



ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL

Available Online: <https://assajournal.com>

Vol. 05 No. 01. Jan-March 2026. Page# 6848-6875

Print ISSN: [3006-2497](#) Online ISSN: [3006-2500](#)Platform & Workflow by: [Open Journal Systems](#)<https://doi.org/10.5281/zenodo.22509346>

Beyond the Benchmark: A Sociotechnical Perspective on Governance, Risk, and Compliance for Trustworthy AI in Cybersecurity Operations

Abdul Rafay

Department of Computer Science, Air University, Islamabad.

2502093@students.au.edu.pk

<https://orcid.org/0009-0000-2341-5645>

Anique Ahmad

Department of Electrical & Computer Engineering, Air University, Islamabad.

anique.ahmad@students.au.edu.pk

Muhammad Saad

Department of Computer Science, Air University, Islamabad.

231682@students.au.edu.pk

Abdullah Saleem

Department of Computer Science, Air University, Islamabad.

2501344@students.au.edu.pk

ABSTRACT

Benchmark accuracy is the most commonly reported evidence for artificial intelligence (AI) systems used in cybersecurity operations, yet accuracy on a fixed test set does not, by itself, establish that a system is safe, governable, or fit for deployment. This paper offers a conceptual, narrative synthesis of the sociotechnical conditions that benchmark-centric evaluation leaves unaddressed: adversarial robustness under realistic threat models, meaningful human oversight, data protection and fairness, and organizational governance, risk, and compliance (GRC). Drawing on a purposively selected set of primary technical standards, regulatory instruments, and peer-reviewed literature and explicitly distinguishing empirical findings, requirements imposed by normative instruments, and the authors' own proposals the paper develops three conceptual artifacts: (1) a four-layer sociotechnical evidence profile (technical, human, data-and-rights, organizational); (2) a lifecycle crosswalk relating that evidence profile to the NIST AI Risk Management Framework, MITRE ATLAS, ISO/IEC 42001, and the EU AI Act; and (3) a five-level maturity model for organizational AI GRC capability, illustrated with a worked application to a hypothetical AI-enabled security-operations-center (SOC) system. That worked application is a hypothetical, formative illustration of how the maturity model would be used, not an empirical validation of it, and its provisional finding restricted, analyst-reviewed decision support rather than expanded autonomy is reported accordingly. Each artifact is presented as an unvalidated design proposal rather than a measurement instrument, and each normative claim in the paper is scoped to stated assumptions about consequence, autonomy, and reversibility rather than asserted as a universal criterion of trustworthiness. The paper closes with an explicit agenda for empirical validation, including inter-rater reliability testing, worked case studies, and comparison against incident and audit outcomes.

Editorial positioning. *This manuscript should be evaluated as a conceptual framework or perspective article informed by a purposive narrative synthesis. It does not claim systematic-review completeness, legal advice, or validation of the proposed assessment artifacts.*

Keywords: Trustworthy AI; Cybersecurity; Governance, Risk, Compliance; Maturity Model; Human Oversight.

1. Introduction

Organizations increasingly deploy machine learning and, more recently, large language model (LLM)–based and agentic systems inside the security operations center (SOC): for intrusion and anomaly detection, malware and phishing classification, alert triage and prioritization, natural-language summarization of incidents, and semi-autonomous response actions. The evidence typically offered to justify these deployments is benchmark performance precision, recall, F1, or area-under-curve figures reported on a held-out or public dataset. Benchmark performance is necessary evidence, but it answers a narrower question than the one procurement, audit, and regulatory processes actually need answered: is this specific system, in this specific deployment context, governable, contestable, and safe enough for the consequences it can produce?

This paper is a conceptual framework informed by a purposive narrative synthesis, not a systematic or scoping review and not a meta-analysis. It does not claim exhaustive or reproducible retrieval of every relevant study; it claims a defensible, purposive account of the literatures that bear on the question above, combined with a transparent statement of the paper’s own evidentiary limits (Section 3). The paper’s contribution is not a new empirical finding but an organizing structure: a way of separating what benchmark evaluation shows from what deployment-level trust claims additionally require, and a way of relating that separation to existing standards and regulation rather than inventing a parallel compliance regime.

1.1 Scope and definitions

“Trustworthy AI” is used in the literature to mean different things reliability, safety, fairness, transparency, accountability, or simple user trust (Jobin, Ienca, & Vayena, 2019). This paper treats trustworthiness as a multidimensional, context-dependent property rather than a single pass/fail label, and it deliberately distinguishes four related but non-identical constructs used throughout the paper:

Construct	What it means here	What it does not mean
Trustworthy	The system’s technical, human-factors, data, and governance evidence is proportionate to the consequence and reversibility of its errors in a stated deployment context.	That the system scored well on a benchmark, or that any single dimension (e.g., accuracy) is high.
Lawfully deployable	The system satisfies the specific legal obligations that apply to its classification and jurisdiction (e.g., EU AI Act risk tier).	That the system is safe, effective, or trustworthy in a broader sense; legal compliance is necessary but not sufficient.
Assurance-ready	The organization can produce documented, verifiable evidence for each claim it makes about the system on request (audit, procurement, regulator).	That the evidence produced is itself sufficient or validated; readiness is about documentation, not adequacy.
High maturity	The organization has repeatable, institutionalized processes for producing and updating that evidence over the system’s lifecycle.	That any individual system deployed by a mature organization is, by that fact alone, trustworthy.

Table 1. Working definitions used in this paper. The four constructs are related but not interchangeable; conflating them is identified in Section 11 as a recurring problem in practitioner and vendor discourse.

“AI-enabled cybersecurity” in this paper covers three overlapping system classes with materially different risk profiles: (a) detection and classification systems (e.g., network intrusion detection, malware classification) that flag but do not act; (b) decision-support and triage systems (e.g., alert prioritization, LLM-generated incident summaries) that shape but do not finalize analyst decisions; and (c) autonomous or semi-autonomous response systems (e.g., agentic SOC tools that can isolate hosts, revoke credentials, or push configuration changes) that act on the environment with limited or delayed human confirmation. Claims made about class (a) systems in this paper should not be assumed to transfer to class (c) systems without re-evaluation; this distinction is used repeatedly in Sections 4–10.

1.2 Contribution and paper structure

The paper makes three contributions, each explicitly labeled as a conceptual proposal rather than a validated instrument: (1) a four-layer sociotechnical evidence profile (Section 8); (2) a crosswalk linking that profile to existing standards and regulation rather than replacing them (Section 9); and (3) a five-level organizational maturity model, illustrated with a worked application (Section 10). Two supporting artifacts make these proposals traceable rather than aspirational: a failure-pathway crosswalk decomposing technical robustness by attack type (Section 4.1) and an assurance-chain template linking any single deployment claim to its evidence artifact, test method, assessor, decision rule, residual uncertainty, and monitoring trigger (Section 8.1). Section 2 situates the paper against related work. Section 3 states the literature synthesis method and its limits. Sections 4–7 synthesize evidence on technical robustness, human oversight and explainability, privacy and fairness, and the GRC and regulatory landscape, respectively, each separating empirical findings from normative requirements and from the authors’ own inferences. Section 11 discusses limitations and scope conditions. Section 12 concludes with a validation agenda.

2. Related Work

Three literatures inform this paper without being fully integrated by any single existing source. First, technical AI security research documents adversarial vulnerabilities in machine learning systems, including evasion, poisoning, model extraction, and membership inference (Biggio & Roli, 2018; Goodfellow, Shlens, & Szegedy, 2015). NIST’s adversarial machine learning taxonomy (Vassilev, Oprea, Fordyce, & Anderson, 2024) consolidates this literature into a common vocabulary of attack and defense categories. Second, human-factors and explainable-AI (XAI) research addresses whether people can understand, calibrate trust in, and productively override model outputs (Doshi-Velez & Kim, 2017; Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016), with a security-specific concern about alert fatigue and automation bias in SOC contexts. Third, AI governance research and a fast-growing body of regulatory and standards instruments the EU Artificial Intelligence Act, the NIST AI Risk Management Framework (NIST AI 100-1, 2023), ISO/IEC 42001:2023, and MITRE ATLAS address organizational accountability, risk management, and threat modeling for AI systems specifically.

Prior surveys of AI-ethics principles (Jobin, Ienca, & Vayena, 2019) show substantial convergence on high-level values such as transparency, fairness, and accountability, but persistent divergence on operational definitions and measurement. Prior cybersecurity-specific reviews tend to survey either detection techniques or governance frameworks, rarely both, and rarely make explicit which of their claims are backed by controlled empirical evaluation versus by the text of a standard versus by the reviewing authors’ own judgment. That gap not a lack of any single missing framework is the specific space this paper occupies.

2.1 Positioning relative to existing surveys

- (P) The paper makes a bounded organizational contribution rather than an empirical one. It does not claim to be the first integration of these literatures; its narrower claim is that the selected sources support a useful, explicitly provisional organization of evidence. Existing technical surveys of machine learning in cybersecurity, such as Sarker, Kayes, Badsha, Alqahtani, Watters, and Ng's (2020) overview of cybersecurity data science, catalog techniques, data types, and application areas across intrusion detection, malware analysis, and related tasks, but are not organized around AI-specific governance instruments such as the EU AI Act or ISO/IEC 42001, which postdate much of that literature. Conversely, Apruzzese et al.'s (2023) discussion of the operational role of machine learning in cybersecurity practice addresses adoption barriers and practitioner concerns without building a crosswalk to formal GRC instruments. This paper's contribution is to connect technical/human evaluation evidence (Sections 4–6) with AI-specific governance instruments (Section 7) inside one evidence-organizing structure (Sections 8–10); it does not report new empirical findings in either the technical or the governance literature. We have not conducted a systematic search sufficient to claim that no other source performs this integration, and we state the narrower, defensible version of the claim: within the purposively selected sources reviewed for this paper (Section 3, Appendix A), we did not find an existing source that performs it.

3. Method and Evidentiary Scope

This paper is a purposive, narrative synthesis. It does not follow a PRISMA-style systematic review protocol, was not pre-registered, and does not report a reproducible database query log with the precision that a systematic review would require. This section states plainly what was done and what was not, so that readers can weigh the paper's claims accordingly.

3.1 What was done

- Sources were selected purposively to cover four domains identified in Section 1.1: adversarial robustness, human oversight/explainability, privacy and fairness, and AI governance/regulation, with deliberate inclusion of primary normative instruments (the EU AI Act text, NIST publications, ISO/IEC 42001, MITRE ATLAS) alongside peer-reviewed technical and human-factors literature.
- Primary legal and technical instruments are cited to their official or maintaining-body source (e.g., the EU AI Act Service Desk, NIST Computer Security Resource Center and National Cybersecurity Center of Excellence, the MITRE ATLAS knowledge base) rather than to secondary commercial explainers, in direct response to the source-quality concern raised about an earlier draft of this work. Appendix A records, for each such instrument, its title, identifier, edition/version, retrieval date, and the specific article or section relied upon.
- Where a claim depends on a fast-changing instrument regulatory deadlines, the count of MITRE ATLAS tactics/techniques, or standards version numbers this paper records the specific version and retrieval context relied upon (Table 5; Appendix A) and flags where the instrument was, at the time of writing, still in draft, pending public comment, or subject to a pending legislative amendment.

3.2 What was not done, and why it matters

- No systematic database search with logged query strings, screening counts, dual-reviewer screening, or a PRISMA flow diagram was performed. Accordingly, this manuscript makes no claim of comprehensive retrieval, prevalence estimation, or systematic-effect synthesis. Any reader who requires that level of rigor should treat this

paper as a scoping and framework-proposal document, not as a systematic review, and should independently verify coverage before relying on it for a completeness claim.

- No formal quality-appraisal tool (e.g., a risk-of-bias instrument) was applied uniformly across sources, because the source set deliberately mixes empirical studies, standards, and regulatory text, which are not commensurable under a single appraisal scheme. Instead, Section 3.3 assigns each major claim an explicit evidence type.
- The four-layer evidence profile, the crosswalk, and the maturity model introduced later in the paper are original conceptual syntheses. They have not been empirically validated, piloted with independent assessors, or tested for inter-rater reliability. The worked application in Section 10.1 is a formative illustration of how the instruments would be used, not a validation study. They are presented as design proposals for future validation, not as measurement instruments ready for operational use (Section 12).

3.3 Evidence-typing convention used throughout the paper

Major claims in Sections 4–10 are typed using the scheme below. Where a sentence combines empirical evidence, a normative requirement, and an author proposal, those components are separated or labeled individually rather than relying on a paragraph-level label. Where a sentence does not carry an inline qualifier, the surrounding paragraph states the type once for the group of claims that follow.

Type	What it means	How it is worded in this paper
Empirical (E)	Supported by a specific peer-reviewed study, benchmark, or reproducible measurement.	“Studies of X report...” / “X has been shown to...”, with citation
Normative (N)	Stated as a requirement or recommendation by a named standard, law, or framework.	“The EU AI Act requires...” / “ISO/IEC 42001 specifies...”
Author proposal (P)	An inference, synthesis, or design choice made by this paper’s authors, not independently validated.	“We propose...” / “This paper suggests...” / explicit “conceptual” or “unvalidated” labeling

Table 2. Evidence-typing convention. This replaces the earlier draft’s undifferentiated narrative synthesis, in which empirical findings, standards language, and author recommendations were not consistently distinguished.

3.4 Source-and-claim evidence table

To make the narrative synthesis auditable, Table 3 lists a representative (not exhaustive) sample of the major claims made in Sections 4–6, the specific source behind each, its evidence type (Table 2), and how directly it bears on cybersecurity operations specifically, as opposed to AI systems in general. Claims not listed in Table 3 but appearing later in the paper carry an inline (E)/(N)/(P) marker and a citation or explicit “we propose” framing, following the same convention.

Source	Source type	System / population	Claim supported	Evidence type	Directness	Key limitation
Sommer & Paxson (2010)	Peer-reviewed (IEEE S&P)	Network intrusion detection	Benchmark-tuned ML-based IDS can perform poorly against realistic,	E	Direct — cybersecurity	Predates many current architectures; some diagnosed

			non-stationary attack traffic.			problems partially addressed since.
Arp et al. (2022)	Peer-rev. (USENIX Sec.)	ML-based security systems (malware/IDS)	Common experimental pitfalls (e.g., spurious correlations, unrealistic splits) inflate reported ML security performance.	E	Direct — cybersecurity	Focused on experimental-design flaws, not a robustness benchmark itself.
Biggio & Roli (2018)	Peer-rev. survey	General ML systems	Benchmark accuracy does not, by itself, establish robustness to adversarial manipulation.	E	Indirect — general ML	Survey synthesis, not a single new experiment.
Sundaramurthy et al. (2016)	Peer-rev. (SOUPS)	SOC analysts	Documents analyst workload strain and burnout relevant to human-oversight capacity.	E	Direct — cybersecurity	Single-organization qualitative case study.
Kokulu et al. (2019)	Peer-rev. (ACM CCS)	SOC organizations	Identifies mismatches between SOC design assumptions and analyst practice.	E	Direct — cybersecurity	Small sample of SOCs; qualitative.
Alahmadi, Axon, & Martinovic (2022)	Peer-rev. (USENIX Sec.)	SOC analysts	Documents a very high false-positive burden and alert fatigue in a real SOC environment.	E	Direct — cybersecurity	Single-organization study; generalizability untested.
Greshake et al. (2023)	Peer-rev. workshop paper	LLM-integrated applications	Demonstrates indirect prompt injection as a practical attack vector, supporting Section 4's architecture-dependent framing.	E	Direct — LLM security	Fast-moving area; specific exploits/mitigations may be superseded.

Vassilev, Oprea, Fordyce, & Anderson (2024)	Normative taxonomy (NIST)	General AI/ML systems	Establishes the shared taxonomy of adversarial ML attack/mitigation categories used in Section 4.	N	Indirect — general AI	Taxonomy document, not an empirical benchmark or control standard.
National Institute of Standards and Technology (2025)	Normative, preliminary draft (NIST)	AI systems in cybersecurity programs	Proposes a CSF 2.0-aligned structure (Secure/Defend/Thwart) for organizing AI-related cybersecurity risk.	N	Direct — cybersecurity, but not yet final	Initial preliminary draft only; open for comment as of the reviewed version, subject to material revision.
This paper's evidence profile, crosswalk, maturity model, and worked application (Sec. 8–10)	Author proposal	N/A	Organizes the above evidence into a lifecycle assessment structure and illustrates its use on one hypothetical system.	P	N/A	Unvalidated; no inter-rater or outcome data exists yet (Section 12).

Table 3. Representative source-and-claim evidence table. “Directness” flags whether a source studies cybersecurity operations specifically or general AI/ML systems from which this paper extrapolates; several claims in Sections 4–6 rely on the latter and are marked accordingly in the text.

4. Technical Reliability and Adversarial Robustness

- (E) Benchmark accuracy on a static test set does not, by itself, establish robustness to distribution shift, novel attack techniques, or adversarially crafted inputs; this gap is documented across the adversarial machine learning literature generally (Biggio & Roli, 2018; Goodfellow, Shlens, & Szegedy, 2015) and specifically for network intrusion detection, where systems that scored well on benchmark datasets have been shown to degrade sharply against non-stationary, realistic attack traffic (Sommer & Paxson, 2010). A methodological review of published ML-security evaluations further finds that common experimental choices unrealistic train/test splits, spurious correlations between features and labels, and evaluation on non-representative data can inflate reported performance figures for malware and intrusion-detection classifiers specifically (Arp et al., 2022). We treat this as a documented risk that reported benchmark numbers can overstate real-world performance, not as a claim that benchmarks are uninformative; a well-constructed benchmark result is still evidence, just insufficient evidence on its own. “Robustness” is not one property: evasion (crafted inputs that cause misclassification at inference time), poisoning (corrupted training or fine-tuning data), model extraction and membership inference (attacks on model confidentiality and training-data privacy), and

generalization failure under distribution shift are distinct failure modes with different detection and mitigation strategies (Vassilev, Oprea, Fordyce, & Anderson, 2024).

4.1 Failure-pathway crosswalk

(P) Because these pathways are not interchangeable “robustness” tests, we separate them further than a single paragraph allows. Tables 4.1a-b decompose five failure pathways evasion, poisoning, extraction or confidentiality loss, privacy leakage, and distribution shift by attacker access, the object manipulated, the property affected, an illustrative detection signal, a representative mitigation, and the residual uncertainty that remains even where the mitigation is applied. This crosswalk organizes and extends, rather than replaces, the NIST adversarial machine learning taxonomy (Vassilev, Oprea, Fordyce, & Anderson, 2024)

and the sources cited above; it is our own organizing synthesis, not a claim that any cited source presents the pathways in exactly this tabular form.

Pathway	Attacker assumed access	Manipulated / targeted object	Affected property
Evasion	Query (black-box) or white-box access at inference time	Input features / traffic at inference time	Decision boundary / classification output
Poisoning	Access to training, fine-tuning, or retrieval-index content	Training data or retrieved knowledge-base content	Model parameters or retrieval outputs (integrity of learned behavior)
Extraction / confidentiality loss	Repeated query access to a deployed model or API	The model itself, or training data reachable via the model	Model or training-data confidentiality; intellectual property
Privacy leakage (membership inference / reconstruction)	Query access, sometimes with partial knowledge of candidate records	Individual training-data records	Data-subject privacy
Distribution shift / generalization failure	None required (also exploitable via gradual environment change)	The operating environment / input distribution over time	Predictive validity of the deployed model

Table 4.1a. Failure-pathway crosswalk, part 1: attacker access, target, and affected property.

Pathway	Illustrative detection signal	Representative mitigation	Residual uncertainty
Evasion	Anomalous input statistics; confidence/calibration drop; adversarial-example detectors	Adversarial training; input sanitization; randomized smoothing; threshold hardening	Adaptive attackers can target the specific defense deployed; efficacy is defense- and threat-model-specific (Biggio & Roli, 2018)
Poisoning	Label/feature distribution drift; retrieval-content	Data-provenance controls; anomaly	Low-rate, subtle poisoning can evade

	provenance anomalies; held-out canary performance	filtering; periodic retraining audits	statistical detection; depends on baseline stability
Extraction / confidentiality loss	Unusual query volume/pattern; anomalies	Query rate limiting; output perturbation; watermarking; access controls	Detection depends on attacker patience and ability to mimic legitimate traffic
Privacy leakage (membership inference / reconstruction)	Not typically observable in production; requires proactive red-team testing (Section 6)	Differential privacy; output aggregation; minimization of memorizable telemetry	Absence of a detected leak is not evidence of absence; requires proactive testing rather than passive monitoring
Distribution shift / generalization failure	Performance/calibration drift on labeled samples; rising abstention/override rates	Continuous monitoring; scheduled re-evaluation; retraining triggers	Detection lags the shift itself; the lag size is not well characterized for SOC-specific systems (Sommer & Paxson, 2010)

Table 4.1b. Failure-pathway crosswalk, part 2: detection, mitigation, and residual uncertainty. Tables 4.1a-b together form a single crosswalk (split across two tables for legibility); each pathway requires a separate threat model, and a system tested against one pathway should not be assumed robust against another. Cell content is illustrative rather than exhaustive and should be adapted to the specific system's architecture and trust boundaries.

- (F) MITRE ATLAS provides a structured, ATT&CK-style taxonomy of adversary tactics and techniques specific to AI and ML systems, including two tactics specific to AI systems AI/ML Model Access and AI/ML Attack Staging layered onto tactics inherited from MITRE ATT&CK, and it has been extended with agentic-AI-specific techniques as agentic tooling has become more common in SOC contexts. Because the atlas.mitre.org landing page is a JavaScript application that does not itself expose a stable, citable count in its rendered text, this paper anchors its count to the versioned, machine-readable data export MITRE publishes alongside it (the mitre-atlas/atlas-data repository) rather than to the live site, so that the figure is reproducible against a dated snapshot rather than a page that can change without a version marker. In the dated ATLAS snapshot used for this manuscript, the export indexed 16 tactics and 101 techniques (MITRE, 2026); an earlier release, version 5.1.0 (November 2025), indexed the same 16 tactics but 84 techniques (up from 66 in a release the previous month). We report the snapshots only as dated examples, not as current or enduring totals: the tactic count (16) has been stable across the releases we checked, while the technique count has grown release-to-release and also varies with counting convention (e.g., whether sub-techniques are counted separately), which is itself the reason this paper treats any single technique count as version-specific evidence rather than a fixed fact. Readers should query the dated data export directly, rather than cite a number from this paper's text, before relying on a specific technique count

elsewhere; this paper does not claim to reproduce MITRE's own tactic/technique labels beyond the two AI-specific tactics named above.

(E/P) For LLM-based and agentic SOC tooling specifically, empirical adversarial evaluation is markedly thinner than for classical classifiers: prompt injection, tool-use manipulation, and unsafe autonomous action chains are active research areas rather than settled ones. We treat the claim that threat modeling has "outpaced" practice in this sub-area as our own (P) assessment based on the comparative maturity of the classical-ML adversarial literature versus the agentic-AI literature, not as an established empirical finding, and we invite the reader to weigh it accordingly.

(E/P) Whether a given cybersecurity AI system exposes an indirect prompt-injection surface is not a property of the model "by construction"; empirical demonstrations of indirect prompt injection against real LLM-integrated applications show that exposure instead depends on the system's ingestion pipeline, what untrusted content reaches the model context, what tools or actions the model can invoke, and how instruction hierarchy and permissioning are implemented (Greshake et al., 2023). A claim that a specific deployed system is or is not exposed to this class of attack should be verified against its actual architecture and trust boundaries rather than inferred from the fact that it uses an LLM.

5. Explainability and Human Oversight

(E) A human being formally assigned to review AI output is not the same as that human exercising meaningful oversight. Field studies of security operations centers document substantial analyst workload strain and burnout under sustained alert volume (Sundaramurthy et al., 2016), structural mismatches between how SOC's are designed and how analysts actually work (Kokulu et al., 2019), and, in at least one studied organization, an extremely high proportion of alerts judged by analysts to be false positives, with corresponding fatigue and desensitization effects (Alahmadi, Axon, & Martinovic, 2022). These are cybersecurity-specific empirical findings, not general human-factors inference; we flag where a claim in this section instead extrapolates from general automation-bias research rather than from a SOC-specific study. Post-hoc explanation methods such as LIME (Ribeiro, Singh, & Guestrin, 2016) and SHAP (Lundberg & Lee, 2017) can support review, but explanation fidelity, analyst comprehension, and actual decision-quality improvement are separate empirical questions; an explanation being locally faithful to a model does not, by itself, establish that it improves an analyst's decision (Doshi-Velez & Kim, 2017), and we did not find a published study directly measuring this link in a SOC setting specifically.

Dimension	Definition	Measurable indicator
Faithfulness	Whether an explanation reflects the model's actual decision process.	Agreement between explanation and controlled perturbation/ablation tests.
Analyst comprehension	Whether the analyst correctly understands what the explanation is telling them.	Comprehension quiz or think-aloud protocol scored against ground truth.
Decision-quality impact	Whether access to the explanation improves the analyst's final decision.	A/B comparison of analyst accuracy/time with vs. without the explanation.
Actionability	Whether the explanation supports a concrete next step	Proportion of explanations linked to a documented analyst action.

	(escalate, dismiss, investigate further).	
Calibration support	Whether the explanation helps the analyst judge how much to trust the specific output.	Correlation between analyst-stated confidence and model reliability, pre/post explanation.

Table 4. Explainability dimensions with operational indicators. Each dimension requires a distinct measurement approach; none is established by the presence of a post-hoc explanation method alone.

5.1 Human oversight as a measurable performance relationship

(P) A human being formally assigned to review AI output is not, by itself, evidence of meaningful oversight; the relevant question is whether the human–AI arrangement improves or degrades decision quality under realistic workload and time pressure. Tables 5.1a-b propose five outcome-level metrics for that relationship, extending Table 4’s focus on properties of the explanation to properties of the oversight relationship as a whole. These are our own proposed metrics, not an established measurement standard; we did not find a published, validated instrument that operationalizes all five for a SOC setting specifically, and we flag this as part of the validation agenda in Section 12.

Outcome	Definition	Why it matters
Missed-threat rate	Proportion of true threats not escalated or acted on under the human-AI arrangement.	Oversight can silently degrade detection if a human defers uncritically or attention is drawn away from correct positives.
Unnecessary-escalation rate	Proportion of system-triggered escalations a qualified reviewer judges unnecessary.	Drives workload burden and alert-fatigue risk (Section 5).
Time to decision	Elapsed time between an alert or recommendation and the analyst’s decision, with and without the AI system.	Tests whether the tool actually accelerates, or instead burdens, triage under real workload.
Override quality	Whether analyst overrides of system recommendations are, on independent review, appropriate.	A raw override rate does not indicate whether the overrides themselves are correct.
Detection of confidently incorrect recommendations	Proportion of confidently worded but incorrect system outputs that analysts correctly identify as incorrect.	Fluent, confident phrasing can mask errors (automation bias).

Table 5.1a. Proposed human-oversight outcome metrics: definitions and rationale.

Outcome	Measurement approach
Missed-threat rate	Retrospective audit against an independently labeled sample.
Unnecessary-escalation rate	Independent secondary review of a sample of escalations.
Time to decision	Instrumented workflow timestamps, compared pre/post deployment or via A/B testing.
Override quality	Blinded second-reviewer adjudication of a sample of overrides.

Detection of confidently incorrect recommendations	Seeded or injected incorrect recommendations evaluated in a controlled setting.
--	---

Table 5.1b. Proposed human-oversight outcome metrics: measurement approach. Tables 5.1a-b together operationalize the distinction, used throughout Section 5, between a human being assigned to review output and a human exercising oversight that measurably changes decision quality. These are our own proposed metrics, not an established measurement standard.

- (P) We propose as our own design recommendation, not as a conclusion drawn from comparative empirical evidence that a cybersecurity AI evaluation distinguish, at minimum, detection systems (which flag; a human decides) from systems that make or execute a consequential decision about a person or an asset (e.g., automatically suspending a user account). We suggest that the latter category warrants a higher evidentiary bar for oversight, on the reasoning that the cost of an unreviewed error differs in kind, not only in degree; we did not find an empirical study that quantifies the appropriate oversight threshold as a function of consequence, and this paper does not supply one. Readers should treat this recommendation as a hypothesis for the case-study work proposed in Section 11.2, not as an established requirement.

6. Privacy, Fairness, and Security

- (E) Security telemetry and detection systems process data that can be sensitive even when it is not classically “personal data” in the narrow sense (e.g., behavioral baselines, authentication patterns, communication metadata), and model training or fine-tuning on such data raises membership-inference and data-retention questions studied in the privacy-preserving machine learning literature. (N) Where a cybersecurity AI system does process personal data or produces effects on identifiable individuals (for example, flagging a specific employee as an insider-threat risk), data protection law and organizational data-governance obligations apply in the ordinary way; using the system for a security purpose does not create an exemption.
- (F) “Fairness” in the algorithmic-fairness literature refers to a family of formally distinct statistical criteria (e.g., demographic parity, equalized odds) that can conflict with one another; there is no single metric that satisfies all of them simultaneously in general. (P) Whether any particular fairness criterion is the relevant one for a cybersecurity detection system depends on whether the system’s errors are distributed across protected groups in a way that produces a cognizable harm (e.g., systematically higher false-positive rates flagging certain employee groups as threats). We caution against assuming that fairness metrics developed for lending or hiring decisions transfer without adaptation to detection-oriented security tooling, and we did not find, in the sources reviewed, a validated fairness benchmark specific to SOC detection systems.

7. Governance, Risk, Compliance, and the Regulatory Landscape

This section separates enacted law, in-force standards, and pending or draft instruments, because conflating them was identified as a specific risk in an earlier draft of this work. Table 5 records the version and status of each instrument, its legal or normative force, and the retrieval context relied upon; it should be re-verified against the authoritative source before any compliance decision is based on it.

Instrument	Type	Status as of writing	Key caveat
EU AI Act (Regulation (EU) 2024/1689), as amended by the	Binding law (EU)	In force since 1 Aug 2024; prohibitions and AI-literacy duties applicable since 2 Feb 2025; GPAI	This paper’s account of the amendment reflects the enacted Regulation as of this paper’s retrieval

Digital Omnibus on AI (Regulation (EU) 2026/1744)		duties since 2 Aug 2025. The amending Regulation (EU) 2026/1744 entered into force 27 Jul 2026 and postpones the Annex III high-risk obligations (originally due 2 Aug 2026) to 2 Dec 2027, and the Annex I product-safety-component high-risk obligations to 2 Aug 2028.	context; verify against the Official Journal text for any article-level detail (e.g., transparency/Art. 50 timing) before relying on it.
NIST AI RMF 1.0 (NIST AI 100-1)	Voluntary framework (US)	Published 2023; in active use as a reference framework.	Voluntary; adoption does not itself satisfy any specific binding legal obligation.
NIST Cybersecurity Framework 2.0	Voluntary framework (US)	Published 2024.	Not AI-specific; requires interpretation to map to AI-specific risks.
NIST Cyber AI Profile (NIST IR 8596)	Voluntary, initial preliminary draft (US)	Released 16 Dec 2025 for public comment; comment period closed 30 Jan 2026; not yet a final or initial public draft as of this paper's retrieval context.	Not sector-specific guidance and not yet a stable publication; structure, focus areas, and CSF-function mapping may change materially before finalization.
ISO/IEC 42001:2023	Certifiable management-system standard	Published 2023; certification schemes active.	Certifies management-system conformance, not system-level safety or security assurance.
MITRE ATLAS	Public threat taxonomy / knowledge base	Actively versioned; data-export version 2026.05 (modified 27 May 2026) indexed 16 tactics and 101 techniques, up from 16 tactics and 84 techniques in version 5.1.0 (Nov. 2025), with agentic-AI techniques added on recent releases.	Tactic count has been stable; technique counts change between releases and with counting convention (e.g., sub-techniques). Cite the specific dated data-export version relied upon, not the live site.

Table 5. Status of major GRC-relevant instruments referenced in this paper, as of the stated retrieval context (sources retrieved and checked during 2025–2026; see Appendix A). The “Type” column distinguishes binding law, voluntary framework, certifiable standard, and preliminary draft so that legal force and maturity are not implied where none exists. Regulatory and standards timelines are subject to amendment or revision; see the discussion of the EU AI Act and the Cyber AI Profile below.

7.1 The EU AI Act

- (N) The EU AI Act entered into force on 1 August 2024 and applies on a staggered timeline. Prohibited practices and AI-literacy obligations became applicable on 2 February 2025, and obligations for general-purpose AI (GPAI) models became applicable on 2 August 2025 for providers of GPAI models, among other GPAI-specific provisions; both dates are unaffected by the amendment discussed below. Under the originally enacted text, the substantive requirements for high-risk AI systems listed in Annex III were scheduled to become applicable on 2 August 2026.
- (O) That original high-risk timeline has since been amended. The European Commission proposed a “Digital Omnibus” package targeting the AI Act (COM (2025)0836) on 19 November 2025; the European Parliament and the Council reached a provisional trilogue agreement on 7 May 2026; Parliament adopted the text on 16 June 2026; the Council formally adopted it on 29 June 2026; and the amending instrument Regulation (EU) 2026/1744 was published in the Official Journal and entered into force on 27 July 2026. As enacted, Regulation (EU) 2026/1744 postpones the Annex III high-risk obligations from 2 August 2026 to 2 December 2027, and postpones the high-risk obligations for AI systems that are safety components of products regulated under Annex I (e.g., medical devices, machinery) from 2 August 2027 to 2 August 2028; it leaves the AI-literacy, prohibited-practice, and GPAI dates above unchanged. This paper cites Regulation (EU) 2026/1744 directly for these postponed dates rather than treating them as a pending proposal, correcting an earlier draft of this paper (and of the audit that reviewed it) that could report only the proposal’s existence because, at the time each was written, the amendment had not yet been adopted. We nonetheless flag two things a reader should still verify independently: first, article-level details of the amendment beyond the two headline dates above (for example, the exact grace period for transparency obligations under Article 50, which secondary reporting describes inconsistently) should be checked against the Official Journal text itself rather than against this paper; second, because this is an area that has changed materially within the drafting of this paper, readers should confirm against the European Commission’s AI Act Service Desk that no further amendment has been adopted since this paper’s retrieval context.
- (P) Cybersecurity relevance under the AI Act is not a blanket category: whether a given AI-enabled security tool is “high-risk” depends on its specific use case and classification under Annex III (and, for safety components of regulated products, Annex I), not on the fact that it is used “for cybersecurity” in general. A detection tool used purely for internal network monitoring and a biometric-categorization or critical-infrastructure-safety component used for security purposes can fall under materially different obligations, and classification depends on the system’s purpose, role, and the applicable Annex provision rather than on its technology category alone. (P) We recommend that organizations perform and document a case-specific classification exercise (as Article 6 contemplates) rather than applying a single blanket determination across all “AI cybersecurity” tooling; this recommendation is an illustrative application of the Regulation’s structure, not a legal conclusion, and it does not substitute for legal advice.

7.2 NIST AI Risk Management Framework, Cybersecurity Framework, and the Cyber AI Profile

- (N) The NIST AI Risk Management Framework (AI RMF 1.0) organizes AI risk management around four functions Govern, Map, Measure, and Manage and is voluntary rather than mandatory in the United States, though it is frequently referenced by procurement and audit processes as a common vocabulary (National Institute of Standards and Technology, 2023).

- (O) On 16 December 2025, NIST released an initial preliminary draft Cyber AI Profile (NIST IR 8596, Cybersecurity Framework Profile for Artificial Intelligence), organized around three AI-specific focus areas Secure, Defend, and Thwart layered onto the six functions of the NIST Cybersecurity Framework (CSF) 2.0 (Govern, Identify, Protect, Detect, Respond, Recover) (National Institute of Standards and Technology, 2025). The preliminary draft was open for public comment through 30 January 2026, and NIST indicated that comments received would inform a subsequent initial public draft. As of this paper's retrieval context, we did not find evidence that a later, more final version had superseded the preliminary draft, but this should be re-verified before the document is characterized as settled guidance. We correct an earlier draft of this paper, which described this instrument as "published sector-specific profile guidance"; that description overstated its maturity. The accurate characterization is: NIST released an initial preliminary draft Cyber AI Profile, organized around CSF 2.0 outcomes, to solicit public comment, and it should not be treated as final, sector-specific in the sense of a completed profile, or independently authoritative for compliance purposes unless a later official version is cited.
- (P) Where this paper's crosswalk (Section 9) maps evidence-profile layers to CSF functions, that mapping is our synthesis of the two frameworks' stated scopes, not a claim that NIST has published a single authoritative one-to-one crosswalk covering every control, and it is not based on the Cyber AI Profile's still-evolving focus-area structure; readers implementing this mapping in a specific organization should adapt it to their own control set and should track the Cyber AI Profile's status separately.

7.3 ISO/IEC 42001 and MITRE ATLAS

- (N) ISO/IEC 42001:2023 specifies requirements for an AI management system (AIMS), following the same high-level management-system structure as ISO/IEC 27001. (P) Certification to ISO/IEC 42001 demonstrates that an organization operates a conforming AI management system; it does not, by itself, demonstrate that any specific AI system deployed under that management system is safe, secure, or free of the technical failure modes discussed in Section 4. Treating AIMS certification as equivalent to system-level security or safety assurance is a category error this paper explicitly avoids and cautions readers against. MITRE ATLAS (Section 4) is a threat taxonomy and knowledge base, not a certification or compliance scheme; mapping a system's exposure against ATLAS tactics and techniques is one useful input to a threat model an author recommendation, not a requirement imposed by MITRE, NIST, ISO, or the EU AI Act and we decline to characterize it as a "minimum" control in the absence of an organization-specific risk assessment establishing that it is the appropriate baseline for that organization's risk tolerance.

8. A Sociotechnical Evidence Profile (Conceptual Proposal)

- (P) This section and Sections 9–10 present this paper's own conceptual contribution. Nothing in these sections should be read as an empirically validated instrument, a certification scheme, or a substitute for legal or security advice. The evidence profile organizes deployment-readiness evidence into four layers, motivated by the observation (Sections 4–6) that technical, human, data/rights, and organizational evidence answer different questions and can fail independently of one another.

Layer	Illustrative evidence	Question it answers
Technical	Adversarial/robustness testing against a stated threat model (e.g., relevant MITRE ATLAS techniques); performance under	Does the system work, including under realistic

	distribution shift; calibration and abstention behavior.	stress, not only on a static benchmark?
Human	Analyst workflow integration, override logging, automation-bias mitigation, explanation comprehension testing.	Can a human meaningfully understand, contest, and override the system?
Data and rights	Data provenance and minimization, retention limits, membership-inference/privacy testing, applicable data-subject rights handling.	Is the data the system uses and produces handled lawfully and appropriately?
Organizational/GRC	Ownership and accountability records, incident and drift monitoring, vendor responsibility terms, lifecycle review cadence.	Does the organization have a durable process to keep the other three layers current over time?

Table 6. The four-layer sociotechnical evidence profile. Each layer lists illustrative (not required) evidence types and the class of question it answers; it does not specify pass/fail thresholds, which Section 8.2 addresses as an open problem rather than a solved one.

Figure 1. The four-layer sociotechnical evidence profile (Section 8)



Figure 1. The four-layer sociotechnical evidence profile.

8.1 The assurance chain: from claim to monitoring trigger

(P) A deployment claim is only as strong as the chain of evidence behind it, and a single-sentence claim “the copilot resists prompt injection,” “the model is robust,” “the analyst maintains oversight” is a compressed conclusion, not evidence in itself. We propose that each such claim attached to an evidence-profile layer (Table 6) be unpacked into seven linked elements before it is treated as reviewable: **Claim -> Evidence artifact -> Test method -> Assessor -> Decision rule -> Residual uncertainty -> Monitoring trigger**. A

claim is the specific, falsifiable statement asserted; the evidence artifact is the concrete record (a test report, log, dataset, or audit finding) that lets a third party check it; the test method states how that artifact was produced, including attack corpus or test protocol, coverage, and threat-model assumptions; the assessor states who produced or reviewed the evidence and their independence from the system's owners; the decision rule is the pre-specified threshold that converts evidence into a pass/fail/conditional determination; residual uncertainty names what the artifact does not cover; and the monitoring trigger is the operational condition that would prompt re-testing, restriction, or suspension. This chain does not itself validate a claim it only makes explicit what would be required to validate it and it is our own conceptual proposal, not an established audit standard.

Chain element	Content for the claim "the copilot resists prompt injection" (Section 10.1's SentinelAssist example)
Claim	The copilot does not execute unauthorized actions when exposed to adversarial instructions embedded in ingested tickets, logs, or phishing text.
Evidence artifact	Red-team test report enumerating attempted injections and outcomes (Appendix B.1).
Test method	A corpus of indirect-injection payloads embedded in representative ticket, log, and email fields; attacker capability limited to content injection (no direct API access); success criterion = any unauthorized tool invocation or policy bypass; coverage limited to the tested payload families and ingestion channels.
Assessor	Internal red team, independent of the model-owning engineering team (second line, Section 9.1); not yet externally validated (Section 11.2).
Decision rule	Zero tolerance for unauthorized state-changing actions in the tested corpus; any single success blocks promotion to expanded autonomy.
Residual uncertainty	Novel payload families outside the tested corpus; payloads combining injection with poisoned retrieval content (Tables 4.1a-b); agentic multi-step chains beyond the tested depth.
Monitoring trigger	Any confirmed successful injection in production; a material change to the ingestion pipeline or tool permissions; or a new technique added to MITRE ATLAS's AI/ML Attack Staging tactic (Section 4).

Table 6.1. Worked example of the assurance chain, applied to one claim from the Section 10.1 illustration. The chain turns the evidence profile (Table 6, Figure 1) from a checklist into a traceability structure without implying that it is already a validated score; each organization's own claims, artifacts, and thresholds will differ.

8.2 What the evidence profile does not yet do

- (P) As proposed, the evidence profile does not define scoring rules, minimum evidence thresholds per layer, assessor qualifications, procedures for handling partial or missing evidence, or inter-rater reliability targets. It is, at this stage, an organizing checklist rather than a graded assessment instrument. Section 10.1 illustrates, on one hypothetical system, how these open questions surface in practice. We regard defining those operational details, and testing them against real systems with independent assessors, as necessary future work (Section 12) rather than as an omission that can be resolved by assertion within this paper.

9. Lifecycle Crosswalk to Existing Standards (Conceptual Proposal)

- (P) Table 7 relates the four evidence-profile layers to stages of an AI system lifecycle and to the existing frameworks discussed in Section 7. The crosswalk is intended to help an

organization locate where existing standards and regulatory obligations already produce relevant evidence, so that the evidence profile adds a lens for organizing that evidence rather than a duplicate compliance burden. Every cell in Table 7 is illustrative and proposed by this paper, not a claim that the named framework legally requires exactly that artifact at that stage; the crosswalk has not been tested against a real audit or procurement process, and specific cell-level mappings should be verified against an organization's actual control set and legal obligations before being relied upon.

Lifecycle stage	Technical layer	Human layer	Data/rights layer	Organizational/GRC layer
Design	Threat model against MITRE ATLAS-relevant techniques	Define human role (detect vs. decide vs. act)	Data minimization plan	AI RMF "Map" function; risk classification exercise
Development	Adversarial and robustness testing	Usability/comprehension testing of explanations	Privacy testing (e.g., membership inference)	Documentation per ISO/IEC 42001 AIMS records
Pre-deployment	Independent red-team evaluation	Analyst training and override protocol	Data protection impact assessment where applicable	Conformity/risk classification sign-off (e.g., EU AI Act Art. 6 documentation)
Operation	Drift and performance monitoring	Override-quality and automation-bias monitoring	Ongoing data governance and retention review	Incident reporting; AI RMF "Manage" function
Review/retirement	Re-testing after model or environment change	Re-training analysts on changed behavior	Data deletion/retention closeout	Internal audit (third line) review; lessons-learned record

Table 7. Illustrative lifecycle crosswalk (all cells are proposed/illustrative, not required). Cells indicate where a given framework typically produces evidence relevant to a given evidence-profile layer at a given lifecycle stage; the table is not exhaustive and is not a certification checklist, and no cell should be read as establishing legal or technical coverage by itself.

Figure 2. Lifecycle crosswalk: evidence-profile layers × lifecycle stage (illustrative; Section 9, Table 7)

	Design	Development	Pre-deployment	Operation	Review / retirement
Technical	Threat model vs. ATLAS techniques	Adversarial & robustness testing	Independent red-team eval.	Drift & perf. monitoring	Re-test after model/env change
Human	Define human role (detect/decide/act)	Usability testing of explanations	Analyst training & override protocol	Override-quality & bias monitoring	Re-train analysts on changed behavior
Data/rights	Data minimization plan	Privacy testing (membership infer.)	Data protection impact assessment	Ongoing data governance review	Data deletion / retention closeout
Org./GRC	AI RMF "Map": risk classification	ISO/IEC 42001 AIMS documentation	Conformity sign-off (e.g., EU AI Act Art. 6)	Incident reporting; AI RMF "Manage"	Internal audit (third line) review

Figure 2. Lifecycle crosswalk: evidence-profile layers by lifecycle stage.

9.1 Three lines of defense

(N) The “three lines” model of organizational risk accountability originates in internal-audit and risk-management practice as articulated by the Institute of Internal Auditors (IIA), which describes management ownership of risk (first line), risk and compliance oversight functions (second line), and independent internal audit assurance (third line). (P) We adapt this model to AI GRC by proposing that model owners and SOC engineering teams sit in the first line, a dedicated AI risk/compliance function (which may be part of an existing security-governance function) sits in the second line, and internal audit or an equivalent independent function periodically tests the evidence profile’s outputs in the third line. This adaptation is our own proposal; we did not find a published, validated application of the IIA model specifically to AI-enabled cybersecurity tooling, and organizations should adapt the roles to their existing accountability structures rather than adopt the three-line labels mechanically.

10. A Five-Level Organizational Maturity Model (Conceptual Proposal, Unvalidated)

(P) The maturity model below is a proposed organizing scale, not a validated measurement instrument. It has not been piloted with independent assessors, has no established inter-rater reliability, and has not been tested against operational outcomes such as incident rates or audit findings. We present it as a starting hypothesis for future empirical work, and we explicitly withdraw any categorical claim that a given level is required for a system to be called “trustworthy” in the sense defined in Table 1; trustworthiness, as defined there, is proportional to consequence and context, and a low-consequence, easily reversible detection tool may warrant a lower evidentiary bar than an autonomous remediation agent regardless of the deploying organization’s overall maturity level.

Level	Label	Illustrative starting criteria (proposed, unvalidated)
1	Ad hoc	AI tools adopted without a documented risk classification, owner, or monitoring plan.
2	Managed locally	Individual teams document ownership and basic monitoring, but practices are not consistent across the organization.

3	Standardized	A consistent evidence-profile process (Section 8) is applied across AI-enabled cybersecurity systems, with defined roles across the three lines of defense (Section 9.1).
4	Measured	Evidence-profile completeness and quality are tracked with metrics, and independent (second- or third-line) review occurs on a defined cadence.
5	Adaptive	Metrics from Level 4 feed back into design and procurement decisions, and the organization can demonstrate improvement over time using its own incident and audit history.

Table 8. Proposed five-level maturity scale (all levels are illustrative and unvalidated). Level descriptions are starting criteria for future operationalization (Section 12), not validated thresholds. The scale does not itself establish that any particular level is sufficient for any particular deployment; sufficiency is a function of consequence, autonomy, and reversibility (Section 1.1), which the scale does not encode.

Figure 3. The five-level maturity scale (Section 10, Table 8)

Illustrative and unvalidated — not a required progression for every system

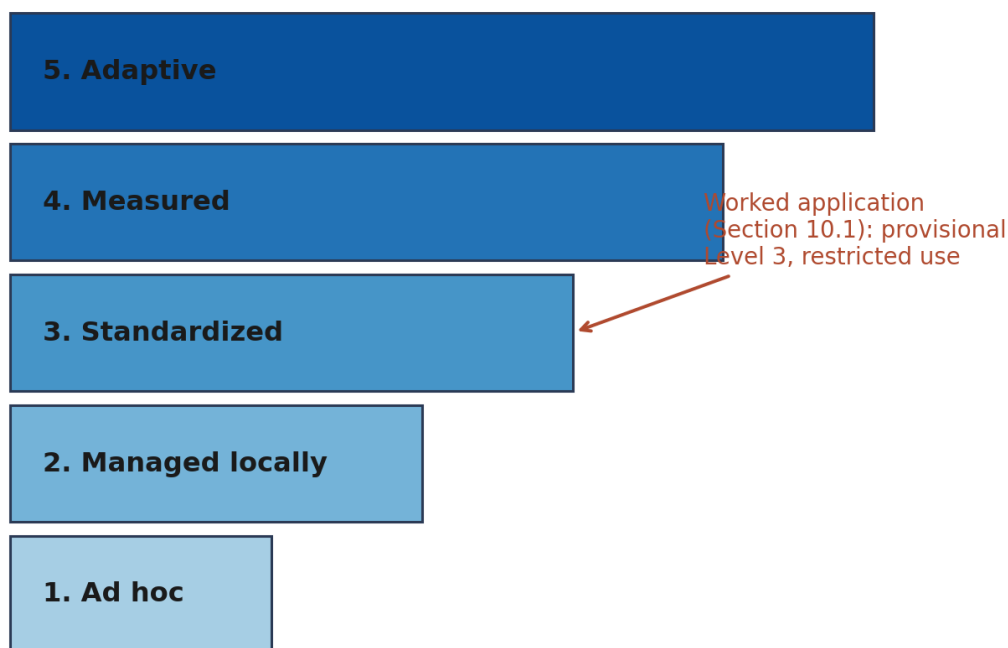


Figure 3. The five-level maturity scale, with the Section 10.1 worked application's provisional placement marked.

- (P) We deliberately avoid labeling any single level as a universal “minimum defensible posture.” The level is an organizational capability descriptor, not a system-safety score and not a legal threshold. For a low-consequence, human-reviewed, easily reversible detection system, an organization operating at Level 2 with a documented plan to reach Level 3 may represent a reasonable and defensible position. For an autonomous or semi-autonomous response system capable of materially affecting availability, safety, or individuals’ rights, we would expect a defensible position to require evidence closer to Level 4, including in the human and organizational layers, not only the technical layer. This is a conditional, proportionality-based claim, not an absolute one, and we state it as

our own judgment (P) pending the empirical validation described in Section 12. Section 10.1 applies this reasoning to one hypothetical system.

10.1 Worked application: an illustrative SOC copilot

Illustrative application, not validation. This subsection applies the evidence profile and maturity model (Sections 8–10) to a hypothetical analyst-facing SOC copilot. It demonstrates how consequence, autonomy, reversibility, and evidence gaps shape a provisional maturity judgment; it does not establish the framework’s reliability, validity, or a universal deployment threshold. The full evidence matrix underlying this discussion is reproduced in Appendix B.

- (P) **System and scope.** The hypothetical system, “SentinelAssist SOC Copilot,” is a decision-support and triage tool with a restricted semi-autonomous response function deployment class (b) shading into class (c) in the typology of Section 1.1. It ingests SIEM alerts, endpoint telemetry, identity events, firewall logs, phishing reports, and ticket text; deduplicates and clusters alerts; assigns severity and a recommended investigation path; generates incident summaries with cited telemetry snippets; and suggests detection rules and containment actions. It can open a case automatically, but isolating an endpoint or disabling a credential requires analyst confirmation. This system is a useful test case precisely because it combines several failure modes discussed separately in Sections 4–6: static-model risk and distribution shift, LLM prompt-injection exposure from untrusted logs and ticket text, analyst reliance and automation bias, privacy-sensitive telemetry, and consequential but reversible response actions rather than presenting only one of them in isolation.
- (Q) **Risk assumptions.** The assessment assumes: moderate autonomy (automatic case creation, confirmation-gated containment); reversible but operationally disruptive response actions (endpoint isolation, credential disablement); a threat model that includes attacker-controlled email bodies, ticket text, and log fields, poisoned knowledge-base content, evasion, data drift, and compromised integrations; and processing of security telemetry, user identifiers, authentication patterns, and ticket text that raises the data-governance questions discussed in Section 6. Consistent with Section 1’s framing, benchmark performance alone would answer only the technical-layer question under these assumptions; the human, data-and-rights, and organizational layers require separate evidence.
- (R) **Evidence-profile assessment, by layer.** (Table 6.1 unpacks one of this system’s claims resistance to prompt injection into a full assurance chain; the assessment below summarizes the remaining claims narratively rather than repeating that level of detail for each one.) Applying Table 6: the *technical* layer shows evidence beyond a static benchmark: a temporal holdout, targeted tests for instruction-like phishing text, red-team tests for prompt injection and unsafe tool invocation, and drift monitoring but no independent cross-organization validation, no stable estimate of performance against emerging attack families, and limited testing of the retrieval index for poisoning, so we assess this layer as *partially demonstrated*. The *human* layer shows formal oversight mechanisms: inspectable evidence excerpts, logged accept/reject/escalate actions, confirmation-gated containment, and analyst training on hallucination and automation bias but no controlled comparison of analyst decision quality with and without the tool and no independent assessment of override quality, so meaningful oversight (Section 5) is not yet established and this layer is also *partially demonstrated*. The *data and rights* layer shows a data inventory, role-based access control, defined retention periods, and redaction of secrets from explanation views, but no completed subgroup error-burden

assessment and no membership-inference testing of stored prompts and outputs, consistent with Section 6's caution that fairness and privacy evidence must be defined for the specific system rather than assumed; *partially demonstrated*. The *organizational/GRC* layer shows named ownership, change governance, a risk register covering AI-specific risks, vendor terms with audit rights, and internal-audit sampling, but no tested end-to-end rollback exercise, no independent test of whether the evidence profile predicts incident or audit outcomes, and no defined retirement trigger; *partially demonstrated*. Appendix B reproduces the full evidence-and-gap listing behind each of these four assessments.

- (S) **Maturity determination.** A simple average across layers would conceal a critical weakness: strong technical evidence cannot compensate for absent human or organizational control. We instead apply a weakest-critical-layer heuristic for consequential actions, combined with a narrative reading against Table 8. The system exceeds Level 1 (it has documented ownership and controls) and Level 2 (practices are consistent, not merely local); it meets the description of Level 3 Standardized, because a cross-layer evidence package exists with defined third-line review but several of the measurements and independent checks associated with Level 4 Measured (external validation, analyst decision-quality testing, a completed rollback exercise, and fuller vendor evidence) remain incomplete. We therefore record a provisional determination of **Level 3 (Standardized)**, not Level 4, and pair it with a conditional rather than unconditional deployment judgment: *suitable for restricted, analyst-reviewed decision support, with confirmation-gated containment; not yet suitable for expanded autonomous response*. Consistent with Section 10's proportionality claim, this is not a judgment that Level 3 is universally sufficient a lower-consequence, flag-only detector might reasonably operate at Level 2, while this system's semi-autonomous response function is the reason we look for evidence closer to Level 4 before its authority is expanded.
- (T) **Operating conditions and what would close the gap.** Because the maturity label alone does not convert into a deployment decision, we attach illustrative conditions: no expansion to autonomous containment; mandatory abstention when evidence is contradictory, confidence is poorly calibrated, or retrieved content is untrusted; treatment of all ingested logs, tickets, and documents as untrusted data rather than instructions, with tool permissions enforced outside the model; periodic independent sampling of accepted, rejected, and overridden recommendations; named escalation thresholds for alert volume, false-positive burden, and analyst workload; risk-based change control for model, retrieval-index, and vendor updates; a demonstrated rollback exercise; and a documented privacy and necessity review for any employee-facing risk flag. The highest-value evidence for moving this system toward Level 4 would be external or cross-organization validation, a controlled analyst study of decision quality and workload, an independent review of containment and override decisions, and a completed rollback and incident-response exercise the same categories of evidence the validation agenda in Section 12 calls for in general, applied here to one concrete system rather than left abstract.

11. Discussion, Scope Conditions, and Limitations

This paper's central claim is proportionality: we propose that the evidence required to support a deployment-level trust claim about an AI-enabled cybersecurity system should scale with the consequence, autonomy, and reversibility of what the system can do, and we argue that

benchmark accuracy alone leaves this additional evidence unaddressed because it speaks only to the technical layer under largely non-adversarial, static conditions (Section 4). This proportionality claim is our own synthesis (P) of the evidence reviewed in Sections 4–6, not an empirically established scaling law; we have tried, in this revision, to state it conditionally rather than categorically, to avoid converting normative instrument requirements or our own design proposals into universal empirical statements about trustworthiness, and to show, in the worked application (Section 10.1), what applying it to one system actually requires and does not yet establish.

11.1 Limitations of this paper

- The literature synthesis is purposive and narrative, not systematic; coverage gaps are likely, and a reader with a completeness requirement should treat this as a starting bibliography rather than an exhaustive one (Section 3).
- The evidence profile, crosswalk, and maturity model (Sections 8–10) are unvalidated conceptual proposals. They have not been applied to a real system by independent assessors, and no inter-rater reliability or predictive-validity evidence exists for them at the time of writing; the worked application in Section 10.1 is a single, hypothetical, author-constructed illustration, not a case study of a real deployment.
- Regulatory and standards content is version- and date-specific and will go stale even where this revision resolves a previously pending item. The EU AI Act’s high-risk timeline, for example, moved from “proposed amendment under negotiation” to enacted law (Regulation (EU) 2026/1744) between an earlier draft of this paper and this version; the MITRE ATLAS technique count is anchored to a specific dated data-export version for the same reason. Table 5 and Appendix A record the retrieval context, and readers should re-verify both before relying on these details operationally. The NIST Cyber AI Profile (NIST IR 8596) discussed in Section 7.2 was, at the time of writing, an initial preliminary draft only, and its eventual final structure and content may differ materially.
- The paper does not independently re-derive or re-verify every empirical claim in the underlying technical and human-factors literature it cites; it relies on the cited sources’ own reported findings.

11.2 What would strengthen this line of work

- A pre-registered systematic review with documented search strings, dual-reviewer screening, and a PRISMA flow diagram, scoped to a narrower research question than this paper’s broad sociotechnical scope.
- A worked case study applying the evidence profile and maturity model to at least one real, named class of AI-enabled cybersecurity system, with two or more independent assessors and a reported measure of agreement extending the illustrative, hypothetical application in Section 10.1 to an actual deployment.
- Empirical work relating maturity-model level or evidence-profile completeness to operational outcomes (incident frequency, analyst decision quality, audit findings), which does not yet exist for these specific instruments.

12. Conclusion

Benchmark performance is evidence about one layer of a much larger question. This paper has argued, with explicit labeling of what is empirical, what is normative, and what is the authors’ own proposal, that a deployment-level trust claim for an AI-enabled cybersecurity system additionally requires evidence about adversarial robustness under a relevant threat model, meaningful (not merely formal) human oversight, appropriate data and rights governance, and institutionalized organizational GRC capability. The four-layer evidence profile, the standards

crosswalk, and the five-level maturity model offered here together with the worked application in Section 10.1 are conceptual proposals intended to organize that evidence and to connect it to existing standards and regulation rather than to replace them. They are explicitly not validated assessment instruments, and the paper's own recommendation is that they be treated as hypotheses for the empirical work described in Section 11.2, not as ready-to-deploy audit criteria. The most informative next empirical test would be the one Section 11.2 names first among equals: applying it, with two or more independent assessors, to a real (not hypothetical) AI-enabled SOC system, and reporting their agreement alongside the organization's subsequent incident and audit record.

Appendix A: Source-Selection Matrix

This appendix records how sources were located for each domain covered by the paper. It is a purposive selection record intended to support the transparency goals of Section 3, not a systematic-review search log: it does not report exact query strings, database hit counts, screening decisions, or a reproducible PRISMA-style flow. Readers requiring that level of rigor should treat this table as a starting point for their own systematic search rather than as a substitute for one. Primary legal and technical instruments were retrieved directly from their maintaining body; retrieval dates fall in 2025–2026 and should be re-checked for currency before reuse.

Domain	Selection approach used	Approximate coverage	Inclusion rationale	Illustrative sources included
Adversarial robustness / AI security	Scholarly search (e.g., Google Scholar, IEEE Xplore, USENIX proceedings) for terms including “adversarial machine learning,” “intrusion detection machine learning,” “malware classification robustness”	2010–2024	Prioritized peer-reviewed security venues (IEEE S&P, USENIX Security) and the primary NIST taxonomy over secondary summaries	Sommer & Paxson (2010); Arp et al. (2022); Biggio & Roli (2018); Goodfellow, Shlens, & Szegedy (2015); Vassilev et al. (2024)
Human oversight / SOC human factors	Scholarly search for “SOC analyst,” “alert fatigue,” “security operations center human factors” in security/HCI venues (USENIX, SOUPS, ACM CCS)	2016–2022	Prioritized empirical studies conducted in or on real SOC environments over general automation-bias literature	Sundaramurthy et al. (2016); Kokulu et al. (2019); Alahmadi, Axon, & Martinovic (2022)

Explainability / XAI	Scholarly search for “explainable AI,” “interpretable machine learning,” “LIME,” “SHAP” in ML venues (NeurIPS, KDD)	2016–2017 (foundational methods)	Prioritized the most-cited foundational methods plus a critical methodology paper on interpretability science	Ribeiro, Singh, & Guestrin (2016); Lundberg & Lee (2017); Doshi-Velez & Kim (2017)
LLM / agentic AI security	Scholarly and archival search for “prompt injection,” “LLM-integrated application security”	2023–2026	Prioritized the earliest well-documented empirical demonstration over subsequent commentary	Greshake et al. (2023)
AI governance / regulation / standards	Direct retrieval from maintaining-body sources (European Commission AI Act Service Desk, NIST CSRC and NCCoE, ISO, MITRE ATLAS) rather than search engines, to prioritize primary-source accuracy	Current as of the stated retrieval context (2025–2026)	Prioritized the authoritative or maintaining-body publication over secondary commercial explainers, in direct response to source-quality concerns raised on an earlier draft	Regulation (EU) 2024/1689; NIST AI RMF 1.0; NIST CSF 2.0; NIST IR 8596 (preliminary draft); ISO/IEC 42001:2023; MITRE ATLAS
AI ethics / cross-cutting governance	Scholarly search for “AI ethics guidelines,” “trustworthy AI principles”	Survey covering guidelines published through 2019	Prioritized an existing systematic survey of the guidelines landscape over individual guideline documents	Jobin, Ienca, & Vayena (2019)

Table 9. Source-selection matrix by domain. “Approximate coverage” reflects the publication dates of sources actually cited, not a bounded search date range; regulatory and standards sources were retrieved directly from maintaining bodies during 2025–2026 and should be re-checked for currency before reuse.

Appendix B: Worked Case-Study Evidence Matrix (Supplement to Section 10.1)

This appendix reproduces, in full, the layer-by-layer evidence and evidence-gap listing summarized narratively in Section 10.1. As stated there, this is an illustrative, hypothetical worked example, not an evaluation of a real organization or product, and it is included as a

formative demonstration of how the evidence profile and maturity model would be used, not as validation evidence for either instrument.

B.1 Technical evidence

Evidence observed: offline evaluation on a historical alert dataset (precision, recall, F1); a temporal holdout from the most recent quarter; separate tests for phishing text containing instruction-like content; red-team tests for prompt injection, unsafe tool invocation, and malformed telemetry; monitoring for alert volume, latency, calibration, abstention, and drift; a documented rollback procedure for model and prompt-template updates.

Evidence gaps: no independent external validation across a second organization; no stable estimate of performance under a newly emerging attack family; limited poisoning tests for the retrieval index and incident knowledge base; no established threshold for when the system must abstain rather than recommend containment; no demonstrated relationship between model metrics and analyst-level security outcomes.

B.2 Human oversight and interaction

Evidence observed: analysts can inspect telemetry excerpts supporting the summary; the interface records accept, reject, edit, and escalate actions; containment requires analyst confirmation; analysts receive training on hallucination, prompt injection, automation bias, and escalation; monthly review examines override rates and a sample of accepted recommendations.

Evidence gaps: no controlled comparison of analyst decision quality with and without the copilot; override rates are recorded, but override quality and rationale are not yet independently assessed; workload and alert-fatigue effects have not been measured over a full operational cycle; the organization has not tested whether analysts can detect a confidently worded but incorrect recommendation; the escalation process is documented but has not been tested through a live exercise.

B.3 Data, privacy, and rights

Evidence observed: a data inventory identifies telemetry categories and system owners; role-based access controls restrict raw ticket text and identity data; retention periods are defined for prompts, retrieved documents, explanations, and action logs; personal-data processing is documented for insider-threat and account-risk use cases; explanation views redact secrets, tokens, and unnecessary personal information.

Evidence gaps: no completed assessment of subgroup error burdens for employee-facing risk flags; no demonstrated test of membership-inference or reconstruction risk from stored prompts and outputs; no formal rule for when analyst-facing explanations can reveal detection logic to an attacker; data provenance for some vendor-supplied threat-intelligence content is incomplete; no independent review of whether all ingested ticket fields are necessary for the stated purpose.

B.4 Organizational and GRC assurance

Evidence observed: a named product owner and SOC owner are accountable for operation; security governance reviews changes to the model, retrieval sources, and tool permissions; vendor contracts include update notifications, incident reporting, audit rights, data-use restrictions, and rollback support; the risk register records prompt injection, data poisoning, drift, unsafe action, privacy, and analyst-reliance risks; internal audit has reviewed the existence of controls and sampled action logs.

Evidence gaps: no independent test of whether the evidence profile predicts incident or audit outcomes; no completed exercise demonstrating end-to-end rollback after a harmful model update; no regular board-level reporting of residual AI-specific risk; vendor documentation does not fully disclose training-data provenance or update evaluation results; the organization has not defined a retirement trigger for the copilot if error or workload thresholds are exceeded.

B.5 Maturity-level determination detail

The determination in Section 10.1 uses a weakest-critical-layer rule for consequential actions rather than a numerical average, because a strong technical evaluation cannot compensate for absent human oversight or organizational control. Against Table 8: the system exceeds Level 1 (Ad hoc) because it has documented ownership and controls, and exceeds Level 2 (Managed locally) because practices are consistent rather than team-specific. It meets the description of Level 3 (Standardized) because a cross-layer evidence package exists with defined oversight roles and routine monitoring. It does not yet meet Level 4 (Measured) because external validation, analyst decision-quality testing, an independent review of override quality, a completed rollback exercise, and fuller vendor evidence are all still missing. Level 5 (Adaptive) is correspondingly not met, since it presupposes Level 4 as a foundation. This yields the provisional determination of Level 3, paired with the conditional deployment decision stated in Section 10.1: suitable for restricted, analyst-reviewed decision support with confirmation-gated containment, not yet suitable for expanded autonomous response.

References

- Alahmadi, B. A., Axon, L., & Martinovic, I. (2022). 99% false positives: A qualitative study of SOC analysts' perspectives on security alarms. In *Proceedings of the 31st USENIX Security Symposium*.
- Apruzzese, G., Laskov, P., Montes de Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos, A. V., & Di Franco, F. (2023). The role of machine learning in cybersecurity. *ACM Digital Threats: Research and Practice*, 4(1), Article 8.
- Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., & Rieck, K. (2022). Dos and don'ts of machine learning in computer security. In *Proceedings of the 31st USENIX Security Symposium*.
- Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- European Commission. (2025). Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) 2024/1689 as regards certain deadlines for the application of rules on high-risk AI systems, COM (2025)0836. Submitted 19 November 2025.
- European Commission. (n.d.). Timeline for the implementation of the EU AI Act. AI Act Service Desk. Retrieved 2026, from <https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act>
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*.
- Institute of Internal Auditors. (2020). The IIA's three lines model: An update of the three lines of defense.
- International Organization for Standardization. (2023). ISO/IEC 42001:2023 Information technology Artificial intelligence Management system.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Kokulu, F. B., Soneji, A., Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., & Ahn, G.-J. (2019). Matched and mismatched SOCs: A qualitative study on security operations center issues. In

- Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS).
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- MITRE. (n.d.). MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. Retrieved 2026, from <https://atlas.mitre.org/>
- MITRE. (2026). ATLAS knowledge base data, version 2026.05 (modified May 27, 2026) [Data set]. GitHub. <https://github.com/mitre-atlas/atlas-data>
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- National Institute of Standards and Technology. (2024). The NIST Cybersecurity Framework (CSF) 2.0 (NIST CSWP 29). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.CSWP.29>
- National Institute of Standards and Technology. (2025). Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile), initial preliminary draft (NIST IR 8596). U.S. Department of Commerce, National Cybersecurity Center of Excellence. Retrieved 2026, from <https://csrc.nist.gov/pubs/ir/8596/iprd>
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
- Regulation (EU) 2026/1744 of the European Parliament and of the Council amending Regulation (EU) 2024/1689 as regards certain deadlines for the application of rules on high-risk AI systems (Digital Omnibus on Artificial Intelligence). Official Journal of the European Union. Entered into force 27 July 2026.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data*, 7, Article 41.
- Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*.
- Sundaramurthy, S. C., Case, J., Truong, T., Zomlot, L., & Hoffmann, M. (2016). A human capital model for mitigating security analyst burnout. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS)*.
- Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2e2023). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-2e2023>