editing. Mansoor Ali: Investigation, data analysis, and visualization. Qalab Hassnain Agha: Supervision, resources, and critical revision of the manuscript. Akhter Hussain: Validation, project administration, and writing—review and

editing. All authors read and approved the final version of the manuscript.

Data Availability: The data will be available from the corresponding author upon reasonable request.

Competing interests: The authors declare no competing interests.

REFERENCES

- Amershi, S., Weld, D., Vorvoreanu, M., Fournay, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Article 3, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2021). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9(1), 2–11. <https://doi.org/10.1609/hcomp.v9i1.18911>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). *On the opportunities and risks of foundation models*. Stanford Center for Research on Foundation Models. <https://arxiv.org/abs/2108.07258>
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. <https://arxiv.org/abs/1702.08608>
- ENISA. (2023). *Cybersecurity of AI and standardisation*. European Union Agency for Cybersecurity. <https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- International Organization for Standardization. (2023a). *ISO/IEC 42001:2023: Information technology–Artificial intelligence–Management system*. ISO.
- International Organization for Standardization. (2023b). *ISO/IEC 23894:2023: Information technology–Artificial intelligence–Guidance on risk management*. ISO.
- Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10, 18–26. <https://doi.org/10.1007/s13162-020-00161-0>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>

- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-14. <https://doi.org/10.1145/3313831.3376219>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392
- MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. *Journal of Marketing*, 75(4), 136-154. <https://doi.org/10.1509/jmkg.75.4.136>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229. <https://doi.org/10.1145/3287560.3287596>
- OWASP Foundation. (2025). OWASP Top 10 for large language model applications. OWASP. <https://genai.owasp.org/>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253. <https://doi.org/10.1518/001872097778543886>
- Pascoe, C., Quinn, S., & Scarfone, K. (2024). *The NIST Cybersecurity Framework (CSF) 2.0* (NIST CSWP 29). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.CSWP.29>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33-44. <https://doi.org/10.1145/3351095.3372873>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. <https://doi.org/10.1145/2939672.2939778>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503-2511.
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68. <https://doi.org/10.1145/3287560.3287598>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495-504. <https://doi.org/10.1080/10447318.2020.1741118>

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>

Tilbury, J., & Flowerday, S. (2024). Automation bias and complacency in security operation centers. *Computers*, 13(7), Article 165. <https://doi.org/10.3390/computers13070165>

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review*, 4(3), 356–367. <https://doi.org/10.1177/1534484305278283>

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, A., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>.