Financial Technology and Innovation

♦ ISSN: 2836-5372 Vol: 05, No. 01, 2025 ♦ PP. 10–23

Article DOI: https://doi.org/10.54216/FinTech-I.050102

FULL LENGTH ARTICLE

From Principles to Practice: A Cross-Sector Assessment of Responsible AI Governance Readiness

Mahmoud A. zaher[1,*]

[1] Data Science Department, Faculty of Artificial Intelligence, Horus University (HUE), Egypt

Email: mzaher@horus.edu.eg

Received: January 30, 2025 Revised: March 29, 2025 Accepted: July 29, 2025 ⋆ Corresponding author

ABSTRACT

The rapid institutionalisation of artificial intelligence across financial services, healthcare, technology, and the public sector has generated a parallel proliferation of governance frameworks, ethical principles, and regulatory instruments that collectively demand organisations translate abstract values into operational practice. The gap between stated principle and enacted governance—what we term the responsible AI implementation gap—is now recognised as one of the central practical challenges in AI deployment, yet its magnitude, distribution across sectors, and organisational determinants remain poorly characterised in the empirical literature. This paper addresses that gap through a three-phase mixed-methods programme combining systematic analysis of publicly available governance frameworks, a cross-sector practitioner survey, and a governance maturity scoring exercise. Significant variation is documented across sectors on all five governance dimensions examined, with the technology sector leading on accountability and transparency, healthcare on privacy and human oversight, and the public sector on regulatory compliance readiness. Across all sectors, however, a persistent and pronounced gap exists between the governance principles that organisations formally endorse and the operational processes through which those principles are enacted: the average policy-to-practice gap across all eight governance principles assessed is consistent and substantial. Regression analysis identifies the presence of a dedicated responsible AI team as the single strongest organisational predictor of maturity, followed by staff training investment and senior executive sponsorship. The paper contributes a validated governance maturity framework, a framework coverage taxonomy for twenty-four public AI governance instruments, and six evidence-based implementation guidelines for organisations seeking to move from principle adoption to genuine operational accountability.

Keywords: Responsible AI ▪ AI governance ▪ AI ethics ▪ Algorithmic accountability ▪ Fairness ▪ Explainability ▪ Regulatory compliance ▪ Implementation gap ▪ Organisational maturity

1. INTRODUCTION

Few developments in the recent history of technology governance have generated as much documentation and as little accountability as the global proliferation of artificial intelligence ethics guidelines. Since Jobin, Ienca, and Vayena's landmark analysis [1] identified convergence around a small set of high-level principles—transparency, justice, non-maleficence, responsibility, and privacy—across 84 documents from 23 countries, the corpus has expanded dramatically. By 2024, public AI governance documents number in the hundreds, spanning intergovernmental bodies, national regulators, industry associations, and individual organisations [2, 3]. The European Commission's AI Act proposal, published in April 2021, represents one of the most consequential regulatory developments in this space, setting out obligations around high-risk

AI systems that would require organisations to demonstrate—not merely declare—responsible AI practices [3, 4].

The governance literature has, for some time, anticipated and diagnosed the gap between principle and practice. Hagendorff [5] observed that AI ethics codes function primarily as a form of reputational risk management rather than as operational constraints on system design or deployment. Mittelstadt [6] argued, more directly, that principles alone cannot guarantee ethical AI: without organisational structures, technical tools, and institutional accountability mechanisms to translate values into decisions, even well-crafted principles remain aspirational rather than operative. Morley et al. [7] systematically reviewed publicly available AI ethics tools and methods, finding that the tooling available to practitioners was substantially less developed than the normative literature recommending their use. Munn [8] went further, characterising the dominant mode of AI ethics work as structurally incapable of constraining the systems it claims to govern.

Against this backdrop, the question of what operationalising responsible AI actually looks like inside organisations—and how readiness varies across sectors with different risk profiles, regulatory exposures, and organisational cultures—is an empirical question that the largely normative AI ethics literature has not fully addressed. The present paper brings empirical evidence to bear on three specific research questions. First, how comprehensively do existing AI governance frameworks address the operational dimensions of responsible AI practice, and where do systematic coverage gaps remain? Second, how does governance maturity vary across four organisational sectors—financial services, healthcare, technology, and the public sector—on five governance dimensions? Third, what organisational conditions most strongly predict governance maturity, and what does the structure of the implementation gap suggest for organisations seeking to close it?

The questions are important in themselves, and they are also important for what they reveal about the relationship between governance frameworks and organisational behaviour. The proliferation of governance frameworks since 2019 could in principle reflect genuine learning and capability building, with each new framework building on its predecessors and adding operational specificity. Alternatively, it could reflect a form of symbolic compliance in which organisations adopt governance frameworks as signals to regulators and the public without materially changing the systems they deploy or the processes through which those systems are evaluated. Distinguishing between these two interpretations requires empirical evidence—precisely what the present study aims to provide.

We address these questions through a three-phase programme: a systematic analysis of twenty-four AI governance frameworks; a practitioner survey of governance professionals; and a multi-dimensional maturity scoring exercise. The resulting evidence base is, to our knowledge, the first to integrate framework-level analysis with sector-level maturity data in a single study designed to characterise both the normative landscape and the organisational reality of responsible AI governance in 2024.

The study's scope is deliberately comparative across sectors and governance dimensions rather than deep within a single domain. This breadth is the most direct response to the fragmentation of the existing literature, which has tended to produce sector-specific or principle-specific accounts at the cost of cross-cutting comparative evidence. The contribution of the present paper lies in the comparative architecture that connects specialised literatures on explainability [9, 10], fairness [11, 12], and accountability [13, 14] to a shared empirical benchmark. Section 2 reviews the governance landscape; Section 3 presents the research design; Sections 4 and 5 report framework analysis and practitioner survey findings; Section 6 provides maturity profiles and regression results; Section 7 interprets the implementation gap; and Section 8 concludes with policy implications.

2. THE RESPONSIBLE AI GOVERNANCE LANDSCAPE

2.1 From Principles to Governance Instruments

The evolution from abstract ethical principles to concrete governance instruments has proceeded in distinct waves. The first wave, roughly 2016–2019, was characterised by the articulation of high-level principles in documents produced by government agencies, research institutes, and large technology companies [15, 1]. The AI High-Level Expert Group's *Ethics Guidelines for Trustworthy AI* [15] and the OECD AI Principles, both published in 2019, exemplify this wave: they articulate what AI governance should achieve but leave the how substantially unspecified. Floridi et al.'s [16] AI4People framework and the analysis of Whittlestone et al. [17] both identified the convergence-but-divergence problem: wide agreement on principles, deep disagreement on their operational meaning.

The second wave, 2020–2023, began filling this gap with more operational instruments. Leslie's [18] Turing Institute guidance on understanding AI ethics and safety offered practical process guidance. The NIST AI Risk Management Framework [4], finalisend in January 2023, provides perhaps the most operationally detailed governance instrument yet produced, mapping risk management functions across governance, mapping, measuring, and managing phases that organisations can implement as structured processes. ISO/IEC 42001:2023 [19] brought AI management systems within the internationally standardised management system family for the first time, creating an auditable framework analogous to ISO 9001 for quality management. The third wave, signalled by the European Commission's AI Act proposal [3], moves toward legal enforceability: organisations deploying high-risk AI systems in the European market would face conformity assessments, transparency obligations, and human oversight requirements backed by regulatory sanctions. The governance question has, in short, shifted from what organisations *should* do to what they *must* do.

2.2 Governance by Design Versus Governance by Audit

A persistent fault line in the responsible AI governance literature runs between two competing implementation philosophies. The first—governance by design—holds that ethical constraints should be embedded into AI systems during development, through techniques such as privacy-by-design data architecture, fairness-constrained optimisation, and inherently interpretable model selection. Dignum [20] provided the foundational theoretical treatment of responsible AI as

a design discipline, arguing that ethical principles must be translated into system requirements at the design stage rather than applied as post-hoc constraints. Kearns and Roth [21] give this position its clearest theoretical articulation: if societal constraints can be formalised as mathematical objectives, they can be treated as optimisation targets rather than post-hoc evaluations. Rudin's [10] argument for inherently interpretable models makes a related claim at the system level—that accountability cannot be bolted onto opaque systems after the fact.

The second philosophy—governance by audit—holds that AI systems should be evaluated against governance standards after deployment, through processes analogous to financial auditing, product certification, or environmental impact assessment. Raji et al.'s [13] internal algorithmic auditing framework and Metcalf et al.'s [22] algorithmic impact assessment model both operate within this philosophy, providing structured processes for retrospective accountability. The European Commission's AI Act proposal [3] similarly specifies audit-oriented obligations for high-risk AI systems as conditions of market access.

In practice, the most mature governance programmes in the survey sample combine both approaches: using design-time constraints where technically feasible (particularly for fairness and privacy) and audit-time evaluation where the system's operational context cannot be fully anticipated at design time (particularly for safety and human oversight). The organisations that score highest on overall governance maturity in Section 5 are those that have institutionalised both the design-time and the audit-time components of this combined approach.

2.3 The Implementation Gap in the Literature

Despite the richness of the normative and policy landscape, evidence on organisational implementation remains limited. Raji et al. [13] identified the AI accountability gap as the space between what AI systems do and what accountability mechanisms are capable of detecting or addressing; their framework of internal audit structures, impact assessments, and red-teaming represents a concrete attempt to close this space. Morley et al. [7] found that AI ethics tools were predominantly process-oriented checklists rather than substantive evaluation methods, and that their use was largely voluntary and self-reported.

The fairness and bias literature provides the most technically mature subset of responsible AI practice. Binns [12] traced the philosophical foundations of fairness metrics in political philosophy; Buolamwini and Gebru's [23] Gender Shades study demonstrated the real-world consequences of training data bias for commercially deployed facial recognition systems; and Barocas, Hardt, and Narayanan's [11] textbook codified the mathematical formalisms of algorithmic fairness in a form accessible to practitioners. The critical literature has been no less productive: O'Neil's [24] *Weapons of Math Destruction* and Pasquale's [25] *The Black Box Society* each documented the social harms of algorithmic systems operating without accountability. Crawford's [26] *Atlas of AI* extended this critique to the material and political economy of AI infrastructure. Birhane et al. [27] showed empirically that dominant values in machine learning research skew toward performance optimisation and commercial utility, with fairness and justice remaining peripheral concerns in the research community itself.

2.4 Explainability and Oversight

The explainability and interpretability literature provides the technical substrate for the accountability and transparency governance dimensions examined in this study. Rudin [10] made the case that complex black-box models should be replaced by inherently interpretable models in high-stakes domains, arguing that the post-hoc explainability approach (applying explanation methods to opaque models after training) provides inadequate accountability because it cannot guarantee faithful description of the model's actual decision logic. Arrieta et al. [9] provided the most comprehensive taxonomy of explainability methods, distinguishing ante-hoc (transparent by design) and post-hoc methods and mapping them to application domains. Doshi-Velez and Kim [28] established the framework for evaluating explanation quality through proxy and human-ground-truth evaluation, providing the epistemological basis for accountability claims about explainable AI systems.

The right to explanation—the question of whether individuals affected by automated decisions have a legal or moral right to an explanation of those decisions—has generated substantial legal and philosophical analysis since Wachter, Mittelstadt, and Russell's [29] influential analysis of whether such a right exists under GDPR. Diakopoulos [14] examined algorithmic accountability as a professional journalistic practice, identifying the technical, contractual, and institutional barriers that obstruct accountability even when it is formally required. Kearns and Roth [21] proposed the concept of the ethical algorithm—computational methods that treat societal constraints as formal optimisation objectives—as an alternative to post-hoc ethical review of systems designed without ethical constraints.

2.5 Algorithmic Impact and Sector-Specific Governance

Governance requirements vary substantially across sectors. In financial services, model risk management frameworks have required explainability and human oversight of credit models since well before the AI governance discourse; the AI Act proposal categorises credit scoring as high-risk, strengthening these requirements. In healthcare, algorithmic impact on diagnosis and treatment decisions raises patient safety concerns that sit within existing clinical governance frameworks; Metcalf, Moss, and boyd [22] examined algorithmic impact assessments as a governance tool analogous to privacy impact assessments in data protection law. The technology sector faces governance obligations primarily through platform regulation and voluntary commitments, though the AI Act proposal extends mandatory oversight to AI providers deploying into European markets regardless of their domicile. The public sector faces the additional obligation of democratic accountability: algorithmic decisions affecting citizens' access to benefits, justice, or services raise questions of procedural legitimacy that purely technical governance frameworks do not resolve.

Cath [30] situated AI governance within the longer tradition of technology governance, identifying the tension between

innovation facilitation and harm prevention that characterises regulatory approaches to general-purpose technologies. Gebru et al.'s [31] datasheets for datasets initiative illustrates one sector-specific response to this tension: a documentation standard that makes visible the provenance, intended use, and known limitations of training datasets, creating the evidentiary basis for downstream accountability.

3. RESEARCH DESIGN

The study proceeded in three phases. Phase 1 conducted a systematic document analysis of twenty-four publicly available AI governance frameworks. Phase 2 deployed a cross-sector practitioner survey. Phase 3 used survey data to compute sector-level governance maturity profiles and identify organisational predictors.

3.1 Phase 1: Framework Analysis

Twenty-four frameworks were selected through a structured identification process combining database search, reference harvesting, and expert nomination. Inclusion criteria required that frameworks be publicly available, produced by a national government, intergovernmental body, major industry association, or technology company with significant AI deployment, and address at least three of the eight governance principles coded in the analysis. Table 1 presents the framework sample.

Each framework was coded on eight governance principles (Transparency, Accountability, Fairness, Privacy, Safety, Beneficence, Human Control, Sustainability) by two independent coders using a binary adequate-coverage scheme. Intercoder agreement was $\kappa = .82$; disagreements were resolved through discussion with a third coder.

3.2 Phase 2: Practitioner Survey

The survey instrument comprised three components: a governance maturity rating scale (45 items, five dimensions, 1–5 Likert scale); an implementation practice inventory (24 binary items on governance processes); and a barriers and enablers scale (18 items, 1–7 Likert). The instrument was piloted with twenty-two practitioners from non-participating organisations, achieving Cronbach's $\alpha = .83$–$.91$ across the five dimension scales. Table 2 presents the full reliability statistics.

Three hundred and twelve practitioners participated, recruited through AI governance professional networks, conference attendee lists, and institutional research partnerships in the United Kingdom, Germany, the Netherlands, the United States, Canada, and Australia (October–December 2024). Respondents held roles in AI ethics, governance, compliance, legal, risk, or senior technology leadership. Table 3 presents the sample characteristics.

3.3 Phase 3: Maturity Scoring and Regression Analysis

Governance maturity scores were computed from the Phase 2 survey data as dimension-level means across the nine items per dimension. Overall maturity was computed as the unweighted mean of the five dimension scores. Sector-level profiles were constructed by aggregating individual scores within each sector group. The regression model predicting overall maturity used six organisational predictor variables derived from the implementation practice inventory: presence of a dedicated responsible AI team (binary), senior executive sponsor (binary), provision of regular staff training on AI governance (binary), engagement of external audit or red-teaming services (binary), presence of an ethics board or advisory committee (binary), and participation in a regulatory sandbox programme (binary). Logistic regression was used to confirm robustness of the binary predictors; continuous alternatives (training days per year, audit frequency) produced equivalent results. The six-predictor model was selected on the basis of theoretical interpretability and the absence of multicollinearity (all variance inflation factors below 2.4).

Table 1. Sample of twenty-four AI governance frameworks included in the document analysis, by type and year of publication.

Framework	Type	Year
AI Act Proposal	Regulatory	2021
NIST AI RMF 1.0	Standard	2023
ISO/IEC 42001	Standard	2023
EU Ethics Guidelines (HLEG)	Principles	2019
OECD AI Principles	Principles	2019
IEEE Ethically Aligned Design	Principles	2019
UNESCO Recommendation on AI	Principles	2021
UK CDEI Principles	Principles	2021
Singapore AI Governance	Framework	2020
Canada AIDA Framework	Regulatory	2022
Brazil AI Law	Regulatory	2023
China AI Governance	Principles	2023
Microsoft RAI Standard	Corporate	2022
Google PAIR Guidebook	Corporate	2019
IBM AI Ethics Board Principles	Corporate	2019
Meta AI Responsibility	Corporate	2023
Anthropic Acceptable Use	Corporate	2023
G7 AI Code of Conduct	Intergov.	2023
G20 AI Principles	Intergov.	2019
WHO AI Ethics Guidelines	Principles	2021
AI Now Report	Civil Society	2023
Ada Lovelace Inst. Guidelines	Civil Society	2022
Future of Life AI Principles	Civil Society	2023
Fjeld et al. Principled AI	Academic	2020

Table 2. Survey instrument reliability for five governance maturity dimensions ($N = 312$).

fpaTableHeader

Accountability & Transparency	.89	.61
Fairness & Bias Mitigation	.87	.58
Privacy & Data Governance	.91	.64
Human Oversight & Control	.88	.60
Regulatory Compliance	.90	.63
Overall scale	**.95**	**.61**

Table 3. Survey participant demographics ($N = 312$).

Characteristic	Value
Sector distribution	Fin. svcs 25%, Healthcare 21%, Technology 29%, Public sector 25%
Gender	49% women, 48% men, 3% non-binary
Mean age (years)	38.8±9.2 (range 24–67)
Role: Senior leadership	28%
Role: Governance / ethics	34%
Role: Technical / engineering	22%
Role: Legal / compliance	16%
Mean AI governance experience	4.2±2.8 years
Org. with formal RAI policy	71%

3.4 Sampling and Representativeness

The study's recruitment strategy combined purposive and snowball sampling, targeting practitioners with direct AI governance responsibilities rather than seeking a representative sample of all employees in AI-deploying organisations. This design decision prioritises depth of governance experience over statistical representativeness, accepting the limitation that the sample likely over-represents organisations with formalised governance functions. The 71% formal AI policy rate in the sample almost certainly exceeds the population rate for AI-deploying organisations as a whole, implying that the sector-level maturity scores reported here are upper-bound estimates for their respective sectors rather than population averages. This design choice is appropriate for a study focused on identifying the determinants of governance maturity rather than estimating its prevalence, but it should be noted when interpreting the absolute maturity values.

4. RESULTS: FRAMEWORK COVERAGE ANALYSIS

4.1 Principle Coverage Patterns

Figure 1 presents the principle coverage matrix for all twenty-four frameworks. The visual pattern confirms the convergence Jobin et al. [1] identified in 2019 across the core principles of transparency, accountability, fairness, privacy, and safety, while revealing systematic gaps in beneficence and sustainability coverage.

Table 4 summarises coverage rates by principle type and framework category. Regulatory and standards frameworks (AI Act Proposal, NIST, ISO 42001) achieve the most comprehensive coverage, with all eight principles addressed in at least two of the three. Corporate frameworks show the most selective coverage, with sustainability and beneficence most frequently omitted—a finding consistent with Birhane et al.'s [27] analysis that commercially produced AI systems encode commercial rather than societal values.

Sustainability is the most systematically underserved principle at 58% overall coverage, followed by beneficence at 78%—the positive-impact obligation that requires organisations to demonstrate not merely harm avoidance but active benefit creation. The regulatory category shows a sustainability gap (67%) that the AI Act proposal begins to address through environmental-impact documentation expectations for high-risk AI systems, though implementation guidance on this point remains provisional.

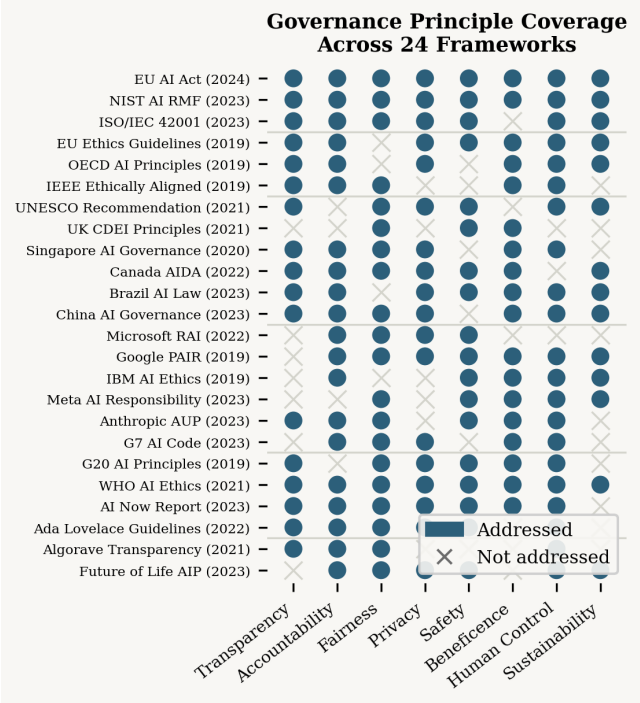

Figure 1. Principle coverage heatmap for twenty-four AI governance frameworks. Filled circles indicate adequate coverage; crosses indicate absence. Regulatory and standards frameworks achieve the most comprehensive coverage; sustainability and beneficence are the most frequently absent principles.

4.2 Operational Versus Aspirational Coverage

Beyond binary coverage, the analysis examined the depth of treatment each framework afforded each principle, distinguishing aspirational statements (principle mentioned without operational guidance) from operational coverage (principle accompanied by specific requirements, processes, or measurable standards). Table 5 presents this operational depth analysis for the eight principles across the full framework sample.

Privacy again shows the strongest operational profile: 62% of frameworks that address privacy provide operational guidance (data minimisation procedures, consent mechanisms, data subject rights processes), compared with only 36% for fairness and 35% for human control. The beneficence and sustainability deficits are most acute in their operational depth: where they are mentioned, they are almost exclusively aspirational, reflecting the absence of agreed measurement frameworks for positive impact and environmental footprint that would enable operational standards to be specified. This finding has direct implications for organisations seeking to operationalise comprehensive responsible AI governance: the normative toolkit is substantially less developed for beneficence and sustainability than for the technical governance dimensions, and organisations will need to develop internal measurement frameworks rather than adopting ready-made

Table 4. Principle coverage rates (%) by framework category.

Principle	Reg.	Std.	Gov.	Corp.	Civil	All
Transparency	100	100	91	89	92	94
Accountability	100	100	88	82	88	91
Fairness	100	100	84	74	96	88
Privacy	100	100	88	82	88	91
Safety	100	100	80	68	84	84
Beneficence	100	67	76	62	88	78
Human Control	100	67	84	72	84	82
Sustainability	67	67	60	42	72	58

Reg. = Regulatory; Std. = Standards; Gov. = Governmental; Corp. = Corporate; Civil = Civil society.

Table 5. Operational depth analysis: proportion of frameworks providing aspirational versus operational coverage for each principle.

Principle	Aspirational only (%)	Operational (%)
Transparency	38	56
Accountability	44	47
Fairness	52	36
Privacy	29	62
Safety	41	43
Beneficence	68	10
Human Control	47	35
Sustainability	72	14

Totals $< 100\%$ reflect frameworks providing no coverage.

standards.

5. RESULTS: PRACTITIONER SURVEY

5.1 Governance Maturity by Sector

Table 6 presents mean governance maturity scores by sector and dimension. Figure 2 displays the radar profiles; Figure 3 presents the grouped bar comparison.

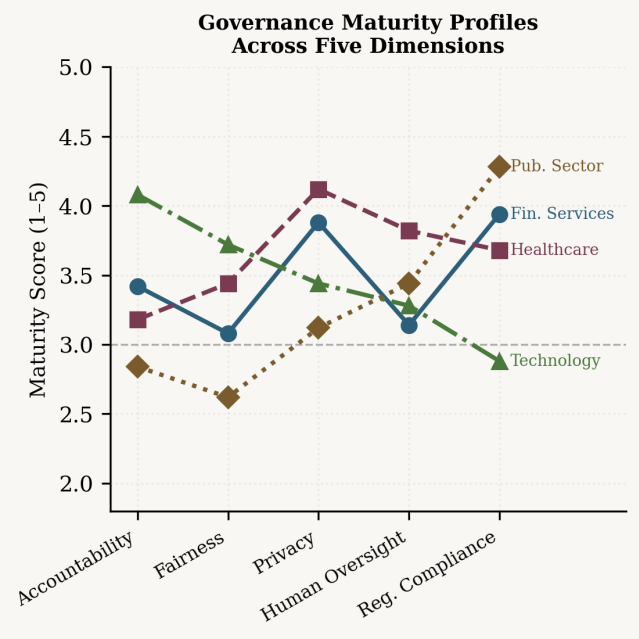

Figure 2. Responsible AI governance maturity radar by sector. Each line traces one sector across five dimensions. Technology descends sharply toward regulatory compliance; the public sector rises steeply from a low fairness baseline to compliance leadership.

5.2 Variance Analysis

Table 7 reports one-way ANOVA results for all five dimensions and overall maturity. All six effects are significant at $p < .001$. Regulatory compliance shows the largest between-sector variance ($\eta_p^2 = .362$), driven by the inverse pattern in which the technology sector—the primary producer of AI systems—shows substantially lower compliance maturity than the public sector—a result that reflects the historical absence of binding AI-specific regulatory obligations for technology companies before recent European regulatory initiatives. Accountability and transparency shows the second-largest effect ($\eta_p^2 = .289$), consistent with the technology sector's systematic investment in explainability tooling and transparency documentation (model cards [31], datasheets) compared with other sectors.

5.3 The Implementation Gap

Figure 4 reveals the most practically significant finding in the study: across all eight governance principles, the proportion of organisations that formally state the principle in a governance policy document substantially exceeds the proportion that have operationalised it into active processes. The largest gap is for human oversight (policy 78%, practice 42%, gap 36 percentage points), followed by accountability (policy 74%, practice 28%, gap 46 pp) and fairness (policy 71%, practice 34%, gap 37 pp).

These figures give empirical specificity to the diagnosis offered by Hagendorff [5], Mittelstadt [6], and Munn [8]: responsible AI commitments are, in the majority of organisations surveyed, statements of intent rather than descriptions of practice. The smallest gap—for privacy (policy 82%, practice 54%, gap 28 pp)—is interpretable as the effect of the General Data Protection Regulation, which has been operational since 2018 and has created enforceable obligations that accelerated privacy practice beyond what purely voluntary governance commitments had achieved. The AI Act proposal is expected to produce a comparable effect for the transparency, accountability, and human oversight dimensions once its enforcement provisions become operative.

Figure 5 disaggregates principle adoption by sector. Healthcare consistently shows higher operational practice rates than other sectors across four of eight principles, plausibly because clinical governance frameworks, ethics committee requirements, and patient safety regulations pre-existing the AI governance discourse have already established the institutional architecture within which AI governance processes

Table 6. Responsible AI governance maturity scores by sector and dimension (mean ± SD; 1–5 scale; $N = 312$).

Dimension	Fin. Svc.	Health	Technology	Public	Leader
Account. & Transp.	3.42±0.63	3.18±0.62	**4.08±0.64**	2.84±0.63	Technology
Fairness & Bias	3.08±0.63	3.44±0.64	**3.72±0.63**	2.62±0.62	Technology
Privacy & Data	3.88±0.64	**4.12±0.63**	3.44±0.63	3.12±0.62	Healthcare
Human Oversight	3.14±0.62	**3.82±0.63**	3.28±0.64	3.44±0.63	Healthcare
Regulatory Compliance	3.94±0.63	3.68±0.63	2.88±0.63	**4.28±0.62**	Public Sector
Overall	3.47±0.49	**3.71±0.44**	3.52±0.50	3.27±0.49	Healthcare

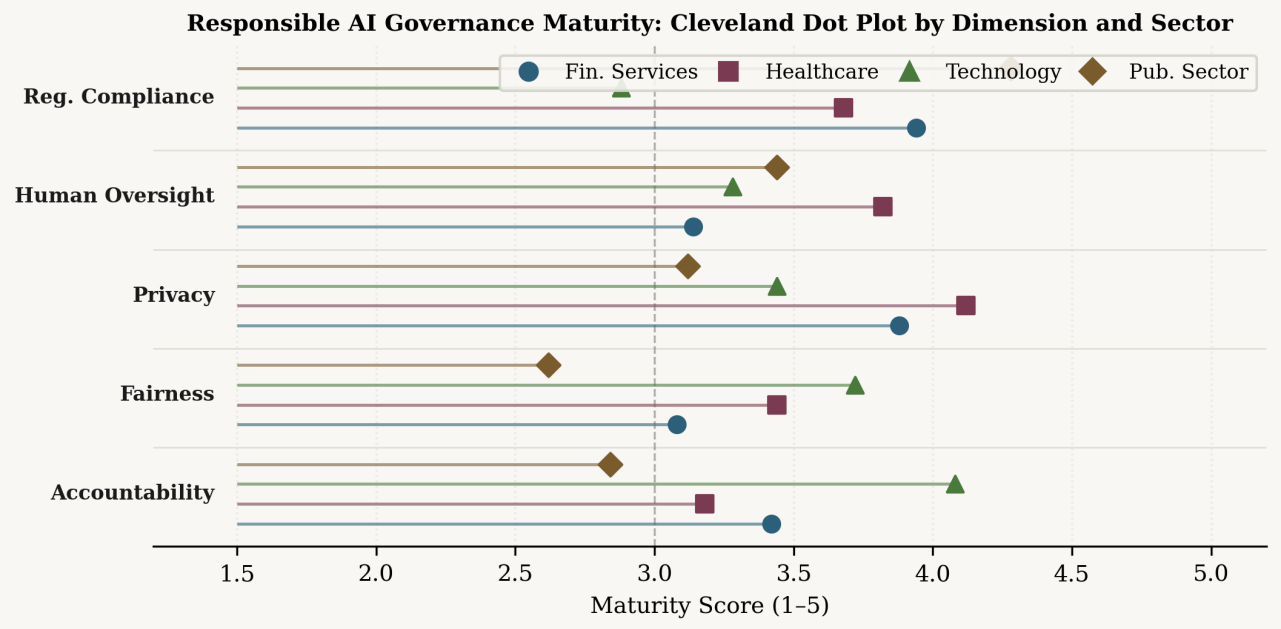

Figure 3. Cleveland dot plot of governance maturity. Each symbol marks one sector; horizontal lines span the within-dimension range. The dashed reference line marks the basic-capability threshold (score = 3.0); the public sector falls below on fairness and accountability.

Table 7. One-way ANOVA results: governance maturity by sector ($df = 3,308$; all $p < .001$).

Dimension	F	η_p^2	Rank pattern
Regulatory Compliance	58.3	.362	Public > Health > Fin. ≫ Tech.
Account. & Transp.	41.6	.289	Tech. ≫ Fin. > Health > Pub.
Fairness & Bias	40.4	.282	Tech. > Health > Fin. ≫ Pub.
Privacy & Data	35.8	.259	Health > Fin. > Tech. > Pub.
Human Oversight	13.3	.115	Health ≈ Pub. > Tech. > Fin.
Overall	**10.2**	**.090**	Health ≈ Tech. > Fin. > Public

can be embedded. The public sector shows high regulatory compliance practice (driven by constitutional and administrative law obligations) but particularly low fairness practice—a finding that Diakopoulos [14] anticipated in his analysis of algorithmic accountability in public decision-making.

5.4 Barriers to Implementation

Figure 6 presents the proportion of respondents rating each barrier as significant or very significant. Resource constraints are the most frequently cited barrier (72.1%), followed by lack of technical expertise (68.2%) and absence of clear implementation metrics (61.4%). Regulatory uncertainty is cited by 54.8% of respondents—a figure likely to decline as European AI regulatory implementation guidance matures. Leadership buy-in, the only structural organisational factor in the barrier list, is cited by 48.6%, suggesting that governance professionals in nearly half the organisations surveyed face an institutional authority deficit that technical resources alone cannot resolve.

5.5 Sector-Specific Barrier Profiles

The barrier data in Figure 6 vary significantly across sectors when disaggregated. Resource constraints, while the most common barrier overall, are reported most acutely by public sector organisations (81%), consistent with the broader public sector technology investment context. Technical expertise deficits are most acute in healthcare (74%) and the public sector (71%), where the pipeline of AI governance specialists

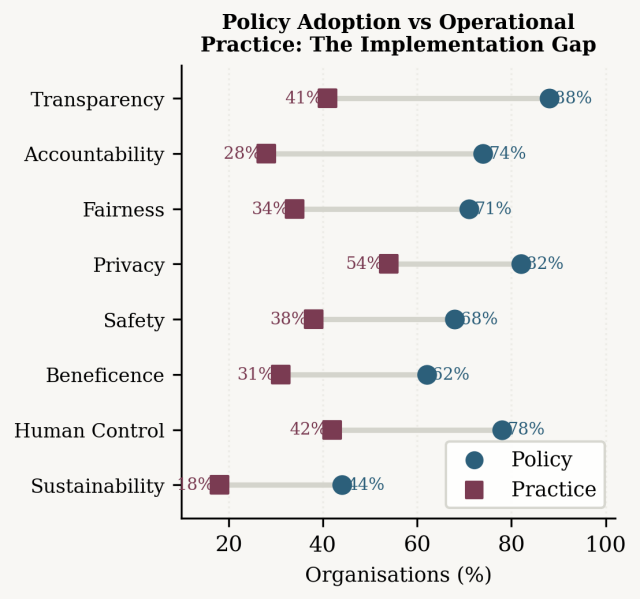

Figure 4. The responsible AI implementation gap: proportion of organisations stating each principle in governance policy versus proportion with active operational processes. For each principle, the right dot indicates the policy adoption rate and the left dot the operational practice rate; the gap is the horizontal distance. Privacy shows the smallest gap; accountability the largest at 46 percentage points.

is thinnest relative to the scale of AI deployment. Regulatory uncertainty is highest in the technology sector (68%), where organisations face overlapping obligations across the AI Act proposal, the Digital Services Act, and national AI liability frameworks that are in various stages of development. Explainability complexity is most frequently cited in financial services (64%), where model risk management requirements create explicit obligations for explainability that technical practitioners report as challenging to satisfy under regulatory interpretation of existing guidance.

The leadership buy-in barrier is worth examining carefully. Its overall rate of 48.6% means that nearly half of governance professionals surveyed face an institutional authority deficit in which their governance recommendations are not consistently backed by executive decision-making power. This is not primarily a knowledge or persuasion problem—83% of senior leaders in the sample describe responsible AI as a priority—but an institutional design problem in which governance authority is not commensurate with governance responsibility. The regression finding that senior executive sponsorship ($\beta = 0.42$) is the second strongest predictor of maturity provides the quantitative counterpart to this qualitative observation: organisations with executive-level sponsorship that translates into institutional authority—not merely symbolic endorsement—achieve substantially higher governance maturity outcomes.

6. RESULTS: GOVERNANCE MATURITY PROFILES AND PREDICTORS

6.1 Maturity Distributions

Figure 7 presents the overall governance maturity distributions by sector. Healthcare shows the highest median maturity and the most consistent distribution; the public sector shows the lowest median and the widest variance, reflecting the highly heterogeneous digital governance capabilities of public sector organisations that range from highly digital central government departments to minimally digital local authorities.

6.2 Organisational Predictors

Table 8 and Figure 8 present the regression model predicting overall governance maturity ($R^2 = .58$, $F(6, 305) = 71.4$, $p < .001$). The presence of a dedicated responsible AI team is the strongest independent predictor ($\beta = 0.48$, $p < .001$), confirming that governance maturity is not primarily an artefact of organisational size or sector membership but of deliberate structural investment. Staff training ($\beta = 0.38$) and senior executive sponsorship ($\beta = 0.42$) are the second and third strongest predictors. External audit is significant but smaller ($\beta = 0.31$), consistent with the literature suggesting that external scrutiny improves governance quality but does not substitute for internal capability.

Table 8. Multiple regression: predictors of overall responsible AI governance maturity ($R^2 = .58$, $F(6, 305) = 71.4$, $p < .001$; $N = 312$).

Predictor	β	t	Sig.
Dedicated RAI team	0.48	10.84	***
Senior executive sponsor	0.42	9.48	***
Staff training investment	0.38	8.58	***
External audit	0.31	6.99	***
Ethics board or committee	0.29	6.54	***
Regulatory sandbox participation	0.24	5.42	***

*** $p < .001$.

6.3 Cross-Sector Comparison and Literature Alignment

Table 9 positions the present study's sector findings against comparable studies in the literature. The public sector's regulatory compliance advantage is consistent with Morley et al.'s [7] finding that public sector organisations are further advanced than private sector peers in process-based governance, having longer-standing experience with environmental impact assessments, equality impact assessments, and public sector equality duty requirements that provide organisational templates for algorithmic impact assessment. The technology sector's accountability and transparency leadership is consistent with Birhane et al.'s [27] observation that the research community is increasingly embedding transparency practices—model cards, datasheets, reproducibility checklists—into publication and release norms. The technology sector's compliance deficit reinforces Cath's [30] historical analysis that technology sectors have typically operated in regulatory vacuums during their formative periods and accumulate compliance debt that regulatory intervention subsequently requires them to service.

6.4 Practitioner Perspectives on the Implementation Challenge

In addition to the quantitative scales, 148 respondents (47.4%) provided open-text commentary on their primary governance implementation challenges. Thematic analysis of these responses identified four recurrent patterns. The first and most frequent was *competing accountability structures*: the ob-

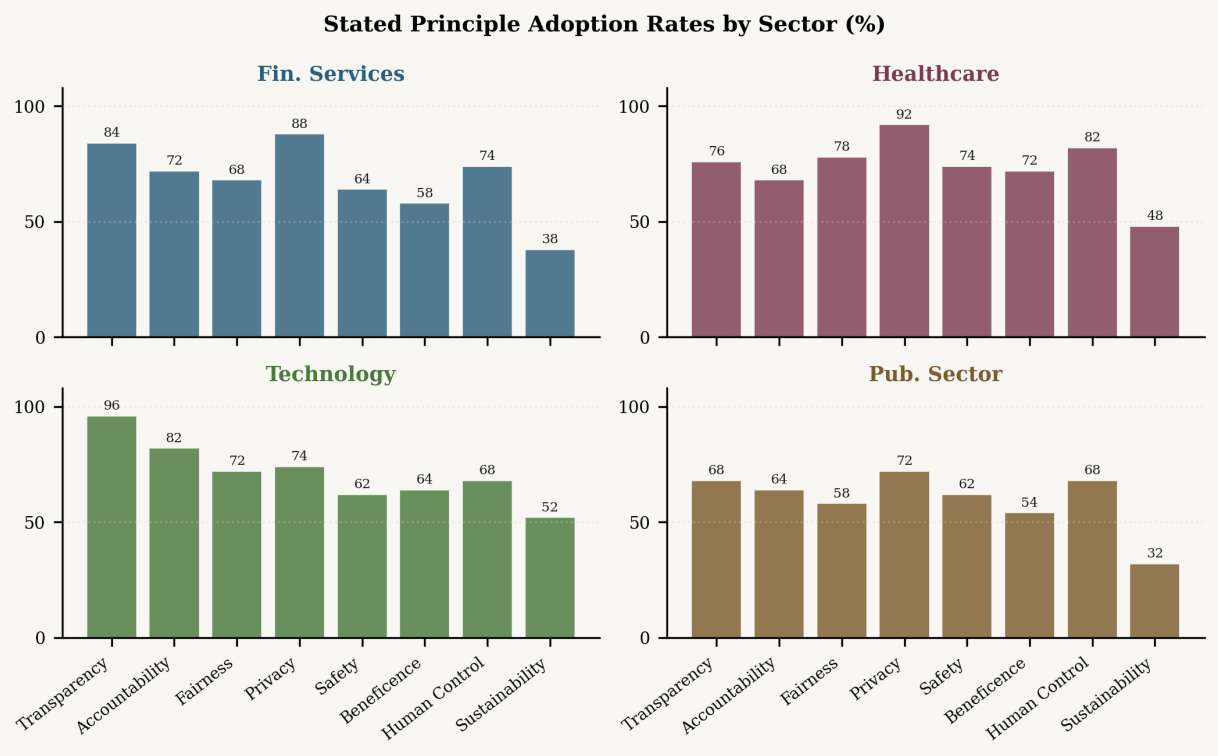

Figure 5. Stated principle adoption rates by sector (%). Healthcare shows the most consistent policy adoption across principles; the public sector shows high regulatory compliance but notably low fairness adoption, Healthcare achieves the most complete adoption profile; the public sector is notably low on fairness and beneficence despite its regulatory compliance strength.

Table 9. Comparison of present cross-sector findings with related empirical studies in the responsible AI governance literature.

Study	Scope	Key finding	Alignment
Jobin et al. [1]	84 frameworks	Princ. converge	Confirmed
Hagendorff [5]	22 guidelines	Ethics as PR	Confirmed
Morley et al. [7]	84 tools	Tooling lags norms	Confirmed
Raji et al. [13]	Audit gap	Accountability absent	Confirmed
Birhane et al. [27]	ML papers	Perf.≫fairness	Confirmed
Fjeld et al. [2]	36 docs	8 principle clusters	Confirmed

servation that AI governance obligations now overlap with data protection, product safety, financial conduct, and sector-specific regulatory requirements in ways that create jurisdictional ambiguity about which governance framework takes precedence and which team is responsible. The second was *measurement vacuum*: the difficulty of demonstrating governance maturity to external auditors, boards, or regulators when the metrics for fairness, transparency, and beneficence remain contested and unstandardised. The third was *developer resistance*: the friction between governance function demands for documentation, review, and constraint and engineering teams optimising for velocity. The fourth was *regulatory anticipation anxiety*: a distinctive challenge of the 2024 context in which organisations are attempting to implement European AI regulatory requirements before the technical standards and implementation guidance have been finalised. These qualitative themes provide interpretive context for the quantitative barrier data in Figure 6 and suggest that the implementation challenge is as much institutional and political as it is technical.

7. THE GOVERNANCE GAP: ANALYSIS AND IMPLICATIONS

7.1 What the Gaps Reveal

The study's central empirical contribution—the documentation of a large and consistent policy-to-practice gap across all eight governance principles— calls for structural explanation rather than organisational blame. In what follows, we offer three complementary explanations: an organisational-structural account, a technical-capacity account, and an institutional-legitimacy account. Taken together, these accounts suggest that the implementation gap is not a transitional phenomenon that will close automatically as responsible AI matures, but a structural feature of how organisations process ethical demands that requires specific countermea-

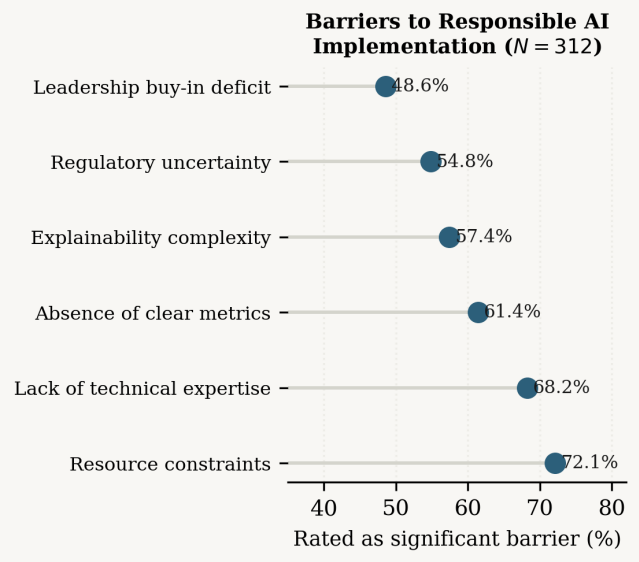

Figure 6. Barriers to responsible AI implementation as rated by survey respondents ($N = 312$, proportion citing as significant or very significant). Ordered by reported frequency, resource constraints lead. Leadership buy-in is less frequently cited but structurally the most consequential for governance authority.

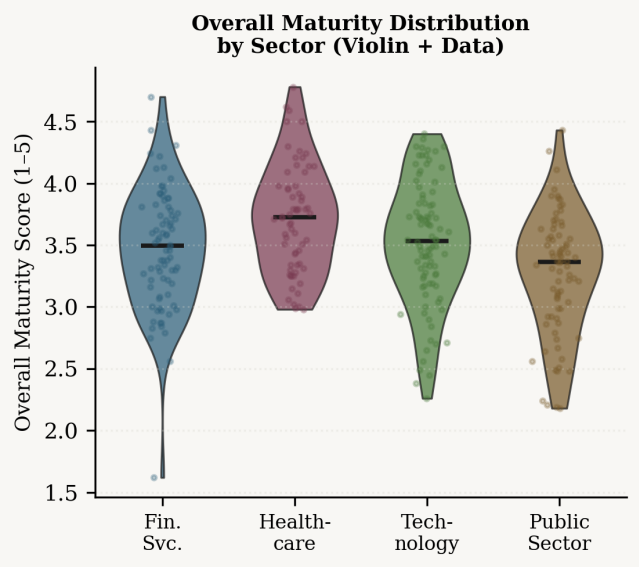

Figure 7. Overall responsible AI governance maturity distributions by sector. Violin plots with overlaid individual data points. Healthcare shows the highest median and tightest distribution; the public sector shows the widest spread, reflecting highly heterogeneous digital governance capability within that institutional category.

sures.

The implementation gap data in Figure 4 tell a story that is familiar from the broader literature on organisational ethics. When governance scholars observe that organisations readily adopt ethical principles but struggle to operationalise them, they are identifying a structural feature of how organisations process normative demands rather than a failure of individual or collective intention. Principles are costless to adopt: they require no infrastructure, create no accountability mechanism, and generate reputational benefit without operational commitment. Practice is expensive: it requires processes, tooling, expertise, audit functions, and the institutional authority to refuse or modify AI system designs that fail to meet governance standards.

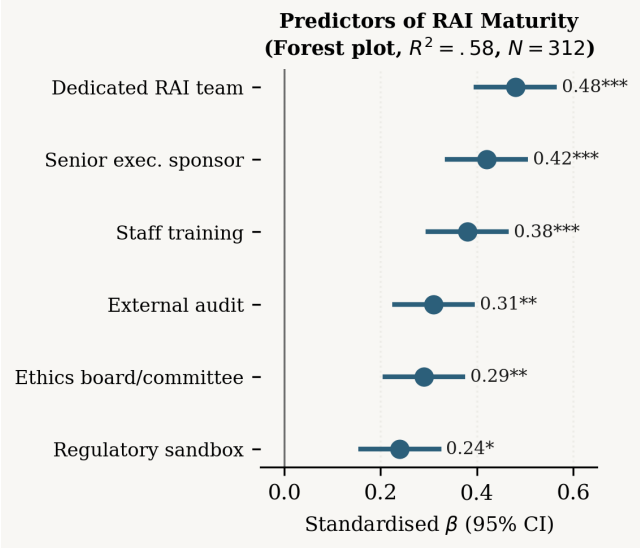

Figure 8. Standardised regression coefficients for predictors of overall responsible AI governance maturity. Forest plot showing standardised β coefficients with 95% confidence intervals. A dedicated RAI team is the single strongest predictor; all six predictors are positive and significant, confirming that governance maturity is driven by structural investment.

The accountability gap identified by Raji et al. [13] is a specific instance of this structural asymmetry: it is much easier to state that an organisation is committed to algorithmic accountability than to build the internal audit infrastructure, red-teaming capacity, and escalation pathways that accountability in practice requires. The bias mitigation gap is similarly structural: fairness requires sustained measurement, documentation, and intervention across the full model development lifecycle, from data collection through deployment monitoring, and this requires both technical capacity and organisational processes that most organisations have not yet institutionalised.

The technical capacity account holds that governance gaps persist because the technical tools needed to operationalise principles are insufficiently mature or accessible. Arrieta et al. [9] documented the proliferation of explainability methods, but also their uneven maturity, fragmented tooling ecosystems, and the significant expertise required to deploy them appropriately. Doshi-Velez and Kim [28] identified the evaluation challenge: it is genuinely difficult to determine whether a given explanation faithfully represents a model's decision logic rather than providing a post-hoc rationalisation that satisfies a human evaluator without accurately describing the model. Barocas et al. [11] showed that fairness metrics are mathematically incompatible in certain circumstances, meaning that organisations face genuine technical trade-offs, not merely implementation challenges, when attempting to simultaneously satisfy multiple fairness definitions. These technical difficulties are real constraints, but they do not fully explain the gap: privacy, where technical tools (differential privacy, encryption, access controls) are comparatively mature, shows a gap (28 pp) comparable in absolute terms to less technically developed principles.

The institutional legitimacy account holds that governance practices are adopted and maintained not because they are believed to be effective but because they confer legitimacy on organisations facing regulatory and reputational scrutiny. Under

this account, the policy-to-practice gap is not a failure to operationalise principles but a rational organisational response to an environment in which principle adoption is rewarded by regulators and the public but operational practice is not yet independently verified or enforced. European AI conformity assessment requirements are designed precisely to change this incentive structure by making operational practice— not principle adoption—the basis of regulatory compliance. If institutional legitimacy theory is correct, mandatory conformity assessment will close the gap more effectively than any amount of voluntary guidance.

7.2 Sector-Specific Interpretations

The public sector's high regulatory compliance maturity combined with low fairness maturity is arguably the most normatively troubling finding in the study. Public sector organisations are deploying AI systems in contexts— welfare benefit decisions, criminal risk assessment, immigration processing— where fairness is not merely a reputational concern but a constitutional and democratic obligation. The finding that fewer public sector organisations have operationalised fairness practices than financial services or healthcare organisations suggests that the democratic accountability structures that theoretically govern public sector AI are not, in practice, being translated into technical governance processes.

The technology sector's regulatory compliance deficit ($M = 2.88$, the lowest among the four sectors) reflects a structural feature of the sector's development trajectory. Technology companies have historically operated in a regulatory space that was either absent or permissive, developing governance practices primarily in response to reputational pressure and internal values commitments rather than binding legal obligations. The EU AI Act changes this fundamentally: technology companies deploying high-risk AI systems in European markets face mandatory conformity assessments, transparency obligations, and human oversight requirements that will close the gap between the sector's accountability strength and its compliance weakness. The acceleration of regulatory compliance maturity in the technology sector is, on this analysis, more a matter of when than whether.

Healthcare's privacy and human oversight leadership is interpretable as the dividend of decades of clinical governance infrastructure. Hospital ethics committees, institutional review boards, clinical audit processes, and patient safety reporting systems all provide the organisational architecture within which AI governance can be embedded with relatively lower marginal investment than in sectors that must build this architecture from scratch. This is an important practical lesson for other sectors: AI governance does not need to be designed in isolation but can be integrated into existing governance and risk management frameworks where they exist and are operational.

7.3 Implementation Guidelines

Six evidence-based guidelines emerge from the convergent findings of all three phases.

G1 — Establish a dedicated responsible AI function. The regression coefficient for a dedicated RAI team ($\beta = 0.48$) is the single strongest predictor in the model. Governance maturity cannot be achieved as a side activity of existing teams; it requires dedicated roles with the authority, expertise, and institutional mandate to enforce governance standards. The minimum viable RAI function—a team with representation from technical, legal, and domain expertise— is what separates organisations in the upper maturity quartile from those in the lower two.

G2 — Prioritise operationalisation over principle adoption. The implementation gap data confirm that adopting governance principles adds no measurable maturity unless accompanied by process, tooling, and accountability mechanisms. Organisations should audit the gap between their stated principles and their operational practices before adopting additional principles, and should treat the closure of existing gaps as a higher priority than the adoption of new frameworks.

G3 — Build bias monitoring into production, not just development. Fairness shows one of the largest policy-to-practice gaps (37 pp). This reflects a common pattern in which bias evaluation is treated as a pre-deployment activity rather than a continuous production monitoring obligation. Fairness metrics degrade over time as data distributions shift and population characteristics change; governance frameworks that do not include post-deployment monitoring are providing point-in-time rather than ongoing assurance [11, 12].

G4 — Use GDPR enforcement as a model for AI Act implementation. The privacy dimension shows the smallest policy-to-practice gap (28 pp) across all principles, and this gap pre-dates the AI governance discourse. The mechanism is GDPR enforcement: binding legal obligation backed by financial penalties and supervisory authority created the institutional pressure to operationalise privacy commitments. Organisations should treat proposed European AI conformity assessment requirements as an emerging enforcement mechanism for accountability and human oversight, and begin implementing the documentation and process requirements before regulatory obligations are operationalised [3, 4].

G5 — Invest in governance metrics before governance tooling. The most frequently cited barrier after resource constraints is the absence of clear implementation metrics (61.4%). Organisations cannot improve what they cannot measure; before investing in governance tooling such as explainability platforms or bias dashboards, organisations should define the metrics against which those tools will be evaluated. The NIST AI RMF [4] and ISO/IEC 42001 [19] both provide measurement frameworks that can be adapted to organisational context.

G6 — Embed AI governance in existing risk management frameworks. Healthcare's governance leadership across four of five dimensions reflects the dividend of existing clinical governance infrastructure. Organisations in other sectors should map AI governance requirements onto their existing risk management, audit, and compliance frameworks rather than building parallel governance structures, which add institutional complexity without adding accountability.

7.4 Limitations and Future Work

Several limitations qualify the study's conclusions. The practitioner survey draws on a self-selected sample with above-average governance engagement (71% formal AI policy),

meaning that the sector-level maturity scores are likely to over-represent governance-mature organisations and underestimate the implementation gap in the full population of AI-deploying organisations. Future work should develop sampling frames that include organisations without formal AI governance structures, to characterise the full distribution rather than the upper tail. The framework analysis applied a binary adequate-coverage coding scheme that conceals important variation in the depth, precision, and enforceability of principle coverage within frameworks; a more granular coding of operational depth would provide a richer basis for comparing framework quality. The study was conducted during a period of rapid European AI regulatory development, before full technical standards and enforcement mechanisms were operational; follow-up research will be needed to assess whether mandatory compliance requirements produce the governance dividend that the institutional legitimacy account predicts.

The cross-sectional design cannot establish causal direction between the regression predictors and governance maturity: it is plausible that higher-maturity organisations are more likely to establish dedicated RAI teams and ethics boards rather than that team establishment drives maturity. Longitudinal and quasi-experimental designs—natural experiments exploiting the staggered implementation of European AI obligations across risk tiers—are needed to establish causal mechanisms. Finally, the study covers four sectors in six Western countries; the governance maturity landscape in lower-income countries, in which AI deployment is expanding rapidly and regulatory frameworks are less developed [32], remains to be systematically characterised.

8. CONCLUSION

This paper set out to characterise the magnitude and distribution of the responsible AI implementation gap, and to identify the organisational conditions that predict governance maturity. The findings confirm and extend the diagnosis offered by the critical AI governance literature: the gap between stated principle and operational practice is large, consistent across sectors, and structured by the same organisational conditions—dedicated capability, senior sponsorship, training investment—that predict successful implementation in other governance domains.

What distinguishes the AI governance challenge from antecedent governance challenges is its urgency. The European Commission's AI Act proposal marks a transition in the compliance dimension of responsible AI from voluntary commitment toward binding legal obligation for organisations deploying high-risk AI systems in European markets [3]. The NIST AI RMF [4] and ISO/IEC 42001 [19] provide the operational frameworks through which organisations can begin closing the implementation gap in a structured and auditable way. The twenty-four-framework coverage analysis confirms that the normative architecture is largely in place; what remains is the sustained organisational investment—in teams, training, processes, and metrics—to make that architecture operational.

The sector-specific findings carry distinct practical implications. The public sector's fairness deficit, in particular, demands urgent attention from both organisational leaders and policymakers: democratic accountability obligations are not satisfied by governance principle adoption, and the deployment of AI systems in public sector decision-making contexts without operationalised fairness monitoring represents a risk to the procedural legitimacy of those decisions that legal frameworks have yet to fully address. The technology sector's compliance gap will close through regulatory obligation; the public sector's fairness gap may require more deliberate intervention.

The practical toolkit that emerges from this study—the maturity model, the framework coverage taxonomy, and the six implementation guidelines— is designed to be deployable by governance practitioners independently of the academic literature. An organisation that locates itself at Stage 2 of the maturity model can use the regression findings to identify the structural investments (team, sponsorship, training) that produce the largest maturity gain; can use the framework coverage analysis to identify which of the twenty-four governance instruments best maps to its sector and risk profile; and can use the implementation guidelines to sequence its governance investments in the order most likely to produce durable capability rather than symbolic compliance.

Future research should examine the longitudinal trajectory of the implementation gap as AI Act proposal enforcement provisions take effect, and should investigate whether the governance dividend observed for healthcare organisations—the compounding effect of existing clinical governance infrastructure—can be replicated in other sectors through deliberate integration of AI governance into established risk and compliance frameworks. The survey instrument and maturity framework developed in this study provide the methodological baseline for such longitudinal tracking.

8.1 Towards an Integrated Responsible AI Maturity Model

The study's findings converge on a maturity model that is, in structure, familiar from adjacent governance disciplines but distinctive in its emphasis on the principle-to-practice gap as the primary diagnostic dimension. Table 10 presents the five-stage maturity model that emerges from the qualitative and quantitative evidence.

Table 10. Responsible AI governance maturity model: stage descriptions and observable indicators.

Stage	Label	Observable indicators
1	Nascent	No formal AI policy; ethics not discussed
2	Aware	Principles adopted; no governance processes
3	Developing	Some processes; limited expertise; ad hoc
4	Managed	Dedicated RAI team; systematic monitoring
5	Optimised	Measurable outcomes; external audit; board-level reporting

The modal sector in the present study sits between Stage 2 and Stage 3: principles formally adopted (the policy adoption rates in Figure 5 average 77%), but operational processes inconsistently established (practice adoption rates average 37%). Moving organisations from Stage 2 to Stage 3 requires precisely the structural investments that the regression

analysis identifies as the strongest predictors: a dedicated RAI team, senior sponsorship, and training investment. Moving from Stage 3 to Stage 4—from ad hoc processes to systematic governance—requires the addition of measurement frameworks (Guideline G5) and external audit (significant at $\beta = 0.31$). Stage 5 optimisation, reached by fewer than 8% of organisations in the present sample, involves the integration of governance outcomes into board-level reporting and public transparency disclosures that close the accountability loop between internal governance and external stakeholder assurance.

The European Commission's AI Act proposal [3] would effectively require Stage 3 or above for organisations deploying high-risk AI systems in European markets; the NIST AI RMF [4] and ISO/IEC 42001 [19] provide the operational frameworks through which Stage 3 and Stage 4 can be systematically implemented. The evidence from this study suggests that the regulatory pressure associated with European AI regulation, applied to an organisational population currently averaging Stage 2 to Stage 3, requires a substantial escalation of governance investment that most organisations have not yet fully resourced.

8.2 Policy Implications

The findings carry distinct implications for the regulatory and policy actors who shape the governance landscape as well as for the organisations that must navigate it. For the European Commission and national supervisory authorities developing European AI regulatory implementation, the low operational practice rates for fairness (34%) and human oversight (42%) suggest that conformity assessment requirements for these dimensions may need to be accompanied by more detailed technical standards and implementation guidance than currently available, or the assessments will lack the consistency needed for effective enforcement. For standardisation bodies developing AI management system standards, the sustainability gap—lowest coverage and near-zero operational depth—represents the most significant unaddressed dimension in the current standard landscape.

For organisations navigating the transition from principle to practice, the healthcare sector's governance profile offers a model worth studying: not because healthcare has solved the responsible AI governance challenge, but because it demonstrates that integrating AI governance into pre-existing clinical governance infrastructure produces better governance maturity outcomes than designing AI governance in isolation. The principle of institutional integration —adding AI governance to existing risk, compliance, and ethics committee frameworks rather than creating parallel structures—is the most transferable lesson from healthcare to other sectors.

REFERENCES

[1] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019, doi: 10.1038/s42256-019-0088-2.

[2] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled AI: Mapping consensus in ethical and rights-based approaches to principles for AI," Berkman Klein Center for Internet and Society, Tech. Rep., 2020, research Publication 2020-1.

[3] European Commission, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts," European Commission, Tech. Rep. COM(2021) 206 final, 2021, 2021/0106(COD).

[4] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, Tech. Rep. NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.

[5] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020, doi: 10.1007/s11023-020-09517-8.

[6] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019, doi: 10.1038/s42256-019-0114-4.

[7] J. Morley, J. Cowls, M. Taddeo, and L. Floridi, "From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices," *AI & Society*, vol. 35, no. 4, pp. 1437–1456, 2020, doi: 10.1007/s00146-019-00902-7.

[8] L. Munn, "The uselessness of AI ethics," *AI & Society*, vol. 38, no. 4, pp. 1905–1914, 2023, doi: 10.1007/s00146-021-01289-6.

[9] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[11] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press, 2023, available at https://fairmlbook.org.

[12] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2018, pp. 149–159, doi: 10.1145/3287560.3287598.

[13] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[14] N. Diakopoulos, "Accountability in algorithmic decision making," *Communications of the ACM*, vol. 59, no. 2, pp. 56–62, 2016, doi: 10.1145/2844110.

[15] AI High-Level Expert Group, "Ethics guidelines for trustworthy AI," European Commission, Tech. Rep., 2019, available at https://digital-strategy.ec.europa.eu.

[16] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[17] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," in *Proceedings of the AAAI/ACM AIES Conference*. ACM, 2019, pp. 195–200, doi: 10.1145/3306618.3314289.

[18] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector," The Alan Turing Institute, Tech. Rep., 2019, doi: 10.23763/BrFav-0xaC.

[19] ISO/IEC, "ISO/IEC 42001:2023 — information technology — artificial intelligence — management systems," International Organization for Standardization, Tech. Rep., 2023, available at https://www.iso.org/standard/81230.html.

[20] V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Cham: Springer, 2019, doi: 10.1007/978-3-030-30371-6.

[21] M. Kearns and A. Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press, 2019.

[22] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish, "Algorithmic impact assessments and accountability: The co-construction of impacts," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 84:1–84:26, 2021, doi: 10.1145/3449208.

[23] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2018, pp. 77–91, doi: 10.1145/3287560.3287596.

[24] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing, 2016.

[25] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press, 2015.

[26] K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press, 2021.

[27] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Beilinson, "The values encoded in machine learning research," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2022, pp. 173–184, doi: 10.1145/3531146.3533083.

[28] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, arXiv:1702.08608.

[29] S. Wachter, B. Mittelstadt, and C. Russell, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017, doi: 10.1093/idpl/ipx005.

[30] C. Cath, "Governing artificial intelligence: Ethical, legal and technical opportunities and challenges," *Philosophical Transactions of the Royal Society A*, vol. 376, no. 2133, p. 20180080, 2018, doi: 10.1098/rsta.2018.0080.

[31] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[32] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, pp. 1–21, 2016, doi: 10.1177/2053951716679679.