Responsible Artificial Intelligence Governance and Sustainable Organisational Performance: An Integrative Framework for Accountability, Stakeholder Trust, and Digital Transformation

Sandra Ugonna Nwagbo, Prof. Gazi Md Nurul Islam

Universiti Tun Abdul Razak (UNIRAZAK)

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.101100274

Received: 21 October 2026; Accepted: 26 October 2026; Published: 25 November 2026

ABSTRACT

Responsible artificial intelligence (AI) governance is increasingly important as organisations integrate AI into decision-making and service delivery. However, the relationship between responsible AI governance and sustainable organisational performance requires further conceptual clarification, particularly regarding the roles of organisational accountability and stakeholder trust. This paper develops an integrative conceptual framework linking institutional pressures, responsible AI governance adoption, effective governance enactment, organisational accountability, stakeholder trust, organisational AI maturity, and sustainable organisational performance. Drawing on responsible AI governance literature, institutional theory, accountability perspectives, and stakeholder theory, the framework distinguishes the formal adoption of governance policies from their effective implementation in organisational practice. It proposes that accountability and stakeholder trust may operate as parallel mediating mechanisms, while organisational AI maturity may strengthen the relationship between effective governance enactment and accountability. The paper presents eight propositions and outlines a future empirical research design for evaluating the proposed relationships. As a conceptual contribution, it does not report original empirical data or claim that the framework has been empirically validated. The framework provides a basis for future research and offers organisational leaders a structured perspective for examining how responsible AI governance may contribute to accountability, trust, and sustainable performance.

INTRODUCTION

Artificial intelligence (AI) is increasingly embedded in organisational activities, including customer service, decision support, operations, risk assessment, and workforce management. These applications can create opportunities for efficiency and innovation, but they also introduce questions about responsibility, oversight, transparency, stakeholder impact, and risk management. The organisational challenge is not only whether AI is adopted, but how governance arrangements are enacted throughout the AI lifecycle. International guidance, including the NIST AI Risk Management Framework and the OECD AI Principles, provides important reference points for AI risk management and trustworthy practices (National Institute of Standards and Technology [NIST], 2023; Organisation for Economic Co-operation and Development [OECD], 2024). Such guidance establishes normative and operational reference points; it does not, by itself, demonstrate that governance arrangements improve organisational performance.

Responsible AI principles alone do not establish that an organisation has effective governance in practice. Governance requires arrangements connecting policies to decision rights, risk assessment, monitoring, documentation, review, escalation, and remediation. Reviews of responsible AI governance and real-world practices describe a broad range of structural, procedural, and relational practices, while also identifying variation in how these practices are conceptualised and implemented (Bach et al., 2025; Batool et al., 2025; Papagiannidis et al., 2025). Haque (2026) further develops a multidimensional organisational-capability account of responsible AI governance. These contributions mean that the present paper cannot claim novelty

merely for defining responsible AI governance as organisational, multidimensional, or relevant to sustainability. The narrower question pursued here concerns the proposed mechanisms through which effectively enacted governance may be associated with organisational outcomes.

The present paper develops a conceptual framework linking effectively implemented responsible AI governance with sustainable organisational performance. It proposes organisational accountability and stakeholder trust as parallel potential mediating mechanisms. It also considers organisational AI maturity as a possible moderator of the governance–accountability relationship and institutional pressures as contextual influences on formal adoption and implementation. The model draws on accountability scholarship, stakeholder theory, and institutional theory (Bovens, 2007; DiMaggio & Powell, 1983; Freeman, 1984). These perspectives address different analytical levels: answerability and corrective action, relationships with affected stakeholders, and external pressures shaping organisational practices.

The paper's provisional contribution is therefore mechanism-focused rather than a claim to have introduced the first responsible-AI-governance framework. It asks whether accountability and stakeholder trust provide distinguishable explanatory pathways between effective governance enactment and sustainable organisational performance, and whether AI maturity and institutional pressures help explain variation in these relationships. These propositions are theoretical expectations, not established effects. The review is a preliminary integrative narrative review rather than a systematic review: no reproducible database search or formal screening protocol has yet been completed. Novelty remains provisional until a documented comparison of full-text models confirms that this specific combination adds explanatory value.

The remainder of the paper is organised as follows. Section 2 reviews the literature and theoretical foundations; Section 3 presents the conceptual framework and eight propositions; Section 4 discusses theoretical integration and the proposed contribution; Section 5 outlines managerial and policy implications; and Section 6 sets out the research agenda. Section 7 presents a proposed empirical validation design, followed by limitations in Section 8 and the conclusion in Section 9.

Research Aim and Guiding Question

The aim is to develop an integrative conceptual framework explaining how effectively implemented responsible AI governance may be associated with sustainable organisational performance through accountability and stakeholder trust, while recognising organisational AI maturity and institutional pressures as contextual conditions.

Guiding research question: Through what organisational mechanisms and under what contextual conditions may responsible AI governance be associated with sustainable organisational performance?

LITERATURE REVIEW AND THEORETICAL FOUNDATIONS

Review Approach and Scope

This manuscript uses a focused integrative narrative review to develop a conceptual argument. The literature was selected purposively around five pre-specified themes: (1) responsible AI governance frameworks and lifecycle practices; (2) organisational accountability; (3) stakeholder trust and trustworthy AI; (4) institutional pressures and symbolic compliance; and (5) sustainable organisational or ESG-related outcomes. The review prioritised peer-reviewed review articles, theory-building work, foundational theoretical sources, and empirical studies directly relevant to one or more proposed relationships. Candidate sources were identified through targeted searches of scholarly and publisher-indexed records using combinations of terms such as “responsible AI governance,” “AI governance implementation,” “AI accountability,” “trustworthy AI,” “institutional pressure,” and “organisational performance.” Sources were retained for their conceptual relevance to the focal constructs and relationships, rather than to represent every publication in the field.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue XI November 2026

This is not presented as a systematic literature review. No reproducible search protocol, complete database coverage, screening log, duplicate-removal procedure, or PRISMA flow has been established for this draft. Consequently, the manuscript makes no claim that the literature is exhaustive or that the review is systematic. The selection process described above transparently states the review's scope and logic, but it is not a substitute for a reproducible systematic search. Before submission to a journal requiring a systematic or scoping review, the authors should document database-specific search strings, search dates, eligibility criteria, screening decisions, and a source-selection flow.

From Responsible AI Principles to Organisational Governance

Responsible AI research increasingly distinguishes between articulating ethical principles and establishing organisational practices that make those principles operational. Papagiannidis, Mikalef, and Conboy (2025) review responsible AI governance as a combination of structural, procedural, and relational practices, emphasising the need to translate abstract principles into activities across the AI lifecycle. This distinction is central to the present paper: the existence of an AI ethics policy is not equivalent to evidence that responsibilities are assigned, risks are monitored, decisions are traceable, or remedies are available. The proposed framework therefore focuses on governance as implemented organisational practice rather than a symbolic statement of intent. Official guidance provides complementary operational detail: the NIST AI Risk Management Framework organises risk work around governing, mapping, measuring, and managing AI risks (National Institute of Standards and Technology [NIST], 2023), while the OECD AI Principles emphasise human rights, transparency, robustness, accountability, and responsible stewardship across the AI lifecycle (Organisation for Economic Co-operation and Development [OECD], 2024).

Recent systematic reviews also report that governance concepts and implementation practices remain diverse. Batool, Zowghi, and Bano (2025) map the AI governance literature and underline the breadth of governance concerns across organisations and settings. Bach et al. (2025), reviewing responsible-AI practices in real-world settings, identify multiple practice areas and note a gap between discussion of responsible practices and evidence of their implementation. Together, these reviews support a research agenda that moves beyond lists of principles to examine organisational mechanisms, implementation conditions, and outcomes. They do not, by themselves, establish that any single governance arrangement will improve performance in every context. These policy and risk-management frameworks are useful reference points, but they should not be treated as evidence that any organisation has implemented governance effectively; implementation quality remains an empirical question (NIST, 2023; OECD, 2024).

Responsible AI Governance as an Organisational Capability

An organisational-capability perspective treats governance not simply as a policy document but as a repeatable capacity to coordinate roles, processes, oversight, learning, and corrective action. The capability framing is useful because AI systems change over time: data may shift, models may be updated, use cases may expand, and new harms may emerge after deployment. Governance arrangements must therefore support monitoring and adaptation rather than only pre-deployment approval. This interpretation is consistent with lifecycle-oriented governance principles reflected in the NIST AI Risk Management Framework (National Institute of Standards and Technology [NIST], 2023) and the OECD AI Principles (Organisation for Economic Co-operation and Development [OECD], 2024).

A 2026 integrative review by Haque proposes responsible AI governance as an organisational capability and argues that organisational, stakeholder, and environmental sustainability outcomes should be treated as distinct rather than assumed to improve together (Haque, 2026). This recent argument is particularly relevant to the present framework. It cautions against a simplistic 'more governance means better performance' claim: governance can impose coordination costs, and environmental effects depend on system architecture, energy sources, infrastructure, and deployment decisions. The present paper therefore treats sustainable organisational performance as a multidimensional outcome that requires careful operationalisation, and recognises that governance benefits may be conditional and involve trade-offs.

Accountability as an Explanatory Mechanism

Accountability is relevant when decision ownership, answerability, review, and corrective action are connected. Bovens (2007) conceptualises accountability as a relationship in which an actor may be required to explain and justify conduct to a forum that can question and assess it and may impose consequences. In AI-enabled organisations, accountability can become fragmented among executives, business owners, technical teams, vendors, data providers, and oversight functions. A responsible governance system should make these relationships sufficiently explicit for material decisions to be reconstructed and challenged. This paper proposes accountability as a potential mechanism through which governance arrangements may be translated into organisational conduct; whether this mechanism predicts performance remains an empirical question.

A complementary management-theory argument links control and accountability in AI development and use. Grote, Parker, and Crowston (2026) argue that AI's adaptive capabilities and distributed development-use relationships can make control and accountability difficult to align, and propose stakeholder governance and integrative negotiation across the AI lifecycle. This perspective cautions against measuring accountability solely through the presence of a named responsible officer: researchers should also examine whether actors possess the information, authority, and practical ability to intervene or remediate.

Stakeholder Trust and Organisational Outcomes

Stakeholder trust concerns a stakeholder's willingness to rely on an organisation under conditions of uncertainty or vulnerability. Trust in AI is not necessarily uniform across groups: employees, customers, suppliers, regulators, and communities may have different information, interests, and exposure to risk. Governance may support trust when it is visible through credible oversight, transparency appropriate to the context, meaningful human review, and accessible routes for redress. However, trust should not be treated as a proxy for objective safety, fairness, or performance. Stakeholders may trust a system that is poorly governed, or distrust a well-governed system because of prior experiences, limited information, or wider concerns about AI.

Recent work directly examining trustworthy AI and organisational trust strengthens the rationale for treating trust as a distinct mechanism, while also warning against assuming that technical principles automatically create organisational trust. A 2025 scoping review found that organisational context, longitudinal change, and organisational maturity remain important considerations in how trustworthy-AI mechanisms translate into trust (Vining et al., 2025). For the present model, this supports a conditional proposition: governance may contribute to stakeholder trust when implementation is credible and contextually meaningful, but trust remains an outcome to measure rather than a benefit to presume. This review directly supports treating organisational trust as a measured, context-sensitive outcome rather than an automatic consequence of trustworthy-AI principles.

Accordingly, the proposed model identifies stakeholder trust as a possible mediating pathway rather than an inevitable consequence of governance. Future research should specify the stakeholder group, the AI application, and the basis on which trust is assessed. It should also consider competing explanations, including organisational reputation, perceived service quality, prior experience, communication, and the actual performance of the AI system.

Sustainable Organisational Performance and Potential Trade-Offs

Sustainable organisational performance is used here to denote an organisation's capacity to maintain effective operations and value creation over time while managing material risks and sustaining legitimate stakeholder relationships. In this manuscript, the construct is bounded to organisational-level outcomes rather than treated as a synonym for corporate social responsibility or environmental sustainability. Depending on context, its operational domain may include operational effectiveness and quality, innovation and adaptive capacity, resilience, and stakeholder-related outcomes. Environmental or broader ESG indicators should be included only where the research question explicitly covers them and where a

defensible theoretical and measurement rationale supports their integration. Financial performance may be relevant but is not, by itself, sufficient evidence of sustainability.

A central methodological implication is that the proposed outcome requires a defensible operational definition and time horizon. Researchers should not aggregate operational, stakeholder, resilience, and environmental indicators simply because each is desirable. The preferred approach is to report theoretically justified dimensions separately or test a higher-order construct using appropriate measurement evidence. Designs should also account for possible short-term costs—such as assurance expenditure, slower deployment, and compliance workload—alongside potential longer-term benefits such as fewer incidents, organisational learning, legitimacy, and resilience. This avoids presuming that responsible governance improves every dimension of performance in every setting.

Research Gap and Rationale for the Proposed Framework

The literature reviewed identifies several issues warranting further investigation, but it does not establish that these issues are wholly unaddressed. Existing reviews have classified responsible-AI governance practices and developed governance frameworks (Papagiannidis et al., 2025); Haque (2026) explicitly conceptualises responsible AI governance as a multidimensional organisational capability and treats sustainability-oriented outcomes as consequences. Accordingly, a broad claim that governance is fragmented, that an organisational capability is needed, or that governance may support sustainability would not by itself constitute a defensible research gap.

The emerging evidence base also means that the paper should not imply that the governance–performance relationship is entirely unstudied. Xia et al. (2025) report a quasi-natural-experiment analysis of 342 securities firms and find an association between adoption of responsible-AI information-governance mechanisms and improved market valuation in their setting. This is relevant evidence for a possible governance–performance relationship, but it does not by itself establish the proposed accountability and trust mediation paths, generalise to all sectors, or demonstrate multidimensional sustainable performance. The present framework should therefore be positioned as a proposed explanation of mechanisms and boundary conditions, not as the first evidence that governance may relate to performance.

The narrower question pursued here is how effectively enacted governance may relate to organisational performance through two proposed mechanisms—accountability and stakeholder trust—and whether organisational AI maturity and institutional pressures help explain variation in those relationships. The paper distinguishes formal adoption from effective enactment and treats the proposed mediating and moderating relationships as hypotheses for future testing. This is an integrative, mechanism-focused proposition set. Before describing it as novel, the authors must complete a reproducible search and a full-text comparison of existing models, including the constructs, arrows, boundary conditions, and outcome definitions in Papagiannidis et al. (2025) and Haque (2026).

Comparison with Existing Governance Reviews

The proposed framework complements, but does not replace, existing responsible-AI-governance reviews. Papagiannidis et al. (2025) organise responsible AI governance practices and discuss antecedents and effects; Batool et al. (2025) map the breadth of AI-governance concerns; Bach et al. (2025) synthesise practices discussed in real-world settings; and Haque (2026) theorises responsible AI governance as a multidimensional organisational capability. These works overlap with the present paper in treating governance as more than an ethical statement and connecting governance to organisational outcomes. The proposed distinction is therefore not a new governance taxonomy. It is a narrower explanatory architecture: (a) accountability and stakeholder trust are specified as parallel, conceptually distinct mediating mechanisms; (b) AI maturity is positioned as a moderator of the enacted-governance–accountability path; and (c) institutional pressure is separated from governance enactment because external pressure may stimulate formal adoption without ensuring substantive implementation. The added value is a set of linked, falsifiable propositions that can be compared against direct-effect, single-mediator, and symbolic-

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue XI November 2026

compliance alternatives. This is a candidate theoretical contribution, not a confirmed claim of originality; full-text comparison and a reproducible search remain necessary.

Source / approach	Main emphasis supported by the source	Relationship to the present framework
Papagiannidis, Mikalef, and Conboy (2025)	Scoping review and framework for responsible AI governance, including structural, relational, and procedural practices, antecedents, and effects.	Provides a governance-practice foundation. The present model extends the discussion toward accountability and stakeholder trust as candidate explanatory mechanisms for organisational performance.
Batool, Zowghi, and Bano (2025)	Systematic review mapping AI governance concerns, including who is accountable, what is governed, when governance occurs in the AI lifecycle, and how it is implemented.	Supports the need to clarify governance actors and processes. The present model focuses more specifically on proposed pathways to sustainable organisational performance.
Bach et al. (2025)	Systematic review of 45 empirical studies on suggested responsible-AI practices in real-world settings; identifies multiple practice areas and notes a gap between suggested and observed practice.	Highlights the implementation challenge. The present model proposes accountability and stakeholder trust as mechanisms that future studies could measure alongside implementation quality.
Present conceptual framework	Integrates responsible AI governance, accountability, stakeholder trust, sustainable organisational performance, AI maturity, and institutional pressures.	Offers a testable conceptual synthesis. Its propositions are theoretical expectations, not empirical findings, and should be assessed against rival explanations and sector-specific conditions.

Positioning decision for revision: this manuscript retains the accountability-and-trust outcome-mechanism model rather than switching to a six-dimensional RAIG capability-development model. Haque (2026) already advances the latter approach. The two models should not be blended without a separate theoretical argument; the current paper instead tests, conceptually, whether accountability and trust can explain how enacted governance relates to performance.

Note. This comparison is a focused conceptual comparison of the sources cited in this draft, not the result of a comprehensive search of all available AI-governance frameworks.

Evidence boundary note: The sample sizes and study-design descriptions above are taken from the cited published studies. No statistics have been calculated for the present manuscript, and no primary data have been collected for it. The results of these external studies must not be presented as findings from the current conceptual paper.

Together, these studies strengthen the rationale for examining governance and performance while also clarifying the remaining research question. Existing evidence suggests that governance-related practices and information-governance mechanisms can be associated with organisational or sustainability outcomes in particular settings. It does not establish the full proposed model, especially the two parallel indirect pathways through accountability and stakeholder trust or the proposed role of organisational AI maturity. The present propositions should therefore be evaluated as a focused extension to be tested, not as mechanisms already confirmed by these studies. Differences in sector, country, governance operationalisation, outcome measures, and research design should be considered when interpreting or generalising the findings.

A related but distinct stream examines AI, information governance, and sustainability outcomes. Zhou et al. (2025) studied 1,418 Chinese A-share listed firms over 2011–2022 using a staggered difference-in-differences design around national AI pilot zones. The authors report improved corporate ESG performance and identify corporate information-governance channels, including transparency, information sharing, and reduced information asymmetry, as mechanisms. This evidence is adjacent rather than identical to the present study: AI adoption and corporate information governance are not interchangeable with responsible AI governance, and ESG performance is not synonymous with the broader sustainable organisational performance construct proposed here.

Published empirical studies provide useful context for the proposed framework, although their findings do not directly validate every pathway in this manuscript. Xia et al. (2025) examined responsible AI information-governance mechanisms and corporate performance using a sample of 342 securities firms. Their quasi-natural-experiment design used a multi-period difference-in-differences approach and reported statistically meaningful improvements in market valuation among firms adopting the governance mechanisms. This is directly relevant to the broad governance–performance relationship. However, the study does not by itself demonstrate that organisational accountability and stakeholder trust are parallel mediators, nor does its sector-specific sample establish that the same effects generalise to all organisations.

Empirical Evidence Relevant to the Proposed Relationships

Published empirical studies provide a useful basis for motivating the proposed framework, although they do not jointly test the complete model. Xia et al. (2025) analyse 342 securities firms using a quasi-natural-experiment design and report that adoption of responsible AI information-governance mechanisms is associated with statistically meaningful improvements in market valuation measured by Tobin's Q. Their design and sector-specific setting make the study relevant to the governance–performance relationship, but the reported results do not establish accountability and stakeholder trust as parallel mediators.

Zhou et al. (2025) investigate AI and corporate ESG performance in China using the establishment of AI innovation pilot zones as a quasi-natural experiment. The authors report an average increase of 0.727 in corporate ESG scores and identify corporate information-governance channels as part of the explanation. This study is relevant to the sustainability-oriented outcome in the present framework, but AI adoption and responsible AI governance are not interchangeable constructs; the finding should therefore be treated as adjacent evidence rather than direct validation of every proposed path.

Khan and Ladwig (2026) use survey data from 289 Australian managers and PLS-SEM to examine AI-supported performance evaluation. They report positive associations between fairness, accountability and transparency appraisals and trust, with transparency showing the largest coefficient among the three dimensions. Because the study uses a cross-sectional scenario-based survey, its results support the relevance of trust-related mechanisms but do not establish causal effects or generalise automatically to all stakeholder groups and AI applications.

Taken together, these studies provide partial and context-specific support for examining governance, organisational performance, and trust. They do not constitute a direct test of the proposed eight-proposition framework. The present paper should therefore distinguish established findings in prior studies from its own conceptual propositions and avoid implying that the full mediation and moderation model has already been empirically validated.

CONCEPTUAL FRAMEWORK AND RESEARCH PROPOSITIONS

Purpose and Design Logic of the Framework

The purpose of the proposed framework is to explain how responsible AI governance may relate to sustainable organisational performance and to specify organisational mechanisms through which this relationship could occur. Rather than treating governance as a direct and universally beneficial input, the framework distinguishes (i) formal adoption, meaning the presence of policies, committees, codes, or

declared commitments, from (ii) effective implementation, meaning that responsibilities, risk controls, human oversight, monitoring, escalation, and remediation operate in practice. The propositions concerning accountability, trust, and performance refer to effective implementation, not policy adoption alone. Institutional pressure is positioned upstream as a possible driver of adoption, while organisational commitment and enforcement conditions help explain whether adoption becomes enacted practice.

Responsible AI governance is positioned as the focal organisational construct. Organisational accountability and stakeholder trust are proposed as explanatory mechanisms. Sustainable organisational performance is the focal outcome. Organisational AI maturity and institutional pressures are considered contextual conditions that may influence implementation or the strength of proposed relationships. These placements are provisional and require comparison with existing frameworks before the model can be claimed as original.

Core Constructs

Responsible AI Governance

Responsible AI governance refers to the coordinated structures, procedures, oversight, and stakeholder arrangements through which AI use is directed and reviewed. Future measurement should capture actual practices rather than the mere existence of policy documents. Possible indicators include assigned ownership, documented risk assessment, review of material system changes, incident reporting, meaningful human oversight, and accessible escalation procedures. Measurement should draw on validated instruments where available and avoid assuming that all governance practices are interchangeable reflective indicators.

Organisational Accountability

Organisational accountability refers to the extent to which responsibility for AI-related decisions is assigned, documented, reviewable, and connected to corrective action. Potential indicators include clarity of decision ownership, traceability, review and escalation routes, and authority to remediate identified problems. Measures should distinguish formal allocation of responsibility from its practical exercise.

Stakeholder Trust

Stakeholder trust refers to relevant stakeholders' confidence that the organisation manages AI competently, fairly, transparently, and responsively. The respondent group must be specified in empirical research. Employees, customers, regulators, and business partners may form different judgements, so their responses should not be combined without a defensible sampling and measurement strategy.

Sustainable Organisational Performance

Sustainable organisational performance is conceptualised as the organisation's capacity to sustain effective operations and value creation over time while managing material risks and maintaining legitimate stakeholder relationships. Future research should operationalise the construct in a context-sensitive way. Subjective manager assessments, objective operational measures, stakeholder outcomes, and resilience indicators should not be combined without evidence that they form a coherent construct.

Organisational AI Maturity

Organisational AI maturity refers to the extent to which an organisation has developed the skills, processes, infrastructure, leadership support, and learning routines needed to use AI responsibly and effectively. Mature organisations may be better equipped to translate governance requirements into operational practices. Maturity should be distinguished from governance itself to avoid conceptual overlap.

Maturity is also not interchangeable with trust. The former concerns organisational resources and routines for responsible AI use; the latter concerns the judgements of specified stakeholder groups. Treating these as distinct constructs is important because an organisation may have sophisticated technical processes without

having earned stakeholder confidence, or may enjoy reputational trust despite weak governance. Empirical studies should test discriminant validity and consider whether maturity affects implementation quality, trust formation, or both.

Institutional Pressures

Institutional pressures refer to regulatory, professional, contractual, and peer-related influences that shape organisational expectations and practices. Such pressures may encourage governance adoption, but may also lead to symbolic compliance where visible policies are prioritised over effective implementation. Institutional pressures are therefore treated as contextual influences rather than direct evidence of governance quality.

Proposed Relationships

Responsible AI Governance and Organisational Accountability

Governance arrangements can clarify decision rights, assign oversight responsibilities, require documentation, and establish procedures for review and remediation. When implemented effectively, responsibility may become easier to identify and exercise. Policies that do not allocate authority or provide escalation routes may have limited practical effect.

Proposition 1 (P1): Higher levels of effectively implemented responsible AI governance are positively associated with organisational accountability for AI-related decisions.

Responsible AI Governance and Stakeholder Trust

Stakeholders may be more willing to rely on an organisation when AI practices are governed through credible oversight, understandable communication, human review, and accessible redress. The relationship may depend on whether stakeholders can observe or experience these practices. Formal governance declarations alone may not generate trust, particularly where prior conduct has undermined credibility.

Proposition 2 (P2): Effectively implemented responsible AI governance is positively associated with stakeholder trust in the organisation's responsible use of AI.

Organisational Accountability and Sustainable Performance

Clear responsibility and corrective procedures may support sustainable performance by improving incident response, reducing unresolved process failures, and enabling learning. However, disproportionate or overly bureaucratic arrangements may create costs. The proposition concerns a positive association under conditions of effective and proportionate implementation, not an unconditional effect.

Proposition 3 (P3): Organisational accountability for AI-related decisions is positively associated with sustainable organisational performance, subject to the quality and proportionality of implementation.

Stakeholder Trust and Sustainable Performance

Stakeholder trust may support continued engagement, cooperation, and durable organisational relationships. Trust may also reduce uncertainty in interactions involving AI-enabled services. Nevertheless, trust can be misplaced and does not guarantee fair or effective practices. Empirical assessment should examine trust alongside observable outcomes and governance evidence.

Proposition 4 (P4): Stakeholder trust in an organisation's responsible use of AI is positively associated with sustainable organisational performance.

Accountability as a Potential Mediating Mechanism

If responsible AI governance clarifies responsibilities and establishes review and corrective procedures, accountability may help explain how governance is translated into organisational conduct. Governance provides structures and processes, while accountability captures whether actors can be required to explain decisions and respond to identified problems.

Proposition 5 (P5): Organisational accountability mediates the relationship between effectively implemented responsible AI governance and sustainable organisational performance.

Stakeholder Trust as a Potential Mediating Mechanism

Responsible governance may influence stakeholder trust when its practices provide credible evidence of competence, fairness, transparency, and responsiveness. Trust may then relate to the durability of stakeholder relationships and other organisational outcomes. This pathway should be tested rather than assumed, and alternative explanations such as reputation, prior experience, service quality, and organisational size should be considered.

Proposition 6 (P6): Stakeholder trust mediates the relationship between effectively implemented responsible AI governance and sustainable organisational performance.

Organisational AI Maturity as a Contextual Condition

Organisations with greater AI maturity may possess stronger technical expertise, clearer workflows, better data practices, and more developed learning routines. These capabilities may help convert governance requirements into practical controls and make accountability mechanisms more effective. The possibility of overconfidence or lock-in means the moderating role should remain open to empirical scrutiny.

Proposition 7 (P7): Organisational AI maturity strengthens the positive association between effective responsible AI governance enactment and organisational accountability.

Institutional Pressures and Implementation

Regulatory requirements, professional standards, and contractual obligations may motivate organisations to establish AI governance practices. External pressure may, however, produce substantive implementation or symbolic compliance. The relationship may depend on enforcement, leadership, and whether governance is integrated into operational routines.

Proposition 8 (P8): Institutional pressures are associated with adoption of responsible AI governance, but their relationship with effective implementation depends on organisational commitment and enforcement conditions.

Integrated Conceptual Model

The model distinguishes institutional pressure, formal governance adoption, and effective governance enactment. P8 positions institutional pressures as an antecedent of formal adoption and proposes that organisational commitment and enforcement condition the transition from adoption to enactment. Effective governance enactment is linked to accountability (P1) and stakeholder trust (P2); accountability and trust are linked to sustainable organisational performance (P3 and P4), forming two proposed parallel indirect pathways (P5 and P6). Organisational AI maturity moderates the effective-enactment–accountability relationship (P7). The core model does not assert an additional unnumbered direct governance–performance path.

- Effective responsible AI governance enactment → Organisational accountability (P1)
- Effective responsible AI governance enactment → Stakeholder trust (P2)

- Organisational accountability → Sustainable organisational performance (P3)
- Stakeholder trust → Sustainable organisational performance (P4)
- P5–P6: parallel indirect effects through accountability and stakeholder trust. P7: AI maturity moderates the enactment–accountability path. P8: institutional pressure influences adoption, while commitment and enforcement condition implementation.

The model is conceptual and has not been empirically tested. A final diagram should distinguish direct paths, mediation paths, and contextual influences. If submitted as a conceptual article, the authors should explain how the integration advances existing frameworks. If tested empirically later, construct validity, temporal ordering, common-method bias, and alternative explanations will require attention.

Figure 1. Proposed Conceptual Framework and Eight Research Propositions

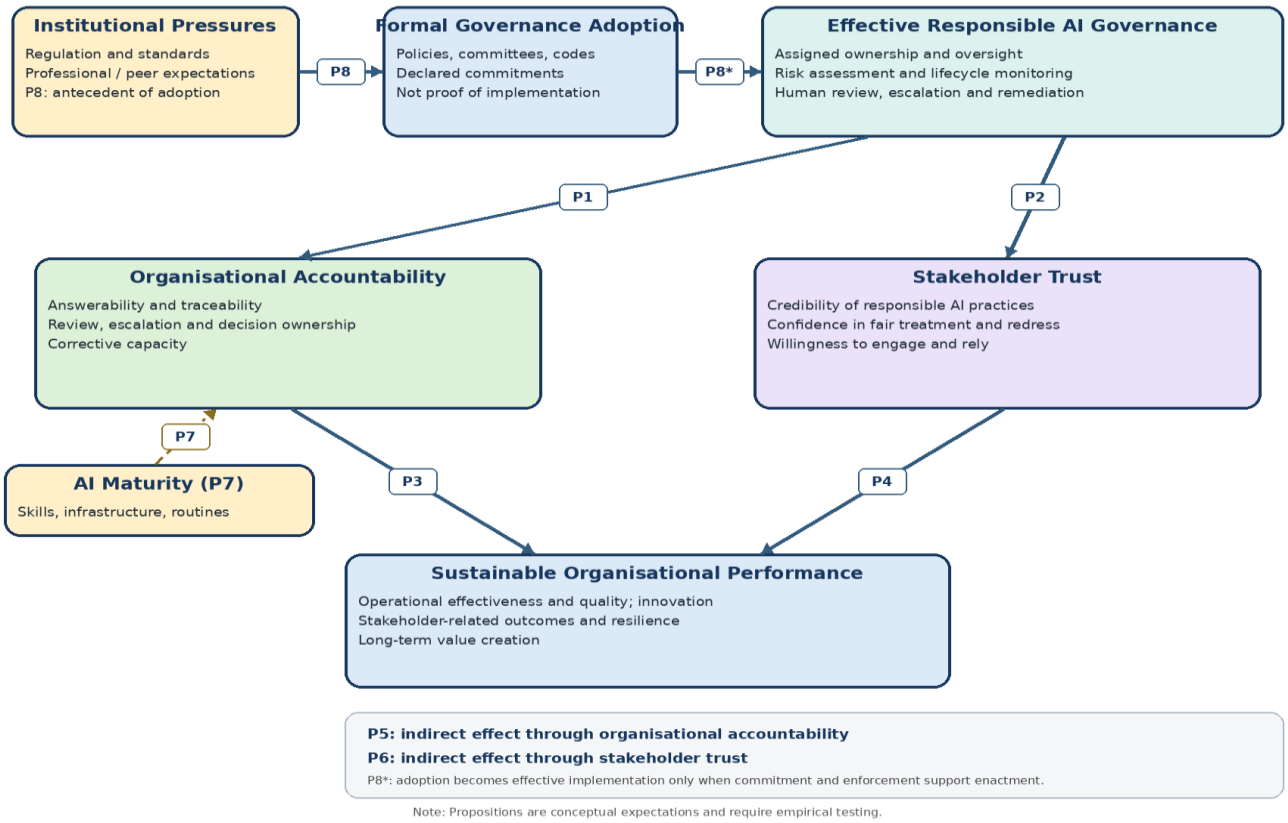

Figure 1. Proposed conceptual framework linking responsible AI governance, accountability, stakeholder trust, and sustainable organisational performance.

Boundary Conditions and Alternative Explanations

The proposed relationships may vary by AI application, sector, organisational size, regulatory environment, and the consequences of errors. A low-risk internal productivity tool may require different oversight from AI used in lending, recruitment, healthcare, or access to essential services. The perceived legitimacy of governance may also vary across stakeholder groups.

Organisational performance is influenced by factors beyond AI governance, including management quality, market conditions, financial resources, workforce capability, competitive pressure, and technology quality. Future studies should account for plausible confounders and avoid interpreting cross-sectional associations as causal effects.

Implications for Future Empirical Testing

A subsequent study could use a multi-respondent design to reduce reliance on one source. Governance practices might be assessed through risk, compliance, or technology managers; stakeholder trust through relevant employee or customer samples; and performance through independently sourced operational indicators where available. This design would require appropriate permissions, ethics review, feasible recruitment, and a sampling plan aligned with the research questions.

If data are collected at one point in time from the same respondents, limitations in causal inference and common-method variance should be acknowledged. Longitudinal data, multiple respondent groups, or objective performance measures could strengthen inference where feasible. The analysis method should follow the final measurement model and data characteristics.

Summary

This section proposes eight propositions connecting responsible AI governance with accountability, stakeholder trust, and sustainable organisational performance. It also identifies organisational AI maturity and institutional pressures as contextual factors. These propositions are theoretical expectations, not empirical findings. Their originality and value must be established through systematic comparison with prior research, and their validity requires subsequent empirical assessment.

THEORETICAL INTEGRATION AND POTENTIAL CONTRIBUTION

Integrating Stakeholder, Accountability, and Institutional Perspectives

The framework brings together three perspectives that address different aspects of responsible AI governance. Its theoretical logic rests on complementarity rather than simply placing three theories side by side: institutional theory explains why organisations may adopt governance arrangements; accountability theory explains how implemented arrangements can create answerability and corrective capacity; and stakeholder theory explains why credible governance may shape stakeholder trust and, in turn, longer-term organisational relationships and outcomes. This sequencing yields an adoption-to-enactment distinction and two separate explanatory pathways, while leaving open the possibility that external pressure produces symbolic rather than substantive compliance.

Accountability scholarship focuses on the relationship between actors who must explain and justify conduct and forums that can question, assess, or require responses to that conduct. Bovens (2007) provides a conceptual account of accountability that helps distinguish responsibility in name from answerability and review in practice. Applied to AI governance, this distinction highlights the importance of identifiable decision ownership, traceable processes, meaningful review, escalation, and corrective capacity.

Institutional theory helps explain why organisations adopt governance policies in response to regulatory expectations, professional norms, and practices adopted by peers. DiMaggio and Powell (1983) show how institutional pressures can produce convergence across organisations. In the AI context, convergence in formal policies does not necessarily demonstrate equivalent implementation quality. The framework therefore distinguishes governance adoption from effective implementation and treats institutional pressure as a contextual antecedent rather than a guarantee of responsible conduct.

These perspectives are complementary but not interchangeable. Stakeholder theory foregrounds relationships and legitimacy; accountability theory foregrounds answerability and corrective processes; institutional theory foregrounds the external conditions shaping organisational adoption. Their integration provides a way to examine governance structures, organisational mechanisms, and stakeholder-facing outcomes in a single proposed model. The framework should not be described as empirically validated or wholly novel until it has been compared systematically with existing responsible-AI and organisational-performance models.

Proposed Theoretical Contribution

The first potential contribution is a mechanism-based account of the relationship between effective responsible AI governance and sustainable organisational performance. Existing responsible-AI-governance reviews already discuss governance frameworks, antecedents and effects, so the contribution is not the general proposition that governance matters. Instead, this paper makes two candidate pathways explicit—organisational accountability and stakeholder trust—and proposes that future studies test whether they explain distinct portions of the governance–performance relationship. The model must remain open to the possibility that one pathway is unsupported, that the mechanisms overlap, or that other mechanisms provide a better explanation.

The second potential contribution is the explicit analytical separation of formal governance adoption from effective enactment. An organisation may publish a policy or establish a committee without embedding review, escalation, meaningful oversight, or corrective action in operational routines. The distinction connects institutional pressure to adoption while reserving the accountability and trust pathways for enacted practice. Because prior frameworks may already recognise implementation gaps, this is a proposed model-level clarification rather than a claim that the distinction has never been made.

The third potential contribution is the joint placement of organisational AI maturity and institutional pressure at different points in the explanatory sequence. AI maturity is proposed to condition the translation of enacted governance into accountability; institutional pressure is proposed to influence adoption but not to guarantee implementation quality. This distinction generates testable alternatives, including symbolic compliance under strong external pressure and the possibility that high maturity does not strengthen accountability where overconfidence or weak challenge processes prevail.

Finally, the framework treats sustainable organisational performance as multidimensional and context-dependent. Operational effectiveness, stakeholder-related outcomes, and long-term resilience may be relevant dimensions, but their combination must be justified conceptually and psychometrically. Researchers should not automatically aggregate these dimensions into a single score without evidence that the measure is appropriate.

Managerial and Policy Implications

For organisational leaders, the framework suggests that responsible AI governance should be operational rather than merely declarative. Governance responsibilities need to be assigned to named roles with sufficient authority, resources, and escalation access. Policies should be connected to the AI system lifecycle, including use-case approval, data and model risk assessment, deployment review, monitoring, incident management, change control, and retirement decisions. The appropriate controls will depend on the system's purpose, risk, and potential impact.

For risk, compliance, and technology teams, accountability requires more than documenting who approved a system. Records should support reconstruction of material decisions, identification of responsible actors, review of exceptions, and evidence that identified problems led to corrective action. Organisations should test whether escalation channels work in practice and whether employees can raise concerns without inappropriate retaliation.

For stakeholder engagement, organisations should communicate the purpose and limitations of material AI uses in language appropriate to the affected audience. Where people may be adversely affected, clear routes for human review, complaint, correction, or appeal can make governance more meaningful. Communication alone should not be treated as proof of trustworthiness; it should be consistent with actual system performance and organisational conduct.

For policymakers and industry bodies, the framework supports attention to implementation evidence as well as formal policy adoption. Evaluation may consider role clarity, documentation quality, monitoring, incident response, human oversight, and accessible remedies. Requirements should remain proportionate to the

context and should avoid encouraging box-ticking practices that impose administrative burden without improving risk management.

These implications are derived from the conceptual logic of the framework and established governance principles; they are not conclusions from a new empirical study. Organisations should adapt them to applicable law, sectoral standards, and the specific risks of their AI systems.

RESEARCH AGENDA AND METHODOLOGICAL CONSIDERATIONS

Construct Operationalisation

Future empirical work should define each construct before selecting indicators. Responsible AI governance may be operationalised through governance structures, risk-management processes, oversight and assurance, and stakeholder mechanisms. Researchers must decide whether these elements form a composite or higher-order construct, rather than presuming that they are interchangeable manifestations of one latent trait. Accountability measures should capture practical answerability and corrective capacity, while stakeholder trust measures should identify whose trust is being assessed and toward which AI use or organisational practice.

Sustainable organisational performance should be specified in a manner appropriate to the sector and unit of analysis. Potential indicators include operational quality, efficiency, innovation, stakeholder outcomes, resilience, and sustained value creation. Where objective indicators and perceptual measures are combined, the rationale and measurement properties should be reported transparently. Baseline performance and non-AI-related drivers should be considered.

Research Design and Analysis

A multi-source design would be preferable where feasible. Governance data could be obtained from risk, compliance, or technology personnel; stakeholder trust could be measured among the relevant employee, customer, or partner group; and performance could be assessed through independent operational records or appropriately validated managerial reports. Data collection should be preceded by ethics review, informed consent where required, and a defensible sampling strategy.

A cross-sectional survey can test associations but is generally insufficient to establish temporal order or strong causal claims. Longitudinal designs, matched multi-respondent data, natural experiments, or carefully justified case comparisons may provide stronger evidence about the proposed mechanisms. Researchers should assess construct validity, discriminant validity, measurement invariance where group comparisons are made, and common-method risks. Mediation and moderation analyses should follow the measurement model, sample characteristics, and pre-specified hypotheses; no particular statistical method should be selected merely because it is fashionable.

Qualitative research can complement quantitative testing by examining how accountability is enacted, how stakeholders interpret governance practices, and why formal policies may become symbolic. Comparative case studies across sectors or AI-risk categories may help identify boundary conditions. Mixed-method research could then connect implementation processes with measured outcomes.

Falsifiability and Competing Explanations

The framework should be treated as open to disconfirmation. Responsible AI governance may increase compliance costs or slow deployment in some settings; stakeholders may not observe governance practices; trust may depend more heavily on prior experience or service quality; and accountability arrangements may become bureaucratic rather than corrective. Organisational performance may also be driven by resources, leadership, competitive conditions, workforce capability, and the quality of the underlying AI system. Strong tests should include theoretically justified competing explanations and report null or contradictory findings rather than interpreting every result as support for the framework.

Proposed Empirical Validation Design

This section translates the conceptual framework into a feasible future empirical study. It is a proposed protocol, not a report of completed data collection or analysis. No sample has been recruited, no original observations have been analysed, and no empirical results are claimed.

Research Design and Unit of Analysis

A multi-industry, cross-sectional survey could be used for an initial test of the proposed relationships, with the organisation as the theoretical unit of analysis and knowledgeable managers or AI-governance professionals as key informants. Where feasible, a stronger design would collect responses from more than one informant per organisation and combine survey measures with documentary evidence of governance practices. A time-lagged design would help separate measurement of governance enactment, proposed mediators, and performance outcomes. Cross-sectional self-report data alone would be vulnerable to common-method bias and would not justify strong causal conclusions.

Target Population and Sampling

The target population should be defined as organisations that use, develop, procure, or materially rely on AI-enabled systems in organisational decision-making or service delivery. Eligible respondents could include senior managers, risk and compliance leaders, data/AI leads, internal auditors, and other staff with direct knowledge of governance arrangements. A sampling frame should be specified before recruitment, with sector, organisation size, and AI-use intensity considered as stratification variables where reliable frame data are available. Convenience recruitment may be practical for a pilot but would limit generalisability and should not be described as representative.

Sample-Size Determination

The final sample size should be justified before data collection, rather than chosen to produce significant results. For covariance-based SEM, the authors should conduct an a priori power analysis using the smallest effect considered theoretically meaningful, the number of predictors of each endogenous construct, the desired statistical power, and the planned significance level. For PLS-SEM, the analysis should also consider model complexity, minimum path coefficients of interest, statistical power, and the precision needed for indirect effects. Monte Carlo simulation is preferable when mediation and moderation are central. The resulting target should allow for unusable responses, non-response, clustering within organisations, and planned subgroup analyses. A numerical target should not be asserted until these assumptions have been specified.

Construct Operationalisation

Responsible AI governance should be measured through enacted practices rather than policy existence alone—for example, assigned ownership, documented risk assessment, human oversight, incident reporting, lifecycle monitoring, escalation, and remediation. Organisational accountability should capture answerability, traceability, reviewability, and authority to correct problems. Stakeholder trust should identify the stakeholder group being assessed and should not combine employees, customers, regulators, and business partners without a defensible measurement model. Organisational AI maturity should capture skills, infrastructure, leadership support, learning routines, and implementation capability while remaining conceptually distinct from governance. Sustainable organisational performance should be operationalised with context-sensitive indicators and, where possible, triangulated with objective outcomes rather than relying exclusively on one informant's overall rating.

Analysis Plan

The analysis should begin with data screening, missing-data assessment, descriptive statistics, and checks for non-response and common-method concerns. Measurement assessment should examine indicator quality,

internal consistency, convergent validity, and discriminant validity. If constructs are specified as formative or multidimensional, their measurement logic must be justified rather than assuming every indicator is reflective. Structural tests should estimate the proposed governance-to-accountability and governance-to-trust paths, the accountability/trust-to-performance paths, and the two indirect effects using bootstrap confidence intervals. The AI-maturity interaction should be tested with a prespecified moderation procedure. Institutional pressures should be modelled as an antecedent of formal adoption, while adoption and effective enactment should be measured separately. Alternative models, plausible confounders, sector differences, and sensitivity checks should be specified in advance. Results should be reported whether paths are supported, unsupported, or contrary to expectations.

Research Ethics and Transparency

Before recruitment, the researchers should obtain the required institutional ethics approval, provide informed-consent materials, protect commercially sensitive information, and explain how responses will be stored and reported. The study protocol, inclusion criteria, measurement choices, exclusion rules, and analysis plan should be documented before inspecting outcome results. If feasible, the authors should preregister the hypotheses and analysis plan and make non-sensitive materials available in a repository.

How Existing Empirical Evidence Informs the Proposed Study

Published evidence supports investigating, rather than presuming, the proposed relationships. Xia et al. (2025) report a quasi-natural experiment involving 342 securities firms and performance improvements associated with responsible AI governance adoption. Zhou et al. (2025) report an average 0.727 increase in corporate ESG scores in their AI-pilot-zone study. Khan and Ladwig (2026), using survey data from 289 Australian managers, report positive associations between fairness, accountability and transparency appraisals and trust. These studies use different constructs and research designs; none alone or together establishes the complete proposed framework.

Proposition-To-Model and Empirical Testing Alignment

Table 1 summarises how each proposition corresponds to the conceptual model and indicates the principal empirical test proposed for future validation. It is a research-design guide, not a report of completed tests or empirical findings.

Proposition	Model relationship	Proposed empirical assessment
P1	Effective governance enactment → organisational accountability	Estimate the governance-enactment-to-accountability path.
P2	Effective governance enactment → stakeholder trust	Estimate the path for a clearly specified stakeholder group.
P3	Organisational accountability → sustainable organisational performance	Estimate the association using prespecified performance indicators.
P4	Stakeholder trust → sustainable organisational performance	Estimate the association using stakeholder trust and performance measures.
P5	Indirect effect through organisational accountability	Estimate the indirect effect and report bootstrap confidence intervals.
P6	Indirect effect through stakeholder trust	Estimate the indirect effect and report bootstrap confidence intervals.
P7	AI maturity moderates effective enactment → accountability	Test the prespecified governance-enactment × AI-maturity interaction.
P8	Institutional pressures → formal adoption; commitment/enforcement condition adoption →	Model adoption separately from enactment and assess the adoption-to-

	enactment	enactment gap.

LIMITATIONS

This manuscript is conceptual and does not report original empirical data. Accordingly, its propositions have not been tested, and the direction or magnitude of the proposed relationships is unknown. The model simplifies a complex organisational phenomenon by focusing on accountability and stakeholder trust as potential mechanisms; other mechanisms, such as employee capability, organisational learning, perceived fairness, or risk-control effectiveness, may also matter.

The constructs are broad and may require different operational definitions across sectors, jurisdictions, stakeholder groups, and AI applications. Sustainable organisational performance is especially context-sensitive, and a single measure may not adequately capture its operational, stakeholder, and resilience dimensions. In addition, the framework draws on established theoretical perspectives but does not, by itself, demonstrate that the integration is original. A systematic and current literature review is needed to establish the precise gap and distinguish the proposed framework from prior models.

The framework's propositions are intentionally expressed as testable expectations rather than findings. Figure 1 and the proposition table should be read consistently: P1–P2 concern effective governance enactment; P3–P4 link the two proposed mechanisms to sustainable performance; P5–P6 specify parallel indirect effects; P7 specifies AI maturity as a moderator of the governance–accountability path; and P8 specifies institutional pressure as an antecedent of adoption, with implementation contingent on organisational commitment and enforcement. The model should not depict institutional pressure as a moderator of the accountability path or imply that formal adoption is equivalent to effective implementation.

CONCLUSION

This conceptual manuscript proposes a mechanism-focused framework for examining how effectively implemented responsible AI governance may relate to sustainable organisational performance through organisational accountability and stakeholder trust. It distinguishes formal governance adoption from effective enactment, considers organisational AI maturity as a possible boundary condition, and treats institutional pressures as influences on adoption whose relationship with implementation quality may depend on organisational commitment and enforcement.

The potential value of the framework lies in specifying pathways and contextual conditions that can be tested rather than assumed. Its contribution is provisional because recent work already offers responsible-AI-governance frameworks and an organisational-capability account. Before submission, the authors should complete a reproducible literature search, compare the full text of relevant models, verify every citation-to-claim match, confirm construct and diagram consistency, and align the manuscript with the target journal's scope. The paper remains conceptual: no empirical findings or causal effects are claimed.

REFERENCES

1. Zhou, X., Li, G., Wang, Q., Li, Y., & Zhou, D. (2025). Artificial intelligence, corporate information governance and ESG performance: Quasi-experimental evidence from China. *International Review of Financial Analysis*, 102, Article 104087. https://doi.org/10.1016/j.irfa.2025.104087
2. Xia, H., Chen, H., Zhang, J. Z., & Kamal, M. M. (2025). Exploring the impact of responsible AI governance on corporate performance: A quasi-natural experiment. *Technological Forecasting and Social Change*, 223, Article 124425. https://doi.org/10.1016/j.techfore.2025.124425
3. Vining, R., Karimian, S., McDonald, N., Ebiele, M., Doyle, B., McKenna, L., Ward, M. E., Bendechache, M., & Brennan, R. (2025). Trustworthy artificial intelligence and organisational trust: A scoping review using socio-technical systems analysis. *Ergonomics*. Advance online publication. https://doi.org/10.1080/00140139.2025.2512426

4. Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), Article 101885. https://doi.org/10.1016/j.jsis.2024.101885
5. Organisation for Economic Co-operation and Development. (2024). *OECD AI principles*. https://www.oecd.org/en/topics/ai-principles.html
6. 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
7. Khan, I., & Ladwig, R. C. (2026). Beyond efficiency: Interactional foundations of fairness, accountability and transparency in AI-supported performance evaluation. *Journal of Enterprise Information Management*. Advance online publication. https://doi.org/10.1108/JEIM-03-2026-0476
8. Haque, A. (2026). Responsible AI governance as an organisational capability: A theory-building integrative review for sustainable organisations. *Sustainability*, 18(20), Article 10501. https://doi.org/10.3390/su182010501
9. Grote, G., Parker, S. K., & Crowston, K. (2026). Taming artificial intelligence: A theory of control-accountability alignment among AI developers and users. *Academy of Management Review*, 51(2), 278–299. https://doi.org/10.5465/amr.2023.0117
10. Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman.
11. DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2), 147–160. https://doi.org/10.2307/2095101
12. Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x
13. Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics*, 5, 3265–3279. https://doi.org/10.1007/s43681-024-00653-w
14. Bach, T. A., Kaarstad, M., Solberg, E. A., & Babic, A. (2025). Insights into suggested Responsible AI (RAI) practices in real-world settings: A systematic literature review. *AI and Ethics*, 5, 3185–3232. https://doi.org/10.1007/s43681-024-00648-7