

Evidence for Whom? Contestability and Socio-Technical Assurance in High-Risk AI Governance

Rohith Reddy Bellibaltu

Florida International University

Manpreet Singh

Boston University

Preeti Uikey

Boston University

Rahul Joshi

`rahulj@sitpune.edu.in`

Symbiosis International University

Research Article

Keywords: Conformity assessment, Algorithmic accountability, AI evaluation, EU AI Act, NIST AI Risk Management Framework, ISO/IEC 42001

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10864713/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Evidence for Whom? Contestability and Socio-Technical Assurance in High-Risk AI Governance

Abstract

In high-risk artificial intelligence (AI) governance, the parties who can produce evidence and those who can use it are not the same. We compare three instruments: the US NIST AI Risk Management Framework, the European Union (EU) AI Act, and ISO/IEC 42001, an AI management-system standard. Our deposited crosswalk relates twelve evaluation-method families to 22 selected requirements, distinguishing artifact-centered evidence from evidence about deployed processes and contestation. Under our coding, several requirements ask chiefly for evidence about deployed processes, while the method families supply evidence about model artifacts. This is a claim about kinds of evidence, not a census of the literature: model tests can contribute to assurance but cannot alone establish oversight performance, evidentiary traceability for contestation, or deployment-impact governance. Two non-author raters recoded a stratified 80-cell sample. They agreed with the mapping on 88.8 percent of sampled cells at the decisive primary-versus-not boundary (binary kappa 0.68–0.70), and diverged more on the finer ordinal distinctions, where four-level exact agreement was 53.8–57.5 percent (unweighted kappa 0.38–0.43). We read that divergence as a finding rather than a defect: the requirements are open to reasonable competing readings, which is the contestability this article is about. We argue that an assurance regime whose evidence reaches only intermediaries wrongs those it governs, procedurally and epistemically, rather than merely leaving a design gap. The crosswalk is an author-constructed analytical mapping, versioned and deposited so that any cell can be contested and recoded, not a validated measurement instrument.

Keywords: Conformity assessment; Algorithmic accountability; AI evaluation; EU AI Act; NIST AI Risk Management Framework; ISO/IEC 42001

1. Introduction: Evidence for Whom?

Suppose a rejected job applicant suspects that an AI screening system misjudged them. The provider may have robustness and fairness reports prepared for conformity assessment. The applicant, however, may be unable to see them, may lack a record adequate for investigating the individual outcome, and may have no evidence that the nominal human overseer could have caught the error. High-risk AI accountability passes through regulators, notified bodies (independent conformity-assessment organizations designated under EU law), auditors, civil-society reviewers, providers, and deployers. For each intermediary, the question is what evidence of trustworthiness deserves credence. An evidence regime therefore shapes which questions can be answered, by whom, and at what stage.

Three instruments define different parts of the assurance landscape for high-risk AI: the NIST AI Risk Management Framework, voluntary guidance for identifying, measuring, and managing AI risk [1, 2]; the EU AI Act, Regulation

(EU) 2024/1689, which binds high-risk systems [3]; and ISO/IEC 42001:2023, a voluntary certifiable AI management system [4]. Section 2 introduces each; Online Resource 1 treats them fully. The article's claim is normative and its apparatus is subordinate to it: the crosswalk exists to locate, in specific requirements, where the evidence that would answer a contestation question is thinly represented, and Table 1 to record who could obtain it. A reader wanting the argument rather than its construction can read Sections 2, 7 and 8 and treat Sections 3 to 6 as the working. They are not one integrated regime and do not impose equivalent duties. We compare them because providers, deployers, auditors, and regulators use their vocabularies together when assembling assurance cases. The comparison asks a distributional question: which claims can be supported by pipeline-produced evidence, and which require observing a deployed workflow, institution, or contested decision? Our crosswalk makes one answer inspectable. The result is consistent with an asymmetry between well-instrumented model properties and weakly instrumented socio-technical processes, but does not establish that the instruments caused the pattern.

We call the umbrella problem a *socio-technical assurance deficit*: the available evidence does not directly test a claim whose truth depends on actors, workflows, and deployment conditions beyond the model itself. We distinguish three core assurance demands: oversight performance, evidentiary traceability for contestation, and deployment-impact governance. Action-trajectory assurance for tool-using systems is treated as an exploratory extension because the crosswalk does not test it. The common issue is evidentiary rather than ontological. A model-level artifact may be relevant to each demand, but cannot by itself settle whether the socio-technical claim holds.

Contributions. First, this article synthesizes contestability, oversight, and organizational-assurance work into three distinct demands for reading regulatory evidence. Second, it contributes a transparent comparative crosswalk that organizes a purposive evidence set through critical framework synthesis, grouping technical evaluation methods into twelve families and grading their relevance to 22 selected requirements. Two non-author raters tested the strongest relevance grade, the primary-code boundary, on a stratified sample, and the data and code permit cell-level recoding. Third, it develops an interpretive institutional hypothesis about why pipeline-produced evidence may not answer questions in deployed institutions. The crosswalk disciplines that argument but does not test its cause. Fourth, and drawing on the first three, it argues that an assurance regime whose evidence reaches only intermediaries wrongs those it governs rather than merely leaving a design gap (Section 7). That argument runs on the access facts recorded in Table 1, not on the cell counts.

Positioning. Existing work supplies legal comparisons, standards mappings, method catalogues, and analyses of individual obligations [5, 6]. Doctrinal scholarship has separately read the Act's architecture closely, including the consequences of building a fundamental-rights instrument on a product-safety chassis [7]; our question is downstream of that critique, asking what evidence the resulting requirements actually attract. Thomas et al. [8] broaden the harms that assurance should address. Laux [9] specifies institutional conditions for effective oversight; contestability-by-design work specifies arrangements through which decisions can be questioned [10, 11]. We do not claim a new theory of audit, oversight, or contestability. That transparency does not confer the ability to contest is Ananny and Crawford's point [19], and that audit can substitute ritual for assurance is Power's [12]. What we add is a systematic account of where, across three named instruments, the evidence that would answer a contestation

question is thinly represented; who can obtain it; and a reconstruction-adequacy criterion that follows from those two facts. Our contribution is synthesis and instrumentation: the crosswalk maps evaluation outputs to heterogeneous requirements through inspectable, graded cell judgments. The distributional question about who can use those outputs is an institutional interpretation layered onto the matrix, not another coded variable.

Scope. The analysis covers *high-risk* AI (the EU AI Act's legal class) as well as *high-stakes* AI (the broader normative category of severe consequence to safety, livelihoods, or rights), including large language models (LLMs) and agentic systems. Because all three instruments are Global-North artifacts, the analysis remains bounded to their governance cultures; Section 9 develops this limitation. This is a critical comparative analysis, not legal advice. Extended methods, tables, and legal notes appear in Online Resource 1. Online Resource 2 is the anonymized data-and-code package: it contains the matrix, the raters' cell-level codes, derived agreement tables, the within-team cross-check ledger, search documentation, and analysis scripts.

2. The Politics of Conformity Evidence

A conformity regime's definition of proof is a political settlement, not a neutral inventory of measurements. Power [12] showed how audit societies replace substantive assurance with *rituals of verification*: certification that a process occurred comes to stand for the protection itself. The pattern recurs in AI governance. Without agreed objects and access, audits invite "audit-washing," in which procedure is certified instead of outcome [13–18]. Fabrication is rarely the problem. Instead, on these accounts, the evidentiary vocabulary tends to be built around what auditees can readily supply, until ease of demonstration becomes the test of proof.

Two lines of scholarship explain why convenient evidence may offer little protection. Ananny and Crawford [19] show that transparency into a system's internals does not, by itself, confer accountability: seeing is not contesting. Selbst et al. [20] explain why model-bound evaluation can miss the social relations in which a technical property acquires meaning. Elish [21] describes the human stationed at the point of automated failure as a "moral crumple zone," while Green [22] documents how difficult mandated oversight is to make effective. Laux [9] locates that difficulty in institutions: effective oversight requires roles, competence, incentives, and capacity for corrective action, rather than a human's mere presence. Sterz et al. [23] independently reach a similar operationalization. On these accounts, assurance turns on organizational and political arrangements as well as technical tests.

Article 22 of Regulation (EU) 2016/679 (GDPR) is a limited legal anchor, not a general right against consequential automated decision-making. Article 22(1) concerns decisions based solely on automated processing that produce legal or similarly significant effects. For processing permitted under Article 22(2)(a) or (c), Article 22(3) requires at least the right to human intervention, to express one's point of view, and to contest the decision. Processing under Article 22(2)(b) is instead governed by safeguards in Union or Member State law [24]. We rely on this protective capacity without applying Article 22 to every high-stakes AI decision. Contestability scholarship asks how a lifecycle can enable decisions to be questioned and revised [10, 11].

Table 1 specifies the three demands, the evidence each requires, and who can obtain that evidence by default.

Table 1. Three socio-technical assurance demands, contrasted with artifact-oriented accuracy.

Assurance demand	Object	Protective capacity	Legal or institutional anchor	Evidence needed	Obtainable by	Boundary
Oversight performance	System-deployer-overseer interaction	Detection, intervention, correction	EU AI Act Art. 14	Deployment-representative human-factors and intervention evaluation	Provider, deployer, authorities; not the affected person	Art. 14 is not itself an individual right of recourse
Evidentiary traceability for contestation	Records linked to an event and later inquiry	Investigation, complaint, explanation, contestation	Art. 12 logging; Arts. 85-86 routes	Trace linking relevant input, version state, human intervention, output, and consequential action	Provider, deployer, authorities; affected person may obtain an explanation, not the logs	Reconstruction adequacy is our proposal, not Art. 12's express rule
Deployment-impact governance	Deployer, system, context, and affected population	Anticipation and mitigation of contextual rights harms	Art. 27 for specified deployers; ISO/IEC 42001 A.5	Contextual evidence, monitoring, and where appropriate stakeholder input	Deployer, authorities; a public summary in some registrations	Art. 27 is outside the coded matrix and does not apply to every deployer
Contrast: accuracy	Performance against a defined task distribution	Reliable output	Art. 15	Performance evaluation	Provider, authorities; declared accuracy reaches deployers	Artifact-oriented under the operationalization used here

*The categories are analytic, non-exhaustive, and not legal terms of art. Access routes. *Oversight: provider holds Art. 11/Annex IV documentation describing Art. 14 measures; deployer retains controlled logs (Art. 26(6)); authorities may require compliance-relevant documentation or data (Arts. 21, 74). Traceability: provider and deployer retain logs within their control (Arts. 19, 26(6)); authorities may obtain provider-held logs on reasoned request (Art. 21) and deployer-held material (Art. 74); a person may complain (Art. 85), and an affected person meeting Art. 86(1)-(3) may obtain an explanation from the deployer, not the logs. Deployment impact: deployer holds an applicable Art. 27 assessment and notifies the market-surveillance authority, subject to the Art. 46(1) exemption; for public-authority registrations outside restricted categories a summary is public in the EU database (Arts. 27(3), 49(3), 71(3)-(4); Annex VIII Section C). Accuracy:** provider holds the validation and testing evidence and declares accuracy in the instructions for use (Arts. 13(3)(b)(ii), 15(3); Annex IV); the evidence enters the applicable Art. 43 route (Section 6). That the AI Act confers no direct affected-person right to it does not exclude rights under data-protection, freedom-of-information, employment, sectoral, procedural, or national law. Regulation (EU) 2026/1744 amended Arts. 111 and 113, to which all AI Act routes remain subject.**

These demands concern different objects: performance in one case, record adequacy in another, and an organizational process in the third. What joins them is an assurance claim whose adequacy depends on deployment

relations, rather than on a model artifact alone. We selected the three because the reviewed governance literature treats effective oversight, investigation of contested outcomes, and contextual impact assessment as protective capacities that artifact-level testing cannot settle. Other socio-technical demands could satisfy the same test.

A claim belongs to the umbrella category only when three conditions hold. It depends on behavior or interaction in an actual or realistically represented deployment context; model-level evidence is insufficient to establish adequacy; and failure impairs a specified protective capacity. A stable, portable artifact-level test that established adequacy across contexts without observing organizational practice, human interaction, or consequential action would falsify the category. Article 15 accuracy provides a negative case because a bounded performance evaluation can answer its principal evidentiary question. Article 11 (technical documentation) provides another: its formal completeness can be assessed as a record. Fairness sits at the boundary. Fairness as a social ideal is socio-technical, but the fairness family in our taxonomy (M2; Section 4) principally operationalizes it through technical metrics; we classify the evidence form, not the ideal.

The matrix cannot independently test the three-demand typology because both arose from the same literature and coding. It can show whether the distinctions organize the selected evidence, and it can generate hypotheses for later research. A stronger test would fix the typology prospectively, draw on an independently selected evidence set, disaggregate the legal provisions, and include deployment actors and affected people in validation.

Table 2 contrasts the three instruments; the **NIST AI RMF 1.0** is voluntary guidance built around GOVERN, MAP, MEASURE, and MANAGE [1, 2]. The **EU AI Act's** core technical obligations are in Chapter III, Section 2, Articles 9 to 15 [3]. We describe them in the present tense as enacted law, not as presently enforceable duties. Regulation (EU) 2026/1744, in force since 27 July 2026, defers Chapter III application to 2 December 2027 (Article 6(2), Annex III) and 2 August 2028 (Article 6(1), Annex I), and amends Article 43(3) [25]. That deferral leaves more time to address the evidentiary gaps identified here. Article 14 requires systems designed so natural persons can effectively oversee their use. Article 12 requires that systems technically allow for automatic event logging appropriate to the intended purpose; it does not require replay of an individual decision or grant access to logs. Articles 85 and 86 add a complaint route and, for specified Annex III decisions (point 2 excepted) with legal or similarly significant effects, an explanation of the system's role rather than access to logs (Table 1). Reconstruction adequacy is therefore our proposed assurance criterion, not Article 12's present legal meaning. **ISO/IEC 42001:2023** specifies an AI management system and can be used alongside ISO/IEC 23894:2023 [4, 26]; certification is optional [27]. The instruments differ in legal force, unit of assessment, responsible actor, and assurance mechanism; neither implementing ISO/IEC 42001 nor certifying against it establishes an individual system's AI Act compliance.

Table 2. The three instruments are not equivalent assessment objects.

Dimension	NIST AI RMF 1.0	EU AI Act (Reg. (EU) 2024/1689)	ISO/IEC 42001:2023
Legal status	Voluntary guidance	Binding law (phased application)	Voluntary, certifiable standard

Dimension	NIST AI RMF 1.0	EU AI Act (Reg. (EU) 2024/1689)	ISO/IEC 42001:2023
Object assessed	The organization's AI risk-management <i>practice</i>	The <i>high-risk AI system</i> , per intended purpose	The organization's <i>AI management system</i>
Responsible actor	Any AI actor (adopter)	Provider (with deployer duties)	The organization adopting the standard (certification optional)
Assurance mechanism	Self-directed; no external conformity assessment	Conformity assessment, self- or notified-body (Art. 43)	Internal/self-assessment, with optional third-party certification audit [27]
Role of technical evidence	Populates the MEASURE function	One input among the quality management system, documentation, and the conformity-assessment route	Input to AI-management-system processes and audits; a certificate attests management-system conformity, not individual-system performance

3. Method: Critical Framework Synthesis and Comparative Crosswalk

This article is a **critical framework synthesis supported by a purposive, citation-verified comparative crosswalk**; the label adapts critical interpretive synthesis [28] to a comparison of governance instruments, not empirical studies. The crosswalk and its recoding check are the article's organizing apparatus, making the synthesis inspectable and correctable, rather than a standalone empirical study. We document evidence assembly through a retrospective provenance accounting (Fig. 1), not a PRISMA or other systematic- or scoping-review reporting standard: the historical search is not reconstructable, so no PRISMA flow diagram can be produced, and the article claims no exhaustive census of evaluation research. The reproducibility artifact supports reconstruction of the deposited crosswalk, coding, adjudication, reliability calculations, tables, and figures—not of the literature search. RQ1 asks which technical-evaluation families the evidence set represents; RQ2, how they can contribute evidence to selected requirements and with what relevance; RQ3, which coverage deficits appear under this coding and which socio-technical assurance problems they illuminate. The third is interpretive: absence from the crosswalk may reflect retrieval, classification, requirement granularity, or genuine method scarcity.

The affected-person interest that motivates this article shaped RQ3 and our coverage-deficit definition (Section 5). The taxonomy and coding derive from technical evaluation literature, and no affected persons or advocates were consulted. Participatory validation remains missing, although participation is not itself a design fix [29]. Empirical work has elicited decision subjects' contestability needs directly [30]. We are aware of no study that tests an evidence crosswalk against those needs.

Evidence set. The retained project records list 152 candidates. We excluded 11 (six unverifiable against a canonical record and five duplicates), leaving 141, then added six separately verified privacy records. The deposited evidence set comprises three non-disjoint pools: **147 indexed records**; a previously published **202-record fairness supplement**, used only for fairness family M2 (family IDs defined in Table 3; Authors [31], cited in blinded form for double-anonymous review); and **five official texts** (NIST AI 100-1 and 600-1; Regulation (EU) 2024/1689; ISO/IEC 42001:2023 and 23894:2023). After a canonical manifest resolves seven duplicate memberships, **347**

unique sources across 354 pool records remain. The English-language search, with snowballing, covered IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and arXiv during 2015–2026. The last search was in June 2026. We deposited the search strings and included records.

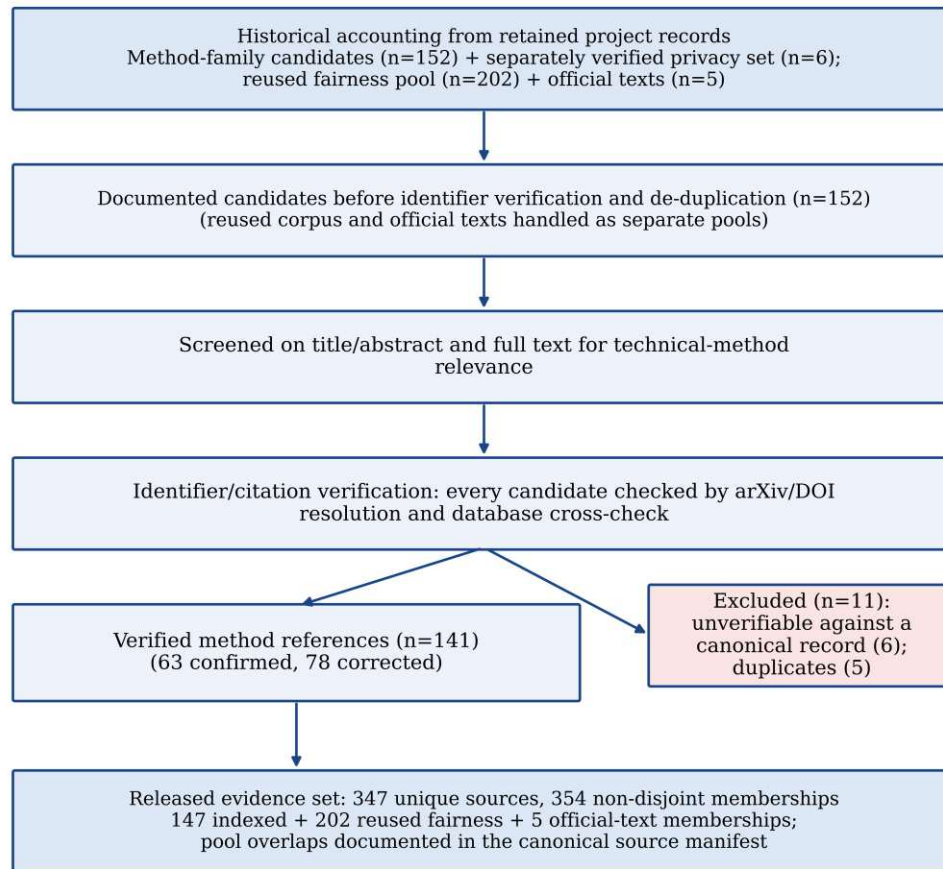


Fig. 1 Evidence-set provenance and accounting diagram. This is a historical accounting assembled from retained project records, not a PRISMA-style systematic-review flow diagram or a reproducible record-level search flow. Of 152 documented candidates,

11 were excluded and 141 retained; per-database yields and a complete record-level exclusion ledger are unavailable. Six separately verified privacy records bring the indexed-record total to 147. Together with a 202-record fairness supplement and five official texts, these records comprise 347 unique sources represented by 354 non-disjoint memberships

Screening and inclusion. Screening was single-reviewer; only the relevance coding was independently recoded, so the agreement figures below exclude screening. Inclusion required a method to map to at least one selected requirement. Density is therefore partly an inclusion artifact. Each included technical-method reference has a verified DOI, publication identifier, or resolving official URL.

Mapping proceeded in three passes: family assignment, relevance grading, and re-examination of over-claims. Family assignment is interpretive; overlapping methods were assigned by principal output, and we make no separate reliability claim for the taxonomy.

Independent recoding. The four-level scale is not-applicable < contextual < supporting < primary, where primary means a method's output is the evidence the requirement chiefly asks for; kappa (κ) measures agreement beyond chance, 0 being chance-level and 1 perfect. Two non-author raters recoded a stratified 80-cell sample, drawn as 20 cells from each of the four first-pass author code categories, using standardized, author-prepared descriptions of what each method produces and each requirement asks for, without seeing the authors' cell codes or consulting each other; both confirmed this in writing. The exercise assesses whether the resulting cell-level relevance judgment can be reproduced from that supplied representation. It does not validate the fidelity of the author-prepared summaries to the official clause text, nor the search, family taxonomy, legal interpretation, socio-technical assurance construct, or coverage-deficit definition.

Agreement was strongest on the distinction that carries the mapping's claims. Against the pre-rater frozen matrix, each rater matched the primary-versus-not boundary on 88.8% of sampled cells (binary $\kappa = 0.68$ and 0.70), and quadratically weighted κ , which credits near-misses on an ordinal scale, was **0.73** and **0.67**. Four-level exact agreement was lower: 57.5% (technical rater, unweighted $\kappa = 0.43$) and 53.8% (legal rater, unweighted $\kappa = 0.38$), with most mismatches one step apart. Part of that spread reflects the kappa paradox under skewed marginals; Gwet's AC2, which resists it, is 0.79 and 0.73 on the weighted scale, and Online Resource 1 tabulates every coefficient with bootstrap intervals. We report the lower figures because they carry information. Experts reading the same requirement description disagree about how much of its evidentiary demand a method family meets, which is the interpretive openness this article attributes to regulatory requirements, observed here on our own mapping. That reading is an interpretation rather than a result. The evidence that would separate principled disagreement from ambiguity in our own task wording is the raters' free-text commentary, which is not published, so we cannot demonstrate the distinction and do not claim to have ruled the second explanation out. The two raters agreed more with each other (weighted $\kappa = 0.82$; 72.5% exact) than with the frozen author matrix, which admits an unflattering reading we should state: if the requirements were simply open to competing readings, the raters should have diverged from one another too, and they did not. The pattern is also consistent with our first-pass coding being the outlier. The consensus rule reached only the seven qualifying cells within the 80 sampled, so any such tendency persists uncorrected in the 184 cells no rater saw; this reflects shared divergence from the authors, not validation of that matrix. Because the sample was balanced across the four code categories, primary codes are over-represented (20 of 80 sampled cells vs. 25 of 264), and chance-corrected coefficients shift with prevalence: these figures describe the balanced sample, not agreement across all 264 cells, let alone every cell or family's validation. M9 (transparency and model documentation) had no sampled cell; M8 (data quality and documentation) had two. That shortfall is structural: the deposited sampler ordered each 20-cell sample by family label, selected at a fixed interval, and stopped at twenty, so the last-sorting families M8 and M9 fell in the discarded tail. In the post-rule matrix, five of the thirteen single-primary-cell columns have a decisive cell neither rater saw: EU Articles 10, 11, and 13 and ISO/IEC 42001 A.7 and A.8, whose decisive codes come from M8 and M9. Those five columns warrant caution, but

they are not unexamined: the within-team cross-check, coded blind to the author matrix across all 264 cells, codes all five decisive cells primary and agrees with the post-rule matrix on 97.7% of the primary-versus-not boundary across the 44 M8/M9 cells. Being within-team rather than non-author, it is weaker evidence than the 80-cell exercise.

After the recoding, and in response to it, we applied a symmetric rule: where both raters crossed the primary/not-primary boundary in the same direction against our first-pass code, their shared judgment entered the revised mapping. The category-balanced sample exposed the rule to 80% of frozen primary cells but only roughly a quarter of the rest; removals were therefore easier to detect than additions. The sample is neither random nor family-representative. It supports no matrix-wide projection, and the direction of the rule's residual effect is undetermined. The rule removed four primary codes and added three across the seven qualifying cells, each identified in Table 6. Because the rule was one choice among several, we also computed the matrix under a looser alternative in which a single rater crossing the boundary is enough. That rule moves eleven cells rather than seven and yields six zero-primary columns rather than five, adding NIST MAP. Article 12 is zero-primary under our rule and under the looser one alike, so that result never depended on which author-side rule was adopted once the raters' judgments were admitted at all; it depended on admitting them, and the re-adjudication then set it aside. It does depend on admitting them: under the frozen matrix M11 remains primary for Article 12, which is the dependency Table 6 records. What the looser rule additionally changes is NIST MAP, and we report that rather than choosing the rule that suppresses it. Against the author post-rule matrix, agreement rose to weighted $\kappa = 0.71\text{--}0.77$ and 97.5% at the binary boundary (binary $\kappa = 0.93$). Against the adjudicated matrix that governs the results, which returns one sampled cell to a code both raters rejected, the same figures are 96.2% at the binary boundary (binary $\kappa = 0.89\text{--}0.90$) and weighted $\kappa = 0.70\text{--}0.76$. Both sets are post-adjudication descriptions, not independent validation. Against the frozen matrix, a within-team cross-check reaches 48.1% exact agreement across 264 cells (unweighted $\kappa = 0.24$, weighted $\kappa = 0.56$, Gwet AC2 = 0.80; Cohen [32]; Gwet [33]; Feinstein and Cicchetti [34]); the weighted figures complement rather than erase the lower ones. Both matrices, the raters' cell-level codes, agreement tables, and a full change ledger are in the deposited package. The authors applied the seven changes under the stated rule, and they were not at that point re-adjudicated by anyone else; the next subsection describes the re-adjudication that followed. We therefore report three stages, all deposited: the frozen pre-validation matrix, the author post-rule matrix, and the adjudicated matrix that governs the results below (Table 6). All three yield the same qualitative pattern, in which the zero- and single-primary columns are the process-oriented requirements. Their zero-primary sets differ at three columns in total, NIST MEASURE 2.10 and NIST MANAGE between the first two stages and Article 12 between the last two, and the re-adjudication resolved Article 12 against our own post-rule code.

Independent adjudication of the seven changed cells. The seven cells the consensus rule moved were re-adjudicated by three adjudicators working from a regenerated blinded packet containing only the method-output and requirement descriptions, with no author codes and the items out of ledger order. The M11 description was corrected first, so that the adjudicators saw the full Table 3 definition rather than the abbreviated wording the raters had received. They confirmed six of the seven outcomes, two of those six by a two-to-one majority rather than

unanimously, and did not confirm the seventh, E233, where all three coded primary against our supporting. That reversal is applied in the governing matrix. Every vote and reason is deposited.

4. The Evidence Core: Twelve Families, Three Frameworks, One Matrix

From this coding we derive twelve method families. Table 3 names them and records the depth of its evidence base. In this corpus, strong relevance to a requirement can still rest on little evidence.

Table 3. Principal outputs of the twelve method families represented in this evidence set.

ID	Family	What it principally evidences	Maturity (LLM/agentic)	Evidence base (records; provenance)
M1	Performance & benchmarking	Accuracy and reliability on held-out and out-of-distribution (OOD) data; performance monitoring	Mature (classical); evolving (agentic)	16 bespoke
M2	Fairness & bias	Group/individual/counterfactual fairness; disparate impact; action-level disparity	Mature (model); emerging (agentic)	reused pool (202 records; not all methods)
M3	Robustness & distribution shift	Stability under shift, corruption, and adversarial perturbation	Mature (vision); emerging (LLM)	14 bespoke
M4	Safety evaluation	Hazard identification, dangerous-capability elicitation, content-harm rates, safety cases	Emerging	10 bespoke
M5	Red-teaming & adversarial testing	Resistance to jailbreaks, (indirect) prompt injection, agent tool-misuse	Emerging, fast-moving	13 bespoke
M6	Interpretability & explainability	Attribution, concept testing, probing, explanation/chain-of-thought faithfulness	Maturing; faithfulness contested	11 bespoke
M7	Uncertainty & calibration	Calibration (expected calibration error, Brier score), conformal prediction, hallucination detection	Mature (calibration); emerging (LLM)	15 bespoke
M8	Data quality, governance & docs	Data quality, representativeness, provenance, contamination/leakage	Maturing	12 bespoke
M9	Transparency & model documentation	Completeness/quality of the evidentiary record (model/system cards, FactSheets, indices)	Maturing	14 bespoke
M10	Human oversight & interaction	Oversight effectiveness, automation bias, appropriate reliance, escalation	Emerging	14 bespoke
M11	Post-deployment monitoring & assurance	Drift detection, runtime monitoring, logging, incident reporting, third-party audit	Emerging	12 bespoke
M12	Privacy evaluation	Leakage detection (membership-inference, extraction), differential-privacy guarantees	Mature (attacks); emerging (LLM/agentic)	6 bespoke (shallowest)

Note: Maturity labels are descriptive judgments conditional on evidence depth, not an independently validated scale; the evidence-base column supplies their basis. Exact per-family counts come from the deposited method index. Eleven bespoke family rows account for 137 indexed records; the remaining 10 of 147 are governance background and index sources not assigned to a method family. The index includes both official NIST documents, AI 600-1 under M11 and AI 100-1 under M1, so each count includes one framework document, not only bespoke methods. M2 has no indexed rows of its own and draws instead on the reused 202-record fairness supplement. That pool is the largest in the study, was assembled by us for a different question, and nonetheless carries graded codes across the matrix on the same footing as bespoke families, including one of the seven adjudicated cells. We flag the asymmetry rather than defend it.

The twelve families fall into three evidentiary groups, and that grouping affects which requirements they can serve.

Artifact-centered families (M1–M7, M12) principally measure a model or bounded system artifact. M1 covers held-out and out-of-distribution accuracy [35, 36]. M2 covers group, individual, or counterfactual fairness [37, 38]; incompatible criteria allow a justified choice, not a universal verdict [39, 40]. M3 tests stability under shift and adversarial pressure [41, 42], while M4 elicits hazards and assembles safety evidence [43]. M5 probes policy violation under adversarial pressure [44]. M6 attributes and tests explanation faithfulness [45]; M7 checks calibration and abstention [46–48]. M12 detects privacy leakage [49–51]. Cluster names denote principal outputs, not ontological claims: fairness, safety, privacy, and explanation remain deployment-dependent.

Record-centered families (M8–M9) concern data and documentation. M8 addresses data quality, representativeness, provenance, and contamination. Documentation creates a record [52, 53]; the crosswalk credits evaluation of its adequacy. M9 concerns the completeness and verifiability of model or system cards and transparency records [54, 55].

Deployment-process families (M10–M11) concern systems in use. M10 evaluates whether an overseer can understand, supervise, and correct a system, including appropriate reliance [56]. M11 covers drift, runtime monitoring, logging, incident response, and audit [57]. Only these two families take a deployed process as their principal object, although artifact-centered methods can also apply here.

Coding levels. Each (method family, requirement) pair receives one level on the Section 3 scale, which was developed against the official clause text. **Supporting** means that a family contributes but is insufficient alone; **contextual** means that it informs without serving as evidence; a blank marks non-applicability. The levels record only technical relevance. We make no separate judgment about legal sufficiency, and the Section 3 reliability exercise tested reproducibility against the author-prepared summaries, not their fidelity to the clause text.

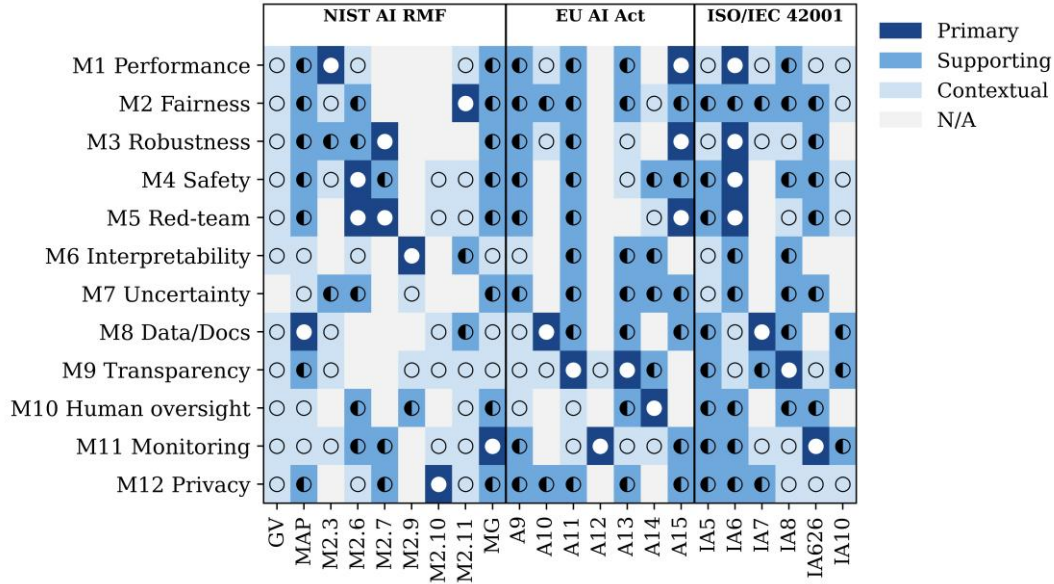


Fig. 2 Technical relevance of twelve method families (rows) to 22 selected requirements (columns). A filled cell is primary, a half-filled cell supporting, an open cell contextual, and a blank cell not applicable. Columns cover NIST GOVERN, MAP, MEASURE subcategories and MANAGE; EU AI Act Articles 9-15; and ISO/IEC 42001 Annex A controls. Density is descriptive because columns differ in breadth; they are not normalized regulatory-adequacy units

For Article 15 ("Accuracy, robustness and cybersecurity"), M1, M3, and M5 have *primary* relevance; M7 is only *supporting*, while M6 does not apply. The primary codes attach facet-by-facet, one family per facet, so none speaks for Article 15 as a whole. Access is not coded in the matrix. Instead, Table 1's obtainability column records the institutional default. A report of this kind may remain in provider, auditor, or authority channels and never furnish an applicant with evidence adequate to investigate an individual outcome.

Table 4. Condensed illustrative slice of the crosswalk: five evidence artifacts, three frameworks. (● primary, ◐ supporting, ○ contextual; values drawn verbatim from the adjudicated matrix that governs the results, as in Table 5 and Fig. 2.)

Evidence artifact (family)	NIST AI RMF	EU AI Act	ISO/IEC 42001
Adversarial-robustness report (M3)	● MEASURE 2.7	● Art. 15	● A.6.2.4
Held-out / OOD accuracy report (M1)	● MEASURE 2.3	● Art. 15 (accuracy)	● A.6.2.4
Fairness / bias audit (M2)	● MEASURE 2.11	◐ Art. 10, ◐ Art. 15	◐ A.7
Model / system card (M9)	◐ MAP documentation	● Art. 11, ● Art. 13	● A.8
Human-oversight study (M10)	◐ MEASURE (safety/expl.)	● Art. 14	◐ A.6.2.6 operation-and-monitoring*

None of the six coded ISO/IEC 42001 Annex A columns takes oversight effectiveness as its object, so M10 rises no higher than supporting. Requirements about the system-overseer relation attract weak codings repeatedly across the matrix.

Table 5. A fuller excerpt of the crosswalk: two or three families from each of the three evidence clusters, against the EU AI Act high-risk obligations (Articles 9–15). (● primary, ① supporting, ○ contextual; blank = not applicable. Articles 10, 11, and 13 rest on decisive M8/M9 cells the 80-cell sample missed (Section 3); the within-team cross-check codes all of them primary, but no non-author rater saw them.)

Method family	Art. 9	Art. 10	Art. 11	Art. 12	Art. 13	Art. 14	Art. 15
M1 Performance & benchmarking	①	○	①		①		●
M2 Fairness & bias	①	①	①		①	○	①
M3 Robustness & dist. shift	①	○	①		○		●
M8 Data quality & docs	○	●	①		①		①
M9 Transparency & model docs	○	○	●	○	●	①	
M10 Human oversight	○		○		①	●	

Reliability note (Tables 4–5): on the sampled primary-versus-not boundary each non-author rater agreed with the frozen pre-validation matrix in 88.8% of cells; Section 3 gives the four-level figures. The 97.5% agreement with the post-adjudication matrix is descriptive only: the raters' codes informed it. It does not guarantee reproducing any individual cell, nor validate the taxonomy or legal interpretation. Sub-primary distinctions remain more interpretive: exact agreement across all four levels rises only to 62.5–66.2% against the post-rule matrix, well below the corresponding primary/not agreement.

Where relevance concentrates. Relevance follows the same three-way clustering within each framework (Fig. 2): measurable model properties draw primary codes, while process obligations draw few. In NIST, most relevance lies in MEASURE 2.x. Under the EU Act, M10 is Article 14's only primary family; Article 9 is thin at every stage; Article 12 has a primary family only because the Section 3 re-adjudication restored M11. In ISO/IEC 42001, A.6.2.4 is near-universal, while no coded column has oversight effectiveness as its object (Table 4 note). Two limits apply. ISO/IEC 42001 is not itself a harmonized standard. Its artifacts may evidence that a control was designed and, when demonstrated, operates, but without a specific assurance case they establish nothing about whether an individual system satisfies EU AI Act requirements.

One scoped evaluation artifact can support claims in more than one framework. The matrix permits that reuse as evidentiary compatibility, not legal identity. A family coded ● in three columns has not satisfied those requirements; it has been judged technically relevant three times.

Contested codings and adjudication. Although oversight methods are immature, $M10 \times \text{Article 14}$ remains primary by evidence type. For $M2/M8 \times \text{Article 10}$, a primary data-quality output is distinct from supporting

fairness evidence. *M9 × Articles 11/13* is primary for the record itself, not for every transparency purpose. Seven sampled cells changed under the consensus rule (Table 6). One of them determines the Article 12 result. The frozen matrix and within-team cross-check coded M11 primary, whereas both non-author raters coded it supporting. Our rule adopted the raters' supporting judgment, on the reasoning that monitoring evidence may use logs without directly evaluating the automatic event-logging capability Article 12 requires. The re-adjudication described in Section 3 overturned that code; Section 5 gives the grounds. The deposited ledger records each change's direction, rater basis, and effect.

5. Reading the Mapping: Evidence Types and Assurance Demands

The analysis that follows reads across matrix columns and uses the Table 1 distinctions rather than framework boundaries. Coverage breadth is computed from our codes, so the deficits are properties of this crosswalk, not the complete literature, and are hypotheses for the field to test. A coding scheme can locate only harms for which it has a category: tacit, experiential, and culturally situated harms may remain invisible because the matrix has no cell to register their absence.

Start with the counts of *primary* codes across the 22 columns (full gap table in Online Resource 1; NIST Measure 2.x numbering is unrelated to families M1-M12). Under the adjudicated matrix, EU Article 15 and ISO A.6.2.4 have three or more primary families; sixteen columns have one or two, fourteen of them resting on a single primary cell; NIST GOVERN, EU Article 9, and ISO A.5 and A.10 have none. Article 14 has one, M10, and Article 12 has one, M11. NIST MAP's single primary is contested: the technical rater coded it supporting, the legal rater primary. The adjudication moved one cell, restoring M11 as primary for Article 12 and taking that column out of the zero set (Table 6). These labels are fragile and non-normalized: a column's breadth, composite character, and institutional function shape how many families map to it. One crude check runs against the narrowness explanation rather than for it: the four zero-primary columns engage 10.8 method families on average through some code, while columns with two or more primaries engage 9.0. On this measure the columns with no chief evidence source are the broader ones, not the narrower. Engaged-family count is a weak proxy for breadth, so this settles nothing on its own, and unequal column breadth remains a rival explanation throughout. Rater-substitution and within-team cross-check analyses are sensitivity checks, not independent votes. The cross-check coded two A.5 cells primary, so A.5's zero is contested too.

Table 6. Every column the adjudication touched, across the three stages of the mapping.

Column	Frozen	Author post-rule	Adjudicated (governing)
NIST MEASURE 2.6	M4	M4, M5	M4, M5
NIST MEASURE 2.10	none	M12	M12
NIST MANAGE	none	M11	M11
EU Article 12	M11	none	M11
EU Article 13	M6, M9	M9	M9
ISO/IEC 42001 A.7	M2, M8	M8	M8
ISO/IEC 42001 A.8	M6, M9	M9	M9

The other fifteen columns are unchanged at every stage. The zero-primary count differs by stage: six under the frozen matrix (NIST GOVERN, MEASURE 2.10, MANAGE; Article 9; ISO A.5, A.10), five under the author post-rule matrix (NIST GOVERN; Articles 9, 12; ISO A.5, A.10), and four after the adjudication (NIST GOVERN; Article 9; ISO A.5, A.10). The adjudication moved one cell. Our own rule had removed M11 as primary for Article 12; three blinded adjudicators unanimously restored it, and the adjudicated column governs the results reported here. Whether the author-side rule's residual effect favoured deficits is undetermined (Section 3).

Not every thin column is a failure. Process requirements such as NIST GOVERN, EU Article 9, and management-system controls are composite governance activities no single technical evaluation should be expected to satisfy; their thinness may reflect a need to assemble evidence, not invent one metric. We therefore distinguish a *method-family coverage deficit*, a descriptive matrix property, from a *socio-technical assurance problem*, a normative and institutional interpretation requiring separate argument. Because the twelve families derive from model-centered evaluation literature, sparse coverage in process-oriented columns partly reflects that starting point. The point is stronger than a caveat. Section 2 makes insufficiency of model-level evidence a membership condition of the umbrella category, the families are drawn from model-centered work, and inclusion required that a method map to some requirement. A reader is entitled to conclude that the pattern is close to analytic under this construction, and that the mapping's contribution is to locate it cell by cell rather than to discover it. A thin column records our judgment that the requirement asks chiefly for evidence about a deployed process while the families in this evidence set produce evidence about model artifacts. It does not show that deployers lack internal audit, escalation, or impact-assessment procedures, on which this study collected no data.

Oversight performance (Article 14). By evidence *type*, exactly one family serves Article 14: the nascent M10. On our reading, only three of the fourteen M10 records measure whether oversight *works*: appropriate reliance [56], reliance drills [58], and explanation effects on oversight [59]. The three-of-fourteen count is an author judgment outside the coded matrix, and its record-level classification is not in the deposited package. Although the column is moderately broad, its sole primary family is immature. The resulting pattern resembles the risk described by Elish [21] and Green [22]: a human safeguard can exist without evidence that it corrects error. For an affected person, the hypothesized consequence is a nominal review with no demonstrable error-catching capacity. This interpretation remains provisional because it depends on one immature family and a coding-sensitive column.

Evidentiary traceability for contestation (Article 12; Articles 85–86). Article 12 is the one cell where the two adjudication stages disagree, and the disagreement is instructive. Our own post-hoc rule removed M11 as primary, following both non-author raters. The independent adjudication restored it, unanimously. Three further signals point the same way: the frozen matrix coded it primary, the blind within-team cross-check of all 264 cells coded it primary, and the abbreviated M11 description supplied to the raters omitted the logging, incident-reporting and audit elements of our own Table 3 definition, which are precisely the elements this cell turns on. The supporting code therefore rested on a single defective elicitation, and we do not carry it forward.

M11 accordingly stands as the primary family for Article 12 in the adjudicated matrix, and Article 12 is not a zero-primary column. We report this at length rather than quietly correcting it because the earlier reading was the paper's

most-quoted result, and because the sequence illustrates the article's own thesis: a coding judgment survived two author-side passes and was overturned only when an independent reading was applied to a corrected description. What remains true, and is unaffected, is the narrower point about what monitoring evidence does. Records that show logs exist and are used are not the same as evidence that logging is adequate to reconstruct a contested decision. Our concern—whether records can connect a contested outcome to relevant input, system state, human intervention, and consequential action—is a normative assurance criterion informed by the Act's complaint and explanation routes; such records can materially enable investigation, but Articles 12, 85, and 86 do not expressly require a technical replay.

Deployment-impact governance (ISO A.5; Article 27). ISO/IEC 42001 A.5 has no primary family under our coding. That may be because impact assessment is a composite organizational process, not a task for one technical method. Article 27 is not in the matrix and applies only to Annex III systems (point 2 excepted) deployed by public bodies or private public-service providers, and to deployers of Annex III point 5(b)–(c) systems. We use it to illustrate a related deployment-context assurance problem, not to claim that the matrix demonstrates an Article 27 deficit.

Exploratory extension: action-trajectory assurance. Agentic action is not a fourth result of the crosswalk: as Section 1 noted, the matrix is family-grained, does not encode output-versus-action evaluation, and has no dedicated agentic legal column. Greshake et al. [44] give initial reason to ask whether tool-use trajectories create risks output-level tests miss, but one source cannot establish a distinct assurance field, and the question awaits a dedicated evidence review.

The matrix illustrates, but does not test, the three core Section 2 distinctions. Article 14 has one emerging primary family. Under our coding, A.5 has none, and Article 12 has one only because the adjudication restored it, so that result rests on a single adjudicated boundary. The institutional-design claim is interpretive. Rival explanations remain, among them a method taxonomy weighted toward technical evaluation, unequal column breadth, composite obligations, different actor responsibilities, and the relative immaturity of deployment-level methods.

6. Interpreting the Pattern: Evidence Production and Institutional Routes

Why the coverage pattern arose cannot be identified from the matrix. The rival explanations at the end of Section 5 remain plausible, so the political-economy account below is a hypothesis rather than a causal result of the analysis.

One plausible mechanism is an asymmetry in evidence generation. A provider can produce performance, robustness, privacy, and documentation artifacts inside its development pipeline, then prototype logging and oversight before deployment. But the external validity of oversight performance, impact assessment, and some traceability forms depends on deployer practice, organizational incentives, affected populations, and use context. Provider-generated evidence is therefore possible but less likely to suffice alone. This reading accords with Maas's [60] account of power relations among developers, users, and end users, and with work locating accountability in organizational routines rather than isolated model properties [8].

Article 43 does not imply universal pre-market self-assessment; its reach is narrower. For points 2–8 of Annex III to Regulation (EU) 2024/1689 (the EU AI Act), Article 43(2) uses Annex VI's internal-control procedure without routine notified-body examination. Biometric systems under point 1 of that Annex follow the conditional alternatives in Article 43(1), and product-embedded systems covered by Annex I Section A follow the applicable sectoral route under Article 43(3). This does not mean evidence is never independently examined. Market-surveillance, documentation-access, testing, complaint, explanation, and corrective routes remain available under Articles 74, 77, 79, 85, and 86. The defensible concern is temporal and institutional: for many stand-alone systems, routine pre-market conformity assessment may not add independent scrutiny of deployment-side evidence.

Harmonized standards may reinforce or counter this asymmetry depending on their content and development. Under Article 40, a standard creates a presumption of conformity only after its reference is published in the Official Journal, and only to the extent it covers relevant requirements. Harmonized standards intended to carry Articles 11, 12, and 14 are still being developed in CEN-CENELEC JTC 21, so the evidentiary vocabulary those requirements will attract is not yet settled. That cuts against reading the thin columns as a settled deficit: some of what looks absent may be evidence that the standards work has not yet had occasion to specify, and the deferral of Chapter III application to 2027 and 2028 removes the immediate demand signal for it. European standardization also provides formal channels for small and medium-sized enterprises, consumer organizations, environmental interests, and social stakeholders under Article 5 and Annex III of Regulation (EU) No 1025/2012 [61]. Formal access does not settle effective influence. Whether resources, technical drafting capacity, continuity of attendance, national-delegation structures, and agenda-setting power differ in practice remains an empirical question. The crosswalk gives reason to study whether these asymmetries shape the evidentiary vocabulary, but does not establish that they do.

One further boundary concerns the shared evidence language itself. By fixing what counts as evidence, it privileges measurable, auditor-legible harms over tacit, experiential, or culturally situated knowledge. The relevant limit is indeterminacy: a crosswalk cannot represent what its source instruments and categories fail to name. Its value is to expose the current vocabulary's boundaries, not to substitute technical legibility for affected people's accounts.

7. What Is Owed to the People the Evidence Is About

The preceding sections describe a distribution. This one argues that the distribution is wrong, and says on what grounds.

Under our coding, the requirements that ask chiefly for evidence about deployed processes are the thinly covered ones, and Table 1 records that for none of the three assurance demands does an affected person have a route to the underlying evidence. Providers hold the documentation. Deployers retain the logs. Authorities may compel either. A person subject to a consequential decision may complain under Article 85 and, if they meet the conditions of Article 86(1)-(3), may obtain an explanation of the system's role. They may not obtain the record that would let them test whether that explanation is true.

We take this to be a wrong done to that person, and not merely an inconvenience or a gap in an otherwise sound design. Two distinct claims support that, and they fail differently.

The first is procedural. A regime that determines something consequential about a person, and that forecloses any route by which that person can examine the basis of the determination, does not become fair by being accurate. Procedural justice research finds that people's judgments of legitimacy turn substantially on whether they had standing in a process, not only on whether they liked its outcome [62]. Binns makes the corresponding move for algorithmic decisions, locating accountability in whether a decision can be justified to those subject to it by reasons they could accept [63]; the access facts in Table 1 are what determine whether such a justification can be tested at all. The affected person here has no standing of that kind. They may be told what happened; they cannot interrogate it, and neither can anyone answerable to them. An explanation that cannot be checked against a record is a claim to be taken on trust. That is precisely the posture an assurance regime exists to avoid demanding of regulators, and it is worth asking why the same posture is acceptable when demanded of the governed.

The second is epistemic. Fricker's account of epistemic injustice describes a wrong done to someone in their capacity as a knower rather than as a subject of harm [64]. Neither of her two forms fits this case exactly, and it is worth being precise about why. The applicant is not disbelieved because of prejudice about who they are, which is testimonial injustice; nor do they lack the concepts to make sense of what happened to them, which is hermeneutical injustice. What they lack is reachable standing in the vocabulary in which claims about them are stated and contested. The exclusion is structural rather than identity-prejudicial. We use Fricker's framing because it names the category of wrong exactly, and we flag the difference in mechanism rather than let her account carry weight it was not built to carry here. Symons and Alvarado have already applied that framework to data-science technologies, arguing they diminish people in their standing as knowers [65]; what we add is narrower, tying the diminution to a specific evidentiary asymmetry that the mapping and Table 1 locate rather than to data science at large. The evidence regimes we examine are structured that way. The applicant in Section 1 supplies the data that the system acts on and is the subject of every claim the assurance case makes, yet the vocabulary in which those claims are stated and contested is available only to intermediaries. Their experience of the decision is not admissible evidence about it. This is what we mean in saying that an assurance regime producing evidence no affected person can reach distributes power as much as it demonstrates safety.

One further bound belongs here. The record-centred requirements closest to this argument, Articles 10, 11 and 13 and ISO/IEC 42001 A.7 and A.8, rest on decisive M8 and M9 cells that the 80-cell sample never reached (Section 3). Their only external check is the within-team cross-check, which we have called weaker evidence than the non-author exercise. The evidentiary basis for this part of the argument is therefore thinner than for the rest of the mapping, and we would rather say so than let the reader assume otherwise. The mapping does not prove either claim, and we do not present it as proof. Both are normative arguments, and their premises are legal and institutional facts about access rather than counts of coded cells. What the mapping contributes is narrower and, we think, still worth having: it shows where in a specific set of requirements the evidence that would answer a contestation

question is thinly represented, and Table 1 shows who could obtain it if it existed. The ethical argument then runs on those facts rather than on the coding.

Two objections deserve answers. The first is confidentiality: logs and model internals implicate trade secrets, third-party data, and security. We accept this, and the criterion proposed in Section 8 is scoped to authorized investigators rather than to public disclosure for exactly that reason. The wrong we identify is not that an individual lacks raw logs. It is that the Act itself creates no route that is both capable of examining the record and answerable to the person the record is about. Confidentiality explains why the first kind of access is limited; it does not explain the absence of the second.

We state that claim about the AI Act specifically, and not about the legal order as a whole. As Table 1 notes, the Act's silence does not exclude rights under data-protection, freedom-of-information, employment, sectoral, procedural, or national law, and Article 22(3) of the GDPR supplies rights to human intervention, to express a view, and to contest. Whether those routes reach the record in practice, and whether the bodies that operate them are answerable to the affected person in the sense used here, is an empirical and doctrinal question this article does not settle. If they do reach it, the wrong we describe is mitigated by instruments outside the assurance regime rather than absent from it, and the design point stands: a regime built to demonstrate trustworthiness should not depend on other law to make its own evidence reachable.

The second objection is that human oversight already supplies this. Article 14 requires that a person be able to oversee the system, and a deployer must assign someone competent under Article 26(2). But oversight assigned is not oversight demonstrated. Article 14 has one primary family in our mapping, and that family is immature by our own reading (Section 5). We do not rest the argument on the record-level count reported there, since its classification is not deposited and therefore cannot be contested in the way the rest of the mapping can. An arrangement that places a person at the point of failure without evidence that they can intervene does not discharge an obligation to the affected party; it relocates responsibility onto an operator who may be no better positioned than the applicant to contest the outcome. Elish's crumple zone is an ethical problem before it is an evidentiary one.

None of this establishes what those affected would themselves say they are owed. Work that elicits decision subjects' contestability needs directly [30] is the closest available check on the argument, and it matters here in a specific way: if what people ask for turns out to be an account they can question rather than the underlying record, the procedural claim above survives and the epistemic one needs restating. We cannot settle that from this study. No affected person was consulted here, and Section 9 treats that as its most consequential limit. The argument above is made on their behalf and not in their voice, which is a weaker thing, and we would rather state the position and be corrected than describe the arrangement and decline to judge it. That preference has the same shape as the arrangement we criticise: an invitation to be corrected is not a channel through which correction can reach us, just as an explanation under Article 86 is not a route to the record. We note the parallel rather than claim to have escaped it.

8. Implications for Policy and Practice

When an evidentiary vocabulary reflects what is convenient to supply, documentary attestation cannot automatically suffice for claims that depend on deployment practice. We offer four institutional design options in response to the reported coverage deficits and the AI Act's provider–deployer allocation of duties. They are not deductions from the crosswalk. The first is targeted independent assurance, which policymakers could consider for provider-side oversight and logging claims in specified high-risk uses. Any reclassification or new review duty would need legislative justification beyond this matrix. Article 27 remains a deployer duty of limited scope; external or competent-authority review must remain within that scope and be presented as reform, not current law.

The second option concerns the Commission's AI standardization request and resulting CEN-CENELEC JTC 21 work [66]. It should include validated approaches to oversight performance and traceability for consequential use. Oversight telemetry carries a hazard: once override rate or review latency becomes a compliance target, it risks measuring appearance rather than oversight and reproducing the ritual of verification criticized here. Standards should use telemetry as a diagnostic input to qualitative organizational audit, not a standalone threshold. Anti-gaming checks should include implausible operator profiles such as near-zero rejection or near-constant latency, randomized reliance drills [58], and qualitative review of rapid overrides. Telemetry is also worker surveillance. It turns the overseer into the measured object, and records may become disciplinary evidence against the person Elish [21] calls the crumple zone. Standards should specify worker notice, aggregate reporting, and limits on individual performance use. Any presumption of conformity under Article 40 should extend only to the cited standard's requirements and facets. Action-trajectory evaluation for agentic systems requires a separate evidence review before supporting a standardization proposal.

The third option, participation reform, can use existing channels. Options include sustained funding for consumer, worker, environmental, civil-society, and affected-community expertise; transparent reporting on committee composition and unresolved minority positions; accessible public-inquiry materials; and procedures recording how fundamental-rights concerns affect technical requirements. Every actor named there is an intermediary. Better-resourced civil-society standing in standardization is an improvement on the status quo, but it does not fill the affected-person column Table 1 records as empty; it funds a substitute for it, and the two should not be confused.

The fourth proposal is a context-sensitive reconstruction-adequacy criterion for systems whose outputs shape consequential decisions. Adequacy depends on the system. A record should allow an authorized investigator to relate a contested outcome to relevant inputs, system and model versions, material human interventions, and consequential actions. The criterion would not itself give an affected person access to confidential logs; its purpose is to prevent an unusable record from defeating investigation, complaint handling, or explanation.

These proposals carry material costs. Logging and reconstruction reach trade secrets and protected model internals, so the criterion is limited to authorized investigators rather than public disclosure and should operate through the Act's existing confidentiality protections (Article 78). Richer logs also create a security liability: they concentrate sensitive inputs and model-behavior traces in a durable store attractive to attackers. Retention limits, data minimization, and access control must be designed in. Administrative burdens are uneven too. Identical demands

weigh more heavily on small and medium-sized providers than incumbents; ignoring that difference can entrench firms best able to absorb compliance overhead. Proportionality, shared tooling, and standardized templates will determine whether the proposals narrow or widen the accountability gap.

Return to the Section 1 hiring example. For a high-risk AI system screening job applicants (Annex III, point 4 to Regulation (EU) 2024/1689, the EU AI Act), the Article 10 and 15 columns point to dataset-representativeness analysis, disaggregated accuracy and subgroup-fairness testing, and adversarial-robustness testing. These artifacts also populate the NIST MEASURE function and ISO/IEC 42001 controls A.6.2.4 and A.7. They show that the model was tested, not that an individual outcome can be investigated or the human review arrangement works. The Article 12 and 14 columns contain candidate evidence for those questions but prove nothing about any applicant's experience. An assurance case still requires a scoped claim, a metric with an a priori acceptance criterion, measured results with uncertainty, a residual-risk statement, and an accountable owner.

The crosswalk is a navigational aid, not an accountability mechanism. It tells an intermediary which evaluation may be relevant, then stops; it cannot decide whether evidence is sufficient. Routed through intermediaries, the proposals redistribute capacity among authorities, standards bodies, auditors, providers, and deployers; validation against technical criteria alone could turn them into further rituals of verification. Participatory research with affected people should therefore shape instrument design, not arrive only after deployment.

9. Limits and the Participatory Horizon

Several limits bound these claims. Screening was single-reviewer, the search English-only, the protocol not prospectively registered, and source quality not systematically appraised. Because the per-database yields were not retained, the evidence set cannot be reconstructed as an exhaustive search. The method taxonomy and socio-technical assurance lens, developed from the same materials, do not independently test one another. We make no absence claims about the literature. Where a column is thin, the claim is that the requirement asks for a kind of evidence these families do not produce.

The evidence pools are asymmetric and overlap has not been removed across pools. Privacy (M12) rests on six bespoke references; fairness (M2) draws on a 202-record reused supplement not assembled for this question. The three instruments are Global-North artifacts, and concepts such as fairness, harm, appropriate oversight, and individual contestation are culturally variable. Regulatory regimes elsewhere, such as China's 2022 Provisions on the Administration of Algorithmic Recommendation and its 2023 Interim Measures for generative AI services [67, 68], allocate evidence and contestation differently; the crosswalk would need re-derivation, not translation, to serve them.

One limit deserves separate treatment: no affected persons were consulted in this study. The distributional question and the three-demand typology are theorized from scholarship on contestability and due process [10, 11, 19, 20, 24], not elicited from people subject to high-risk automated decisions. The crosswalk therefore cannot establish what an

affected person experiences or which capacities they value; it can only generate propositions for participatory testing.

That study should recruit people subject to the kinds of systems used here as illustrations: hiring, credit, benefits, or immigration screening. Participants could revise or reject the Table 1 categories, identify capacities the crosswalk omits, and assess whether the proposed evidence would help an authorized intermediary investigate a contested outcome. Design frameworks for that engagement exist [29, 69]. Until then, the crosswalk should be described as an intermediary-facing research instrument, not a representation of affected people's priorities.

10. Conclusion

This synthesis identifies an uneven pattern in the technical evidence our crosswalk represents: artifact-centered and omnibus technical requirements attract multiple primary method families. Under the adjudicated coding, four columns have none and fourteen of the sixteen with one or two depend on a single primary cell. Article 14 depends on one emerging family, and Article 12 on one, M11, restored by the independent adjudication after our own rule had removed it (Table 6). The pattern is not a normalized measure of regulatory adequacy or proof of its institutional cause.

We interpret the pattern through three socio-technical assurance demands: oversight performance, evidentiary traceability for contestation, and deployment-impact governance. Article 12 does not itself require individual-decision reconstruction; reconstruction adequacy is our proposed criterion for supporting investigation, complaint, and scoped explanation. Article 27 and action-trajectory assurance both sit outside the coded matrix.

The contribution is an auditable evidence-navigation artifact and institutional hypothesis. The title's question receives a provisional answer: under the arrangements reviewed here, conformity evidence is produced by and circulates among intermediaries, and little of it is designed for use by the people subject to the decisions. That answer is an inference from Table 1's access analysis, not a coded result. The crosswalk helps intermediaries identify potentially relevant methods; testing and revising it with affected communities is the next step.

The hypothetical rejected applicant returns as a test for later work: can the evidence the crosswalk identifies help an authorized intermediary investigate the decision, and does that process protect the capacities affected people themselves identify? This study cannot answer either question. It establishes something narrower: a cell-level, auditable account of which technical evidence reaches each of the 22 requirements, where a verdict turns on a single contested code, and, through Table 1, who can obtain that evidence by default. The questions themselves remain for an assurance regime to confront.

Acknowledgements

We thank the two non-author raters who independently recoded the 80-cell sample, and the colleague who performed the within-team cross-check of all 264 cells.

Use of large language models. This project used language models for copy editing the work only.

Statements and Declarations

Data and code availability. The crosswalk, coding ledgers, search documentation, and analysis code are provided as Online Resource 2, which ships derived agreement tables sufficient to recompute every reported reliability statistic. The two non-author raters' cell-level codes are included with their written consent; their free-text comments are not published. The package documents which outputs it reproduces and which parts of evidence discovery cannot be reconstructed.

Ethics approval and consent. Ethics approval was not required for this work. The independent recoding exercise did not constitute human-subjects research: two expert raters applied a coding scheme to author-prepared descriptions of evaluation methods and regulatory requirements. The analytical dataset contained only anonymized professional judgments. Identifying and consent records were collected solely for consent, acknowledgement, and project administration, held separately from the analytical data, and not analyzed. Both non-author raters returned signed confirmation forms consenting to publication of their anonymized cell-level judgments; their free-text comments are withheld because neither form confirmed the accuracy of those comments.

Funding, competing interests, and author contributions. Entered in the journal's submission system rather than in this file, so that the reviewed manuscript carries no author-identifying information. Prior-work details are handled the same way.

References

1. National Institute of Standards and Technology (NIST) (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>
2. National Institute of Standards and Technology (NIST) (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, U.S. Department of Commerce.
<https://doi.org/10.6028/NIST.AI.600-1>
3. European Parliament and Council (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union, OJ L, 12.7.2024.
<https://data.europa.eu/eli/reg/2024/1689/oj>
4. International Organization for Standardization and International Electrotechnical Commission (ISO/IEC) (2023a) ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system.
<https://www.iso.org/standard/81230.html>
5. Hernandez J, Golpayegani D, Lewis D (2025) An open knowledge graph-based approach for mapping concepts and requirements between the EU AI Act and international standards. AI Ethics 5:4463–4474.
<https://doi.org/10.1007/s43681-025-00708-6>

6. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, Ye W, Zhang Y, Chang Y, Yu PS, Yang Q, Xie X (2024) A survey on evaluation of large language models. *ACM Trans Intell Syst Technol* 15(3):39. <https://doi.org/10.1145/3641289>
7. Veale M, Zuiderveen Borgesius F (2021) Demystifying the Draft EU Artificial Intelligence Act. *Comput Law Rev Int* 22(4):97-112. <https://doi.org/10.9785/crl-2021-220402>
8. Thomas C, Roberts H, Mökander J, Tsamados A, Taddeo M, Floridi L (2025) The case for a broader approach to AI assurance: addressing "hidden" harms in the development of artificial intelligence. *AI Soc* 40:1469–1484. <https://doi.org/10.1007/s00146-024-01950-y>
9. Laux J (2024) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI Soc* 39:2853–2866. <https://doi.org/10.1007/s00146-023-01777-z>
10. Almada M (2019) Human intervention in automated decision-making: toward the construction of contestable systems. In: 17th International Conference on Artificial Intelligence and Law (ICAIL), pp 2–11. <https://doi.org/10.1145/3322640.3326699>
11. Alfrink K, Keller I, Kortuem G, Doorn N (2023) Contestable AI by design: towards a framework. *Minds Mach* 33(4):613–639. <https://doi.org/10.1007/s11023-022-09611-z>
12. Power M (1997) *The audit society: rituals of verification*. Oxford University Press, Oxford
13. Goodman EP, Trehu J (2023) Algorithmic auditing: chasing AI accountability. *Santa Clara High Technol Law J* 39(3):289–332
14. Raji ID, Smart A, White RN, 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. In: *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, pp 33–44. <https://doi.org/10.1145/3351095.3372873>
15. Birhane A, Steed R, Ojewale V, Vecchione B, Raji ID (2024) AI auditing: the broken bus on the road to AI accountability. In: *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. <https://doi.org/10.48550/arXiv.2401.14462>
16. Manheim D, Martin S, Bailey M, Samin M, Greutzmacher R (2025) The necessity of AI audit standards boards. *AI Soc* 40(8):6609–6624. <https://doi.org/10.1007/s00146-025-02320-y>
17. Mökander J, Morley J, Taddeo M, Floridi L (2021) Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Sci Eng Ethics* 27(4):44. <https://doi.org/10.1007/s11948-021-00319-4>
18. Mökander J, Schuett J, Kirk HR, Floridi L (2024) Auditing large language models: a three-layered approach. *AI Ethics* 4(4):1085–1115. <https://doi.org/10.1007/s43681-023-00289-2>
19. Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc* 20(3):973–989. <https://doi.org/10.1177/1461444816676645>

20. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: ACM Conference on Fairness, Accountability, and Transparency (FAT*), pp 59–68. <https://doi.org/10.1145/3287560.3287598>
21. Elish MC (2019) Moral crumple zones: cautionary tales in human-robot interaction. *Engag Sci Technol Soc* 5:40–60. <https://doi.org/10.17351/ests2019.260>
22. Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Comput Law Secur Rev* 45:105681. <https://doi.org/10.1016/j.clsr.2022.105681>
23. Sterz S, Baum K, Biewer S, Hermanns H, Lauber-Rönsberg A, Meinel P, Langer M (2024) On the quest for effectiveness in human oversight: interdisciplinary perspectives. In: ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp 2495–2507. <https://doi.org/10.1145/3630106.3659051>
24. European Parliament and Council (2016) Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Off J Eur Union, OJ L* 119, 4.5.2016. <https://data.europa.eu/eli/reg/2016/679/oj>
25. European Parliament and Council (2026) Regulation (EU) 2026/1744 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). *Off J Eur Union, OJ L*, 24.7.2026. <https://data.europa.eu/eli/reg/2026/1744/oj>
26. International Organization for Standardization and International Electrotechnical Commission (ISO/IEC) (2023b) ISO/IEC 23894:2023 — Information technology — Artificial intelligence — Guidance on risk management. <https://www.iso.org/standard/77304.html>
27. International Organization for Standardization (ISO) (2026) Management system standards. <https://www.iso.org/management-system-standards.html>. Accessed 21 July 2026
28. Dixon-Woods M, Cavers D, Agarwal S, Annandale E, Arthur A, Harvey J, Hsu R, Katbamna S, Olsen R, Smith L, Riley R, Sutton AJ (2006) Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Med Res Methodol* 6:35. <https://doi.org/10.1186/1471-2288-6-35>
29. Sloane M, Moss E, Awomolo O, Forlano L (2022) Participation is not a design fix for machine learning. In: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO). <https://doi.org/10.1145/3551624.3555285>
30. Yurrita M, Verma H, Balayn A, Alfrink K, Gadiraju U, Bozzon A (2025) Identifying algorithmic decision subjects' needs for meaningful contestability. *Proc ACM Hum-Comput Interact* 9(CSCW). <https://doi.org/10.1145/3757415>
31. Authors (2025) Reference withheld to preserve double-anonymous review; the 202-record fairness corpus is deposited in Online Resource 2.
32. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46. <https://doi.org/10.1177/001316446002000104>
33. Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* 61(1):29–48. <https://doi.org/10.1348/000711006X126600>

34. Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 43(6):543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
35. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2021) Measuring massive multitask language understanding. In: *International Conference on Learning Representations*. <https://openreview.net/forum?id=d7KBjmI3GmQ>
36. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In: *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, pp 220–229. <https://doi.org/10.1145/3287560.3287596>
37. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems* 29, pp 3315–3323
38. Parrish A, Chen A, Nangia N, Padmakumar V, Phang J, Thompson J, Htut PM, Bowman SR (2022) BBQ: a hand-built bias benchmark for question answering. In: *Findings of the Association for Computational Linguistics (ACL)*, pp 2086–2105. <https://doi.org/10.18653/v1/2022.findings-acl.165>
39. Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. In: *8th Innovations in Theoretical Computer Science Conference, LIPIcs* 67:43. <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
40. Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5(2):153–163. <https://doi.org/10.1089/big.2016.0047>
41. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. In: *International Conference on Learning Representations*. <https://openreview.net/forum?id=rJzIBfZAb>
42. Cohen JM, Rosenfeld E, Kolter JZ (2019) Certified adversarial robustness via randomized smoothing. In: *International Conference on Machine Learning*. PMLR 97:1310–1320
43. Lin S, Hilton J, Evans O (2022) TruthfulQA: measuring how models mimic human falsehoods. In: *Annual Meeting of the Association for Computational Linguistics*, pp 3214–3252. <https://doi.org/10.18653/v1/2022.acl-long.229>
44. 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: *16th ACM Workshop on Artificial Intelligence and Security*, pp 79–90. <https://doi.org/10.1145/3605764.3623985>
45. Jain S, Wallace BC (2019) Attention is not explanation. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp 3543–3556. <https://doi.org/10.18653/v1/N19-1357>
46. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: *International Conference on Machine Learning*. PMLR 70:1321–1330
47. Angelopoulos AN, Bates S (2023) Conformal prediction: a gentle introduction. *Found Trends Mach Learn* 16(4):494–591. <https://doi.org/10.1561/22000000101>

48. Manakul P, Liusie A, Gales MJF (2023) SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 9004–9017. <https://doi.org/10.18653/v1/2023.emnlp-main.557>
49. Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy, pp 3–18. <https://doi.org/10.1109/SP.2017.41>
50. Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A, Raffel C (2021) Extracting training data from large language models. In: 30th USENIX Security Symposium, pp 2633–2650
51. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. In: ACM SIGSAC Conference on Computer and Communications Security, pp 308–318. <https://doi.org/10.1145/2976749.2978318>
52. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, Crawford K (2021) Datasheets for datasets. *Commun ACM* 64(12):86–92. <https://doi.org/10.1145/3458723>
53. Pushkarna M, Zaldivar A, Kjartansson O (2022) Data cards: purposeful and transparent dataset documentation for responsible AI. In: ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp 1776–1826. <https://doi.org/10.1145/3531146.3533231>
54. Richards J, Piorkowski D, Hind M, Houde S, Mojsilović A, Varshney KR (2021) A human-centered methodology for creating AI FactSheets. *IEEE Data Eng Bull* 44(4):47–58
55. Wan A, Klyman K, Kapoor S, Maslej N, Longpre S, Xiong B, Liang P, Bommasani R (2025) The 2025 Foundation Model Transparency Index. *SuperIntelligence–Robotics–Safety & Alignment* 2(4). <https://doi.org/10.70777/si.v2i4.17165>
56. Guo Z, Wu Y, Hartline JD, Hullman J (2024) A decision-theoretic framework for measuring AI reliance. In: ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp 221–236. <https://doi.org/10.1145/3630106.3658901>
57. Greco S, Vacchetti B, Apiletti D, Cerquitelli T (2025) Unsupervised concept drift detection from deep learning representations in real-time. *IEEE Trans Knowl Data Eng* 37(10):6232–6245. <https://doi.org/10.1109/TKDE.2025.3593123>
58. Hunter R, Moulange R, Bernardi J, Stein M (2024) Monitoring human dependence on AI systems with reliance drills. In: NeurIPS Workshop on Socially Responsible Language Modelling Research. <https://openreview.net/forum?id=LAttrv62x8t>
59. Romeo G, Conti D (2025) Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. *AI Soc*. <https://doi.org/10.1007/s00146-025-02422-7>
60. Maas J (2023) Machine learning and power relations. *AI Soc* 38:1493–1500. <https://doi.org/10.1007/s00146-022-01400-7>
61. European Parliament and Council (2012) Regulation (EU) No 1025/2012 on European standardisation. Off J Eur Union, OJ L 316, 14.11.2012. <https://data.europa.eu/eli/reg/2012/1025/oj>
62. Tyler TR (2006) Why people obey the law. Princeton University Press, Princeton

63. Binns R (2018) Algorithmic accountability and public reason. *Philos Technol* 31(4):543-556.
<https://doi.org/10.1007/s13347-017-0263-5>
64. Fricker M (2007) *Epistemic injustice: power and the ethics of knowing*. Oxford University Press, Oxford
65. Symons J, Alvarado R (2022) Epistemic injustice and data science technologies. *Synthese* 200:87.
<https://doi.org/10.1007/s11229-022-03631-z>
66. European Commission (2023) Commission Implementing Decision C(2023) 3215 final of 22 May 2023 on a standardisation request to CEN and CENELEC in support of Union policy on artificial intelligence (subsequently amended).
67. Cyberspace Administration of China (2022) Provisions on the Administration of Algorithmic Recommendation of Internet Information Services. Effective 1 March 2022.
68. Cyberspace Administration of China (2023) Interim Measures for the Administration of Generative Artificial Intelligence Services. Effective 15 August 2023.
69. Delgado F, Yang S, Madaio M, Yang Q (2023) The participatory turn in AI design: theoretical foundations and the current state of practice. In: *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. <https://doi.org/10.1145/3617694.3623261>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM2dataandcodeanonymized.zip](#)
- [ESM1technicalappendixblinded.pdf](#)