The Governance Gap Between AI Policy and Technical Implementation

Nishant Sandeep Bagde

B.Sc. Biotechnology, Vikas College of Arts, Science and Commerce, University of Mumbai

Jasreet Kaur

Undergraduate Researcher, Bachelor of Science in Analytics, Golden Gate University, USA

Dr. Smrite Goudhaman

Global Professor of Practice, Golden Gate University, USA

Abstract

Artificial intelligence (AI) governance has grown rapidly, producing regulation and standards specifying, in increasing detail, what organizations must document, test, log and review before deploying an AI system. Yet engineers, auditors and compliance staff report that these instruments describe obligations current practice cannot straightforwardly produce, or that can be satisfied on paper without changing what a system does. This article examines that governance gap — the difference, at a specific lifecycle point, between the artifact, capability or decision right a policy instrument requires as evidence and what the parties building and operating the system can actually produce, exercise or hold — a narrower, more falsifiable construct than compliance, accountability or legitimacy. Drawing on regulation, software-engineering, sociotechnical accountability and algorithmic-auditing scholarship, the article identifies six recurring translation points along the AI lifecycle — data provenance, risk classification, evaluation, logging, human oversight and incident reporting — and six observable dimensions (availability, completeness, accuracy, currency, access,

authority to act) for assessing closure. It develops a layered policy–organizational–technical model, positions it against prior auditing frameworks, proposes five testable propositions, and illustrates the model using the EU AI Act's obligations on a high-risk employment-screening system. The article concludes that closing the gap requires policy instruments to specify verifiable artifacts and decision rights rather than outcomes alone, treating compliance evidence as a first-class engineering deliverable.

Keywords: AI governance; algorithmic accountability; technical documentation; AI Act; sociotechnical systems; compliance verification

Introduction

Over the past five years, AI governance has shifted from broad ethical-principle statements (Jobin et al., 2019) toward instruments specifying what an organization must produce before an AI system can be trusted to operate. The EU AI Act sets out obligations for risk management, data governance, documentation, automatic logging, human oversight and post-market monitoring (European Parliament and Council of the European Union, 2024). NIST's Risk Management Framework organizes governance functions across the lifecycle (National Institute of Standards and Technology, 2023, 2024); ISO/IEC 42001 is a certifiable standard (International Organization for Standardization, 2023); and the Council of Europe (2024) and UNESCO (2021) tie AI governance to human rights and ethics.

These instruments share an unstated assumption: that a legal requirement maps onto a technical artifact an engineering team can produce and a reviewer can inspect. A requirement to log behavior assumes behavior can be logged meaningfully; to document training data assumes provenance can be traced; for human oversight assumes the reviewer has authority to intervene. This assumption fails more often than policy design anticipates (Costanza-Chock et al., 2022; Rakova et al., 2021; Sovrano et al., 2025; Widder & Nafus, 2023): compliance artifacts are produced but frequently incomplete, produced too late, or disconnected from the systems they describe.

This article treats that failure as a distinct governance problem, separate from the observation that ethical principles are too abstract to guide practice (Mittelstadt, 2019;

Morley et al., 2020). The instruments examined here ask for specific things — a named risk owner, a logged decision, a tested threshold — so the gap is not between values and action, but between what policy text specifies and what technical practice can reliably produce.

Defined precisely, the AI governance gap is the difference, at a specific lifecycle point, between the artifact, capability or decision right a policy instrument requires as evidence and what the parties actually building and operating the system can produce, exercise or hold when asked — narrower than several adjacent terms. Compliance is a claim that the gap is closed; the gap is what that claim should be checked against. Accountability concerns who bears responsibility once a gap is found (Section 4); organizational capacity makes closing it harder or easier (Section 5.8); institutional legitimacy depends on conditions well beyond any single artifact (Section 9). This article's term verifiable compliance names the artifact-level condition the model diagnoses: a point demonstrates it when the record, capability or authority a requirement calls for can actually be produced by someone positioned to do so.

The article makes four contributions: it connects regulatory scholarship with empirical evidence on how engineers build and govern AI systems (Sovrano et al., 2025); it reframes the gap as a structural property of a regulated object that is often modular and multi-party (Widder & Nafus, 2023); it identifies lifecycle-located divergence points; and it positions this model against the auditing and documentation frameworks it builds on (Section 3.7).

The article addresses two research questions. RQ1: Why do increasingly detailed AI policy requirements still fail to produce the practices they describe, and what would let such a requirement be verified as engineering practice rather than merely declared as compliance? RQ2: At which points in the AI lifecycle do policy requirements and technical practice most consistently diverge?

2. Conceptual Approach and Scope

This is a theory-synthesis conceptual article rather than an empirical study or systematic review, contributing by integrating literatures addressing the same problem from different disciplinary starting points and generating propositions later work can test (Jaakkola, 2020; Whetten, 1989). The analysis began with a tension: AI policy has become markedly more technical, yet practitioner and audit research continues to report that specified obligations are not reliably met.

Sources were selected purposively (Snyder, 2019), included where they analyzed a primary regulatory text examined here, reported empirical evidence about producing a policy-required artifact, or offered a sociotechnical account of oversight success or failure. Five bodies of work were connected: comparative regulatory analysis; empirical responsible-AI studies (Rakova et al., 2021; Sovrano et al., 2025; Widder & Nafus, 2023); sociotechnical critique of documentation; auditing research; and scholarship on AI supply chains and SME capacity (Cors & Thiébaut, 2025; Stampernas & Lambrinoudakis, 2025).

The unit of analysis is the compliance artifact: the record, test result or documented decision a requirement calls for and that practice must produce on request. A requirement is only as strong as the artifact it can generate. Some requirements presume an operational capability or a decision right instead. The model in Section 5 keeps the artifact as its primary unit, representing capability through Section 5.8's moderators and decision rights through the authority-to-act dimension (Section 5.1).

The synthesis proceeded in four steps: each source was read for the artifact a policy calls for and the point where it could not be produced; points were grouped by lifecycle location; candidates not separable by an operational test were merged; and each surviving group was restated as a translation point defined by the artifact a reviewer could ask to see, with capacity and jurisdictional overlap carried forward as moderators.

The article uses "AI system" broadly, covering generative and non-generative machine-learning systems (European Parliament and Council of the European Union, 2024; National Institute of Standards and Technology, 2023), and does not resolve frontier-safety or supply-chain labor questions, though these enter indirectly through the

moderators in Section 5.8. Its narrower question is institutional: what has to be true for a policy requirement to be verifiable as engineering practice?

3. Literature Review

3.1 The technical turn in AI policy instruments

Comparative analysis of early AI ethics guidance found convergence around transparency, fairness, accountability and privacy (Jobin et al., 2019), but left open how these would be verified. More recent instruments specify artifacts rather than only values: the EU AI Act requires a risk management system, data governance, documentation, logging, oversight and testing (European Parliament and Council of the European Union, 2024), though Veale and Zuiderveen Borgesius (2021) warn some provisions may prove ineffective. NIST's framework organizes governance functions across the lifecycle (National Institute of Standards and Technology, 2023, 2024); ISO/IEC 42001 adds certifiable requirements (International Organization for Standardization, 2023); the Council of Europe (2024) ties these to human rights.

3.2 A definitional mismatch upstream of implementation

Krafft et al. (2020) found policymakers and engineers define AI systematically differently: policy compares AI to human thought, while engineers center architectures and data. When the regulated object is not defined the way its builders define it, obligations attach unevenly. Selbst et al. (2019) describe an inverse failure, where technical framing strips out social context; Ananny and Crawford (2018) add that visibility is not the same as being able to act.

3.3 What engineers and compliance teams actually do

Rakova et al. (2021) found organizational culture, more than principle content, determines whether responsible-AI work changes practice; documentation is often treated as a cost under delivery pressure. Widder and Nafus (2023) found dislocated accountability: because AI systems pass through many hands, engineers commonly do not see compliance as their own responsibility. Sovrano et al. (2025) find current tools draft partial AI Act documentation but perform poorly on demanding sections. Metcalf et

al. (2021) add that impact assessments depend on methodological choices shaping which impacts are found at all.

3.4 Auditing, assurance and the limits of verification

A second failure point lies in mechanisms meant to check whether a required artifact exists and is accurate. Raji et al. (2020) present an internal audit framework tying documentation to development stages. Mökander et al. (2021, 2024) argue ethics-based auditing is constrained by auditor independence and limited access. Costanza-Chock et al. (2022) found auditors routinely lack standardized access, and legal risk discourages disclosure.

3.5 Documentation as an attempted bridge

Structured documentation formats translate technical characteristics into something a non-specialist can inspect: model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2021). Both act as boundary objects between the technical and policy layers, but their value depends on conditions the formats do not guarantee: mandatory completion, currency, and third-party verification (Sovrano et al., 2025).

3.6 Capacity, concentration and the distribution of technical compliance ability

The gap is not evenly distributed. Cihon et al. (2020) show fragmented governance risks multiplying compliance artifacts without improving any. Cors and Thiébaut (2025) argue SMEs face disproportionate compliance costs adopting a general-purpose model they did not build; Stampernas and Lambrinoudakis (2025) propose an SME framework still awaiting validation. Mohamed et al. (2020), Birhane (2020) and Sambasivan et al. (2021) argue governance and fairness concepts built in resource-rich contexts do not transpose elsewhere.

3.7 Positioning relative to existing auditing and documentation frameworks

The model developed in Section 5 builds directly on five frameworks already in wide use; Table 1 summarizes the comparison.

Raji et al.'s (2020) auditing framework ties checkpoints to an organization's own development process but is silent on artifacts an outside party must produce. Mökander et al.'s (2021, 2024) frameworks identify governance, model and application layers to audit, treated here as one scorable dimension (Section 5.1). Model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2021) supply a documentation template; this article specifies the conditions — mandatory completion, third-party verifiability, currency — under which a completed template closes a translation point, since Sovrano et al. (2025) show even a tool-assisted template can fail these conditions. NIST's framework organizes functions across the lifecycle but does not assess whether an output artifact is accurate and actionable; Section 5.1 supplies that. ISO/IEC 42001's clauses are distinguished from the artifact each produces and the capability a certified organization already has.

None of this displaces the five frameworks; each remains a primary source for what an organization should do. This article adds a way of asking, for any required output, whether the artifact, capability or decision right involved can actually be produced, exercised or held by the party the requirement was meant to reach — and if not, at which dimension and lifecycle point the failure sits.

Framework	Primary contribution	Unit of analysis	What this article adds
Raji et al. (2020) internal auditing framework	Ties audit checkpoints to development stages	Audit checkpoint / internal process step	Extends checkpoints to multi-party, cross-organizational lifecycles, tied to a producible artifact rather than a process stage
Mökander et al. (2021, 2024) ethics-based / three-layered auditing	Identifies governance, model, application audit layers	Audit layer / auditor capability	Locates where in the lifecycle — not only which audit layer — a requirement fails, and scores access and standing as explicit dimensions
Model cards (Mitchell et al., 2019) / datasheets (Gebru et al., 2021)	Standardized documentation formats	Document template	Specifies the conditions (mandatory completion, third-party verifiability, currency) under which a completed template closes a translation point

Framework	Primary contribution	Unit of analysis	What this article adds
NIST AI RMF (2023, 2024)	Organizes governance, mapping, measurement and management functions across the lifecycle	Organizational function	Adds an assessable status (closed / partially closed / open) to each function's output artifact, scored on six dimensions
ISO/IEC 42001 (2023)	Certifiable management-system requirements for continual risk assessment and monitoring	Management-system clause	Distinguishes the artifact a clause produces from the capability and authority it presumes a certified organization already has

Table 1. Positioning the model relative to existing accountability and auditing frameworks

Source: Author's own elaboration.

4. Research Gap

The literature reviewed supplies most ingredients needed to explain why AI policy requirements go unmet, but each strand explains only part of the pattern. Regulatory scholarship documents what instruments require; practitioner research documents what happens when organizations try to comply; auditing research documents why verification often cannot close the difference. Underdeveloped is an account connecting these directly to lifecycle points where requirement and artifact fail to meet.

Three specific gaps follow. A verifiability gap: policy enumerates obligations without asking whether existing methods can produce the named artifact at the stated grain; Sovrano et al. (2025) show AI-assisted tools cannot reliably produce complete Annex IV documentation. An attribution gap: practitioner research documents workarounds (Rakova et al., 2021; Widder & Nafus, 2023) but rarely traces these to drafting choices. A capacity gap: harmonization is treated as a legal problem among well-resourced actors; SME literature (Cors & Thiébaut, 2025; Stampernas & Lambrinoudakis, 2025) does not yet connect capacity to where it is tested.

The opportunity, then, is a specification of where, along an AI system's lifecycle, a policy requirement and a producible artifact must meet, and under what conditions that meeting succeeds or fails. Table 2 maps the literature problems above onto the three governance-gap dimensions developed here.

Literature problem	Governance-gap dimension	Core diagnostic question	Selected foundations
Requirements specify artifacts without testing producibility	Verifiability gap	Can current methods generate the named artifact at the stated grain?	Veale & Zuiderveen Borgesius (2021); Krafft et al. (2020)
Workarounds and dislocated responsibility not linked to drafting choices	Attribution gap	Which requirement produced this workaround, and could it be redrafted around a producible artifact?	Rakova et al. (2021); Widder & Nafus (2023); Metcalf et al. (2021)
Auditors lack access, standards or standing	Verifiability gap	Does the verifier have access and standing to compel remediation?	Costanza-Chock et al. (2022); Mökander et al. (2021, 2024)
Harmonization treated as a legal/technical problem among well-resourced actors	Capacity and distribution gap	Does the regulated party have staff, tooling and data access to produce this evidence?	Cihon et al. (2020); Cors & Thiébaut (2025); Mohamed et al. (2020); Sambasivan et al. (2021); Stampernas & Lambrinoudakis (2025)

Table 2. Literature problems and the dimensions of the AI governance gap

Source: Author's own elaboration.

None of the instruments in wide use is silent on evidence, but each treats it as a legal or technical mapping problem for well-resourced, single-provider systems rather than a lifecycle-located producibility question. This article links artifact producibility, accountable ownership and verification across the lifecycle. Table 2 also serves as a research-question crosswalk: RQ1 is addressed by Propositions 1–5 (Section 6); RQ2 by the translation points (Section 5, Table 4).

5. A Layered Model of the Policy-Implementation Gap

The governance gap is not a single, undifferentiated failure; it recurs at specific, locatable points where a policy requirement must be translated into a technical artifact. This section proposes a layered model connecting three levels: the policy layer of laws

and standards; the organizational layer of procurement, ownership and audit; and the technical layer of data pipelines, models and deployment infrastructure. Six recurring translation points concentrate this difficulty.

Because not every translation point reduces to a document, the model distinguishes three units: the compliance artifact a reviewer can inspect, the operational capability an organization must have, and the decision right to act on what the artifact shows. The dimensions below capture the artifact and access questions; authority-to-act captures decision rights; the moderators in Section 5.8 capture capability.

5.1 Six diagnostic dimensions of translation-point closure

To make failure assessable rather than a diffuse label, this article resolves each translation point's status along six observable dimensions, defined in Table 3 with an indicator and illustrative failure for each.

Dimension	Definition	Indicator	Illustrative failure
Availability	The artifact exists in recorded form — a log, template or documented decision.	The artifact can be retrieved on request without being newly created.	A foundation model's training-data documentation was never created (data provenance).
Completeness	The recorded artifact covers the full scope the requirement names.	The proportion of required fields, decisions or components actually addressed.	Logs capture the final decision but not the intermediate ranking (logging).
Accuracy	The content correctly describes the system, not aspirational text.	An independent re-derivation matches what is on record, within an agreed tolerance.	A model card states a use case never evaluated for the deployed population (evaluation).
Currency	The artifact reflects the system's present state, not an earlier version.	Elapsed time since last verified, relative to how often the system changes.	Documentation describes a version superseded by an unlogged update (data provenance / logging).
Access	The intended party can actually obtain the artifact.	A documented access channel exists and has been exercised, not only granted on paper.	An accurate internal audit log exists but is contractually inaccessible to the deployer (logging).
Authority to act	The party with access also holds standing to require a change.	At least one recorded instance where the artifact's contents changed a decision.	A reviewer sees a flagged case but has no standing to overturn it (human oversight).

Table 3. Six diagnostic dimensions of translation-point closure: definitions, indicators and illustrative failures
Source: Author's own elaboration.

A translation point is closed only where all six dimensions hold; partially closed where an artifact exists and is accessible but fails completeness, accuracy or currency; and open where no party can produce, obtain or act on it. The rule is conjunctive by design for the high-risk case this article develops (Section 7): a log that is complete and current but inaccessible, or unable to change a decision, serves no better than a log that does not exist.

The rule's strength is also its limitation. Treating the six dimensions as equally weighted and jointly necessary may be too rigid outside the high-risk tier, where a partial-closure profile may be an acceptable outcome. Whether the dimensions should be weighted, or the rule relaxed by risk tier, is an empirical question this model cannot settle (Section 11). The conjunctive rule fits the high-risk case developed here, not every AI governance requirement.

5.2 Translation point 1: Data provenance and documentation

Data governance provisions in the EU AI Act, ISO/IEC 42001 and the NIST framework call for documented, traceable data practices; datasheets and model cards illustrate a completed artifact (Gebru et al., 2021; Mitchell et al., 2019). The gap appears where provenance chains cross organizational boundaries: data later used to fine-tune a model often loses its documentation trail once it changes hands (Widder & Nafus, 2023).

5.3 Translation point 2: Risk classification and intended purpose

Risk-tiered regimes such as the EU AI Act classify obligations by a system's stated intended purpose. General-purpose and foundation models complicate this: built without a fixed deployment context, they are repurposed by parties other than their developer (Widder & Nafus, 2023), so classification presumes a purpose that, for many systems, does not exist until a downstream deployer decides how to use the model.

5.4 Translation point 3: Evaluation, testing and benchmarking

Policy instruments require testing sufficient to demonstrate accuracy, robustness or fairness without specifying a technical standard of sufficiency. Benchmarks developed in one context frequently do not transfer to another (Sambasivan et al., 2021), and abstraction traps can make a technically passing evaluation miss the conditions of actual use (Selbst et al., 2019).

5.5 Translation point 4: Logging and traceability

Automatic logging requirements assume a system's behavior can be captured in a stable, reviewable format. Continuously fine-tuned models and version sprawl across supply chains make full reproducibility difficult, consistent with the access limits the auditing literature documents (Costanza-Chock et al., 2022; Mökander et al., 2024).

5.6 Translation point 5: Human oversight and override

Human-oversight provisions require a person able to understand a system's output and intervene when necessary. Ananny and Crawford (2018) show visibility does not guarantee capacity to act, and Metcalf et al. (2021) show review design shapes what is found; the requirement is often satisfied by a human present without the information or authority to meaningfully intervene. This is the clearest case where the artifact — a documented review step — is distinct from the capability (Section 5.8) and decision right (Section 5.1) the requirement depends on.

5.7 Translation point 6: Incident reporting and post-market monitoring

Reporting obligations assume someone in the deployment chain can notice a serious incident and trace it to its cause. Widder and Nafus's (2023) dislocated accountability applies directly: where responsibility is distributed across many hands, incident detection can default to no one.

5.8 Cross-level dynamics and boundary conditions

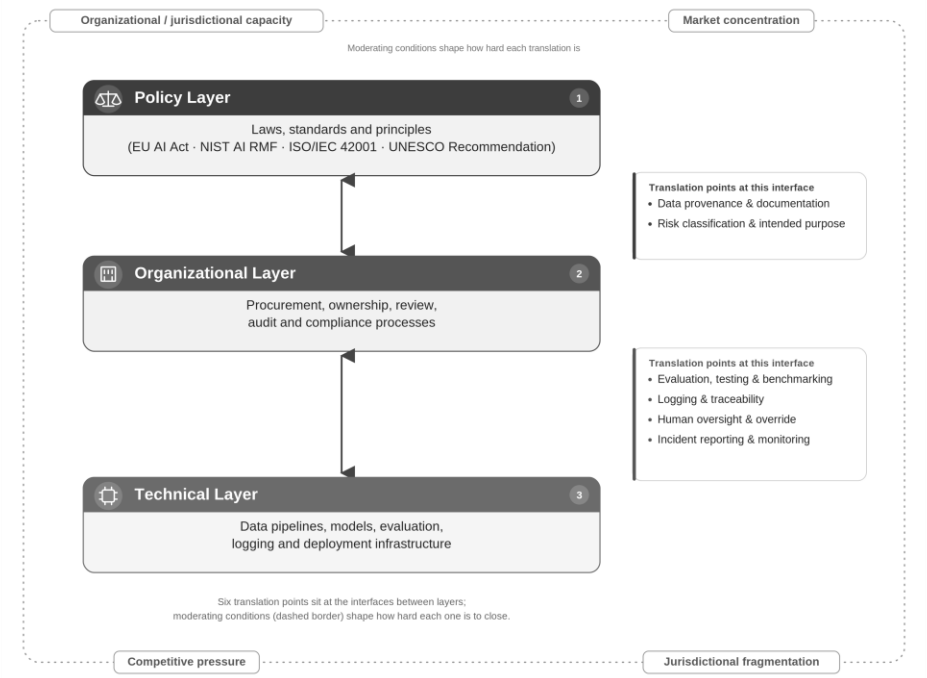


Figure 1. A layered model of the AI policy-implementation gap. Six translation points sit at the interfaces between the policy, organizational and technical layers; moderators shape how difficult each translation is.

Source: Author's own elaboration.

NIST's govern/map/measure/manage functions and ISO/IEC 42001's clause structure give organizations a repeatable process for producing artifacts (International Organization for Standardization, 2023; National Institute of Standards and Technology, 2023), but neither resolves what makes each translation hard. Concentration in foundation-model supply places compliance leverage upstream (Widder & Nafus, 2023);

competitive pressure shortens diligence; fragmentation multiplies artifacts (Cihon et al., 2020); and capacity determines whether an artifact can be produced at all (Cors & Thiébaut, 2025; Mohamed et al., 2020; Sambasivan et al., 2021; Stampernas & Lambrinoudakis, 2025). The gap is predicted widest where a high-risk use sits atop a concentrated supply chain under pressure with limited capacity.

5.9 From translation points to failure mechanisms

The six translation points fail for different reasons. Table 4 states, for each, the primary failure mechanism and what is lost when unresolved, since a fix appropriate to one will not resolve another.

Translation point	Primary failure mechanism	What is lost when unresolved
Data provenance & documentation	Broken provenance chains across organizational boundaries	Training-data claims cannot be checked against what was used
Risk classification & intended purpose	No fixed purpose exists at training time for general-purpose models	Obligations attach to the wrong party, or not until far downstream
Evaluation, testing & benchmarking	Sufficiency of testing is technically underspecified	A passing evaluation may not reflect deployment conditions
Logging & traceability	Non-determinism and version sprawl across supply chains	System behavior at a given time cannot be reconstructed
Human oversight & override	Oversight satisfied by presence, not capacity to intervene	A named reviewer exists but cannot meaningfully change an outcome
Incident reporting & monitoring	Dislocated responsibility across multi-party chains	Incidents go undetected, or are detected too late to trace

Table 4. Translation points, failure mechanisms and what is lost

Source: Author's own elaboration.

6. Conceptual Mechanisms and Propositions

The translation points are not independent; a failure at one often produces or masks a failure at another. Weak data documentation limits what evaluation can test; weak evaluation limits what logging can record; weak logging limits what an incident report can reconstruct. The propositions below state, in testable form, when the gap narrows or widens; each names the mechanism, the dimension it predicts will move, and when the prediction should not hold.

Proposition 1 (specificity). Requirements specified as measurable thresholds or named artifacts show higher closure than those specified as broad performance standards, since a vague standard leaves the choice of evidence to the regulated party while a named artifact removes that discretion (Krafft et al., 2020); the effect is strongest for availability and completeness. Boundary condition: this weakens where a named artifact becomes a target rather than a measure (Costanza-Chock et al., 2022).

Proposition 2 (modularity). Translation-point closure will be lower for systems assembled from externally sourced, general-purpose components than for systems built end-to-end by one organization. Widder and Nafus (2023) find engineers in modular supply chains commonly do not see compliance as their own responsibility, depressing access and authority-to-act specifically. Boundary condition: modularity is sufficient but not necessary for this failure.

Proposition 3 (verification). Independent verification narrows the gap only where the verifier holds both access to artifacts and standing to compel remediation; access without standing, or standing without access, leaves authority-to-act open even as access closes, so reported compliance stays decoupled from behavior (Costanza-Chock et al., 2022; Mökander et al., 2021). Access alone, without standing, is predicted to produce no improvement in authority-to-act closure — a testable interaction rather than an additive effect.

Proposition 4 (documentation). Documentation closes a translation point only when completion is a condition of deployment rather than voluntary, and the completing party can be independently verified. Model cards and datasheets make characteristics inspectable (Gebru et al., 2021; Mitchell et al., 2019), but neither mandates completion: Sovrano et al. (2025) show even tool-assisted drafting still needs expert verification. The dependent construct is completeness and accuracy closure specifically, not availability, which voluntary documentation can already satisfy.

Proposition 5 (capacity). Organizational and jurisdictional capacity moderates whether a clearly drafted requirement is translated into practice, independent of drafting quality — the central finding of the capacity literature (Cihon et al., 2020; Cors & Thiébaut, 2025; Mohamed et al., 2020; Sambasivan et al., 2021; Stampernas & Lambrinoudakis, 2025).

Boundary condition: capacity should matter most where Propositions 1–4's conditions are otherwise met, since that is the case in which remaining variation in closure is most plausibly attributable to capacity rather than to drafting or structural factors already accounted for.

These are relations for future research to test, not empirical findings. Read together, they suggest closure is least likely where an underspecified requirement meets a modular supply chain, verification with access but not standing, voluntary documentation and limited capacity simultaneously, and most likely where the reverse holds on every count.

Evidence matrix

Table 5 traces each cited source to the translation point, failure mechanism and proposition it supports, so a claim above can be checked against its specific source.

Source	Translation point(s)	Failure mechanism supported	Moderator	Proposition supported
Krafft et al. (2020)	Cross-cutting; risk classification	Definitional mismatch between policy and engineering	—	Proposition 1
Widder & Nafus (2023)	Data provenance; risk classification; logging; incident reporting	Broken provenance; no fixed purpose; version sprawl; dislocated responsibility	Supply-chain concentration	Proposition 2
Rakova et al. (2021)	Data provenance & documentation	Documentation treated as a cost, not linked to drafting choices	Organizational capacity	Proposition 5 (contextual)
Sovrano et al. (2025)	Data provenance & documentation	Tool-assisted drafting incomplete on demanding sections	—	Proposition 4
Metcalf et al. (2021)	Human oversight; evaluation	Review design shapes what is found	—	Proposition 4

Source	Translation point(s)	Failure mechanism supported	Moderator	Proposition supported
Raji et al. (2020)	Evaluation; incident reporting	Internal checkpoints not extended to multi-party access	—	Framework baseline, Section 3.7
Mökander et al. (2021, 2024)	Logging; incident reporting	Auditor independence, standards and access limits	Verification access/standing	Proposition 3
Costanza-Chock et al. (2022)	Logging; human oversight	Auditors lack standardized access; legal risk discourages disclosure	Verification access/standing	Propositions 1 and 3
Mitchell et al. (2019); Gebru et al. (2021)	Data provenance & documentation	Format completed but not mandated or verified	—	Proposition 4
Cihon et al. (2020)	Cross-cutting	Fragmentation multiplies artifacts without improving any	Jurisdictional fragmentation	Proposition 5
Cors & Thiébaut (2025)	Cross-cutting; data provenance; evaluation	Disproportionate compliance cost for SMEs adopting general-purpose models	Organizational capacity	Proposition 5
Stampernas & Lambrinoudakis (2025)	Cross-cutting	Lifecycle-mapped SME framework proposed, not field-validated	Organizational capacity	Proposition 5
Mohamed et al. (2020); Birhane (2020)	Cross-cutting	Imported governance reproduces dependency	Jurisdictional/resource capacity	Proposition 5
Sambasivan et al. (2021)	Evaluation & benchmarking	Benchmarks and fairness concepts do not transpose across contexts	Jurisdictional/resource capacity	Proposition 5
Ananny & Crawford (2018)	Human oversight & override	Visibility does not guarantee capacity to act	—	Proposition 3 (authority-to-act)

Source	Translation point(s)	Failure mechanism supported	Moderator	Proposition supported
Selbst et al. (2019)	Evaluation, testing & benchmarking	Abstraction traps miss conditions of actual use	—	Background (Section 3.2)

Table 5. Evidence matrix: sources mapped to translation points, failure mechanisms, moderators and propositions

Source: Author's own elaboration.

7. Illustrative Application: A High-Risk Employment-Screening System

This is a hypothetical illustration, not a report of an existing deployment. Consider a company deploying an AI candidate-screening tool built by fine-tuning a third-party foundation model on its own hiring data, with resume parsing from a separate vendor — falling within Annex III as high-risk (European Parliament and Council of the European Union, 2024). Articles 9–15 set out risk, data, documentation, logging, oversight and testing requirements; Articles 26 and 72 place monitoring obligations on deployer and provider; Article 73 governs incident reporting, with Annex III application set at 2 December 2027 (European Commission, 2026).

Table 6 applies each translation point from Section 5 to this case, naming the actor positioned to close the gap and the evidence that would demonstrate closure.

Translation point	What the Act requires	Where the gap typically appears	Responsible actor	Evidence that would close the gap
Data provenance & documentation	Documented data governance covering training and fine-tuning data	Foundation model's training data undisclosed; provenance documentable only from fine-tuning onward	Foundation-model provider; deploying company	A data-provenance record, contractually obtained from the provider, covering both stages, re-verified at each update
Risk classification & intended purpose	A defined intended purpose supporting high-risk classification	Model wasn't built for hiring; company must retroactively define the purpose	Deploying company, as the party giving the system its intended purpose	A dated, named-owner determination of intended purpose, revisited when the use case changes

Translation point	What the Act requires	Where the gap typically appears	Responsible actor	Evidence that would close the gap
Evaluation, testing & benchmarking	Testing for accuracy/robustness across the deployment population	Benchmarks for the model rarely reflect the company's actual applicant pool or languages	Deploying company	A population-specific evaluation report on the company's own applicant pool, repeated after each update
Logging & traceability	Automatic logs able to reconstruct a screening decision	Vendor-supplied parsing and periodic updates make end-to-end reconstruction difficult	Parsing vendor and model provider; deploying company	A versioned log covering parsing, scoring and ranking, accessible to the company on request
Human oversight & override	A human able to review and override recommendations	High volumes and limited visibility reduce oversight to formal sign-off	Deploying company	A record of reversed recommendations evidencing standing to intervene, not only sign-off
Incident reporting & monitoring	Monitoring and reporting of serious incidents	No single party among provider, vendor and company is positioned to detect a cross-cutting pattern	Deploying company	A documented monitoring process aggregating outcomes across the full pipeline

Table 6. Applying the layered model to a high-risk employment-screening system
Source: Author's own elaboration.

The illustration shows why neither prohibition nor unrestricted adoption serves candidates well. Before deployment, the table becomes a procurement specification — documentation rights, audit access and testing obligations negotiated while the company still has leverage. During operation, the same rows become a monitoring schedule: how many decisions can be reconstructed from logs, how often reviewers reverse a recommendation, how quickly testing follows a vendor update.

8. Implications for Organizational Practice

This article is a conceptual contribution rather than a report of an intervention conducted with an organization; a formal action-research design is left to Section 11. Applied diagnostically, the model suggests three steps.

Diagnosis: a team asks, for each translation point, what artifact it could produce today if a regulator asked. This commonly surfaces a mismatch between the point believed weakest — often logging or documentation, being most visible — and the one actually blocking compliance, frequently risk classification, since no one may have authority to state the system's intended purpose once it has been repurposed from a general-purpose model.

Planning and intervention: the team addresses the translation point actually weakest, designing changes to institutional practice, not only technology — a named owner for the classification decision, a contractual right to updated vendor documentation, a monitoring schedule tied to Table 4. Clarifying an intended purpose should, if it works, increase rather than decrease evaluation costs, since it surfaces deployment contexts requiring their own testing.

Review: the propositions in Section 6 supply review questions — has naming a specific artifact reduced reinterpretation (Proposition 1)? Has independent review changed a decision or only produced a report (Proposition 3)? Reflection should include reviewers and candidates whose concerns were not upheld, since a translation point that only works for successful challenges is not functioning as oversight.

Two cautions apply: coverage should not become a single reported score, since raising it can displace closing the gap; and the model diagnoses missing capacity, not whether an AI use should occur at all — that judgment remains political (Section 9).

9. Governance, Ethical and Practical Considerations

Selecting a response to a diagnosed gap requires specifying the decision at stake, who is affected, and an acceptable level of residual risk.

A concern cutting across every translation point is power asymmetry: providers control training data and update schedules, deployers control context of use, and auditors need access to both, though trade-secret concerns often restrict it (Costanza-Chock et al., 2022) — weighed against the standing regulators and courts require to verify a claim rather than merely record it.

Closing the gap creates real costs — documentation, logging and evaluation require staff, tooling and time, which matters most for smaller organizations and lower-resource jurisdictions (Cors & Thiébaut, 2025; Mohamed et al., 2020; Sambasivan et al., 2021; Stampernas & Lambrinoudakis, 2025). Shared infrastructure helps only where it builds local capacity rather than imposing costs calibrated elsewhere.

Finally, the model has a limited role in questions of value: evidence about verifiable compliance can inform how an AI system is used, but cannot settle whether a given use should be permitted at all. Those are questions of politics and law bearing on legitimacy in the fuller institutional sense (Introduction); the model makes a governance claim visible and contestable, but should not convert a value judgment into a technical score.

10. Limitations

This conceptual study has five limitations. The model has not been empirically tested; the propositions in Section 6 are testable relationships, not validated findings. Sources were selected purposively, so grey literature may be omitted. The six translation points overlap in practice, so the categories are analytical rather than cleanly separable. The model does not comprehensively treat frontier-model safety or environmental impact, entering only indirectly through the moderators in Section 5.8. And the conjunctive closure rule in Section 5.1 is untested against lower-risk requirements, where weighted scoring may be more appropriate.

11. Future Research

A first step is defining measurable indicators for each translation point: completeness and currency of documentation; the proportion of decisions reconstructable from logs; evaluation coverage across languages and populations actually served; the rate at which reviewers reverse an automated recommendation; the time between an incident and its report; and indicators of organizational technical capacity.

A second is comparative case-study research across sectors and organization sizes, building on SME-specific frameworks not yet field-tested (Stampernas & Lambrinoudakis, 2025; Cors & Thiébaut, 2025). A third is research with lower-resource-

jurisdiction institutions, testing whether harmonized standards build local capacity or entrench dependency. A fourth is action research applying the model across audit cycles. A fifth, prompted by Sovrano et al. (2025), evaluates AI-assisted compliance tooling. A sixth tests whether the six dimensions should be weighted rather than uniformly conjunctive.

The central empirical question this research program should keep in view is not whether an organization has adopted a named framework, but whether it can produce, on request, the technical evidence that framework says it should be able to produce.

12. Conclusion

AI policy has become substantially more technical, specifying documented risk assessments, dataset provenance, automatic logs, human oversight and incident reporting rather than only broad ethical commitments. Technical capacity has not kept pace, particularly for modular, continuously updated, cross-jurisdictional systems, and for smaller organizations least equipped to absorb the burden. Specificity in law is necessary but not sufficient, since a well-drafted requirement still fails wherever the artifact it presumes cannot be produced.

The contribution here is deliberately modest: the layered model does not replace the EU AI Act, NIST or ISO/IEC 42001; it asks whether a requirement can be satisfied by an artifact, capability or decision right current practice can generate, and, if not, where capacity must be built. It reduces to four questions: What would demonstrate this requirement has been met? Who can produce it? What happens when it cannot be? Who bears the cost of building that capacity? If it cannot answer these, its compliance may not be verifiable.

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

Birhane, A. (2020). Algorithmic colonization of Africa. SCRIPTed, 17(2), 389–409. https://doi.org/10.2966/scrip.170220.389

Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the future: Investigating architectures for international AI governance. Global Policy, 11(5), 545–556. https://doi.org/10.1111/1758-5899.12890

Cors, M. S., & Thiébaut, R. (2025). Artificial intelligence and the impact of the EU AI Act in business organizations. AI Magazine, 46, e70039. https://doi.org/10.1002/aaai.70039

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1571–1583). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533213

Council of Europe. (2024). Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

European Commission. (2026). Regulatory framework proposal on artificial intelligence: Implementation overview. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L, 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

International Organization for Standardization. (2023). Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023).

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. AMS Review, 10, 18–26. https://doi.org/10.1007/s13162-020-00161-0

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (pp. 72–78). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375835

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 735–746). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445935

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. Science and Engineering Ethics, 27, Article 44. https://doi.org/10.1007/s11948-021-00319-4

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. AI and Ethics, 4, 1085–1115. https://doi.org/10.1007/s43681-023-00289-2

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics, 26, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

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). 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

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 7, 1–23. https://doi.org/10.1145/3449081

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining algorithmic fairness in India and beyond. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 315–328). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445896

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the

Conference on Fairness, Accountability, and Transparency (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Sovrano, F., Hine, E., Anzolut, S., & Bacchelli, A. (2025). Simplifying software compliance: AI technologies in drafting technical documentation for the AI Act. Empirical Software Engineering, 30, Article 91. https://doi.org/10.1007/s10664-025-10645-x

Stampernas, S., & Lambrinoudakis, C. (2025). A framework for compliance with Regulation (EU) 2024/1689 for small and medium-sized enterprises. Journal of Cybersecurity and Privacy, 5(3), Article 40. https://doi.org/10.3390/jcp5030040

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU Artificial Intelligence Act. Computer Law Review International, 22(4), 97–112. https://doi.org/10.9785/cri-2021-220402

Whetten, D. A. (1989). What constitutes a theoretical contribution? Academy of Management Review, 14(4), 490–495. https://doi.org/10.5465/amr.1989.4308371

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. Big Data & Society, 10(1). https://doi.org/10.1177/20539517231177620