

RESPONSIBLE AI IN ORGANIZATIONS: EXPLORING MANAGERS' PERCEPTIONS, CHALLENGES AND GOVERNANCE PRACTICE

¹Yuslina Yusoff, ²Norazmila Yusuf,

¹Faculty of Business Management, Universiti Teknologi MARA Kelantan, Malaysia, ²Academy of Contemporary Islamic Studies, Universiti Teknologi MARA Kelantan, Malaysia

yuslinayusoff@uitm.edu.my, norazmila@uitm.edu.my

*Corresponding Author: Yuslina Yusoff (yuslinayusoff@uitm.edu.my)

Received 24 July 2026; Revised 10 August 2026; Accepted 20 Sept 2026; Published 01 Oct 2026

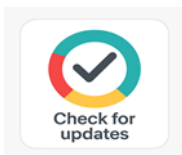
Vol: 2, Issue 2 (2026)

Doi: 10.5281/zenodo.23004978

Abstract

Keyword:

Responsible AI; AI governance; managers' perceptions; AI ethics; organizational practice; qualitative research; ATLAS.ti



Artificial intelligence (AI) is being adopted rapidly across organizations, yet the translation of high-level ethical principles into everyday managerial practice remains poorly understood. This study explores how managers perceive responsible AI (RAI), the challenges they face in implementing it, and the governance practices they adopt. Using a qualitative exploratory design, semi-structured interviews were conducted with 11 middle and senior managers from the banking, telecommunications, healthcare, public sector, higher education and manufacturing sectors in Malaysia. Data were analysed using reflexive thematic analysis supported by ATLAS.ti 24. Four themes emerged: (1) responsible AI as accountable and trustworthy decision-making, (2) the principle–practice gap, (3) emergent governance mechanisms, and (4) governance maturity as a continuum. Managers understood RAI primarily through the lenses of accountability, fairness, transparency and human oversight, and several anchored these in organizational and religious-ethical values such as *amanah* (trustworthiness). Key challenges included limited AI literacy, vendor opacity, regulatory ambiguity, resource constraints and unclear ownership. Governance practices were largely emergent, including cross-functional AI committees, risk-tiered approval, human-in-the-loop controls and vendor due diligence. The study contributes a manager-centred framework linking perceptions, challenges and governance maturity, and offers practical recommendations for organizations and policymakers seeking to operationalize responsible AI.



This is an open-access article under the CC BY-SA license.

Doi: 10.5281/zenodo.23004978

Introduction

Artificial intelligence has moved from experimental pilots to core organizational infrastructure. Organizations now use machine learning and generative AI for credit scoring, recruitment screening, customer service, clinical triage, fraud detection and content production (Dwivedi et al., 2021). This shift has been described as a change in the nature of management itself, because AI systems increasingly act with a degree of autonomy, learn from data, and remain partly inscrutable to the people who deploy them (Berente et al., 2021). As a result, decisions that were once made and justified by human managers are now shaped, and sometimes made, by algorithmic systems whose behaviour is difficult to predict or explain.

This expansion has been accompanied by growing concern about the “dark side” of AI, including biased outcomes, privacy violations, opacity, and the erosion of accountability (Mikalef et al., 2022). In response, governments, international bodies and firms have produced a large number of ethical principles and guidelines. A review by Jobin et al. (2019) identified 84 such documents and found convergence around transparency, justice and fairness, non-maleficence, responsibility and privacy. Frameworks such as the European Commission’s *Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on Artificial Intelligence, 2019), and more recently binding regulation in the form of the EU Artificial Intelligence Act (European Parliament & Council of the European Union, 2024), have made responsible AI a matter of compliance as well as ethics.

Principles alone, however, do not guarantee ethical outcomes. Mittelstadt (2019) argued that AI ethics lacks the professional norms, fiduciary duties and enforcement mechanisms that make principles work in fields such as medicine, while Hagendorff (2020) and Munn (2023) showed that guidelines often have little measurable effect on practice. Researchers have therefore called for a shift “from what to how”, that is, from articulating principles to operationalizing them through tools, processes and organizational structures (Morley et al., 2020). Organizational AI governance has emerged as the field concerned with this translation (Mäntymäki et al., 2022; Papagiannidis et al., 2025).

Within organizations, managers occupy a pivotal position in this translation. They approve AI projects, select vendors, allocate resources, define accountability, and mediate between technical teams, compliance functions and senior leadership. Yet most empirical work on responsible AI practice has focused on data scientists, engineers and AI practitioners (Madaio et al., 2020; Rakova et al., 2021), leaving the perspective of managers, who are often non-technical, comparatively under-explored. This gap is especially pronounced in emerging economies. In Malaysia, the Ministry of Science, Technology and Innovation (2024) introduced the *National Guidelines on AI Governance and Ethics*, but these guidelines are voluntary, and little is known about how managers interpret and apply them.

This study addresses that gap by exploring how managers perceive responsible AI, what challenges they encounter in putting it into practice, and what governance practices they adopt. Using qualitative interviews analysed with ATLAS.ti, the study makes three contributions. First, it provides a manager-centred account of how RAI is understood in practice, including value-based framings rooted in local culture. Second, it identifies the organizational, technical and institutional barriers that sustain the principle–practice gap. Third, it proposes a governance maturity continuum that organizations can use to assess and advance their RAI practices.

Literature Review

This section reviews the literature in six areas: how responsible AI has been conceptualized, the growth of ethical principles and standards, the gap between principles and practice, the emerging field of organizational AI governance, mechanisms for accountability and oversight, and the empirical research

on those who implement RAI inside organizations. It ends with a synthesis that identifies the research gap and the conceptual lens used in this study.

2.1 Conceptualizing responsible AI

Responsible AI refers to the design, development, deployment and use of AI in ways that are ethical, lawful, and aligned with human values and societal well-being. The concept overlaps with related terms such as “trustworthy AI”, “ethical AI” and “human-centred AI”, and these terms are often used interchangeably in both policy and research (Jobin et al., 2019). What they share is the idea that the value of AI cannot be judged only by its technical performance, such as accuracy or efficiency. It must also be judged by its effects on the people and communities it touches.

One of the most influential conceptual frameworks was proposed by Floridi et al. (2018) through the AI4People initiative. Drawing on established principles from bioethics, they proposed four principles, namely beneficence, non-maleficence, autonomy and justice, and added a fifth principle specific to AI: explicability. Explicability combines two ideas: *intelligibility*, meaning that people can understand how a system works, and *accountability*, meaning that it is clear who is responsible for how it works. Floridi et al. argued that explicability enables the other four principles, because it is difficult to judge whether a system is fair or harmful if nobody can understand it or be held responsible for it.

Arrieta et al. (2020) extended this view from a technical perspective. In their review of explainable AI (XAI), they argued that explainability is a necessary part of a broader concept of responsible AI, which also includes fairness, privacy, accountability and security. They emphasized that explanations must be tailored to their audience: what a data scientist needs to debug a model is very different from what a manager, regulator or affected citizen needs to trust a decision. Taken together, these conceptualizations show that RAI is not a single property of a technical system. It is a socio-technical outcome that depends on how systems are designed, governed and used within organizations. This view places organizational actors, including managers, at the centre of RAI.

2.2 The proliferation of principles, guidelines and standards

Since around 2016, the number of AI ethics guidelines has grown dramatically. In a landmark review, Jobin et al. (2019) analysed 84 guidelines issued by governments, companies, research institutions and civil society. They found convergence around five principles, namely transparency, justice and fairness, non-maleficence, responsibility and privacy, but substantial differences in how these principles were interpreted, why they were considered important, and how they should be implemented. They also noted that most guidelines came from economically developed countries, with the Global South, including much of Southeast Asia, underrepresented. Hagendorff (2020), in a separate evaluation of 22 major guidelines, found that they rarely specified enforcement or implementation mechanisms and tended to omit issues such as labour conditions, environmental costs and the concentration of power in large technology firms.

At the regional and international level, several frameworks have become particularly influential. The European Commission’s High-Level Expert Group on Artificial Intelligence (2019) set out seven requirements for trustworthy AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. These requirements later informed the EU Artificial Intelligence Act (European Parliament & Council of the European Union, 2024), the first comprehensive, binding AI law. The Act adopts a risk-based approach, prohibiting certain “unacceptable-risk” practices and placing obligations on providers and deployers of “high-risk” systems, including requirements for risk management, data governance, documentation, transparency and human oversight.

Standards bodies have tried to make principles more operational. The NIST *AI Risk Management Framework* (National Institute of Standards and Technology, 2023) describes the characteristics of trustworthy AI, including validity and reliability, safety, security and resilience, accountability and transparency, explainability, privacy and fairness, and organizes practice around four functions: *govern*, *map*, *measure* and *manage*. ISO/IEC 42001:2023 goes further by specifying requirements for an AI management system that organizations can implement and be certified against, using the familiar management-system logic of policies, roles, risk assessment, monitoring and continual improvement (International Organization for Standardization, 2023). In Malaysia, the Ministry of Science, Technology and Innovation (2024) introduced the *National Guidelines on AI Governance and Ethics*, which set out seven principles, including fairness, reliability and safety, privacy and security, inclusiveness, transparency, accountability, and the pursuit of human benefit, and describe the responsibilities of end users, policymakers and technology developers. Unlike the EU AI Act, however, these guidelines are voluntary. This raises the question of how organizations, and the managers within them, interpret and act on them.

2.3 The principle–practice gap

A consistent finding in the literature is that ethical principles do not translate easily into organizational behaviour. Mittelstadt (2019) made one of the most influential arguments on this point. He compared AI ethics with medical ethics, which is often held up as a model, and found four important differences. AI development lacks the common aims and fiduciary duties that bind doctors to patients, lacks a long professional history with established norms, lacks proven methods for turning principles into practice, and lacks strong legal and professional accountability mechanisms. Without these supports, he argued, principles alone cannot guarantee ethical AI and risk producing only the appearance of ethical behaviour.

Morley et al. (2020) examined the problem from a practical angle, asking how principles could be moved “from what to how”. They mapped publicly available tools and methods against both ethical principles and the stages of the machine learning lifecycle. They found that tools were unevenly distributed: many focused on explicability and on early stages of development, while few supported later stages such as deployment and monitoring. Many tools were also difficult for non-specialists to use. Munn (2023) offered a stronger critique, arguing that many AI ethics principles are ineffective because they are meaningless, being too abstract to guide action; isolated, being disconnected from the business and engineering contexts where decisions are made; and toothless, lacking consequences when they are ignored.

Empirical studies with practitioners support these concerns. Rakova et al. (2021) interviewed practitioners working on responsible AI in industry and found that most organizational practices were ad hoc and reactive, often dependent on individual champions who took on ethical work in addition to their formal roles. They distinguished between prevalent practices, which were largely reactive; emerging practices; and aspirational practices, in which responsible AI is proactively built into organizational structures. This distinction suggests that the principle–practice gap is not the same in every organization and may narrow as organizations develop. What remains less clear is how this gap is experienced by managers, who approve and oversee AI systems but are rarely the ones building them.

2.4 Organizational AI governance

In response to the principle–practice gap, researchers have increasingly turned to organizational AI governance. Mäntymäki et al. (2022) defined it as a system of rules, practices, processes and technological tools used to ensure that an organization’s use of AI technologies aligns with its strategies, objectives and values, fulfils legal requirements, and meets ethical principles. This definition is important because it moves the focus from what principles say to how organizations act. It also recognizes that governance must be connected to organizational strategy rather than existing as a separate ethics exercise.

Birkstedt et al. (2023) reviewed the emerging AI governance literature and found it fragmented across disciplines, with much of the work being conceptual or focused on policy and regulation. They identified a significant knowledge gap regarding how governance is actually enacted within organizations and called for more empirical research at the organizational level. Papagiannidis et al. (2025) responded with a review and research framework for responsible AI governance. They distinguished between three types of governance practices: *structural* practices, such as committees, roles and reporting lines; *procedural* practices, such as risk assessments, review processes and monitoring; and *relational* practices, such as stakeholder engagement, communication and collaboration. They noted that most existing research still concentrates on principles rather than practices, and that relational practices are especially under-studied.

From a more engineering-oriented perspective, Lu et al. (2024) developed a responsible AI pattern catalogue based on a systematic review of academic and industry literature. The catalogue organizes reusable solutions at three levels: multi-level governance patterns for industry, organizations and teams; process patterns for trustworthy development; and product patterns for building RAI into system design. Their work provides a practical bridge between high-level principles and concrete organizational routines. At the level of management theory, Berente et al. (2021) argued that AI presents distinctive challenges because of its autonomy, its capacity to learn and its inscrutability, which together make traditional forms of managerial control less effective. This helps explain why governance of AI requires new structures and processes rather than simply extending existing IT governance.

2.5 Accountability, auditing and human oversight

Accountability is a recurring theme across principles, standards and regulation, and several mechanisms have been proposed to operationalize it. Raji et al. (2020) proposed an end-to-end framework for internal algorithmic auditing, known as SMACTR, which stands for Scoping, Mapping, Artifact collection, Testing and Reflection. The framework assigns documentation and review responsibilities at each stage of development so that ethical risks can be identified and addressed before deployment, rather than discovered afterwards. Importantly, it treats auditing as an internal organizational process rather than something done only by external regulators.

Mökander and Floridi (2021) argued for ethics-based auditing, a structured process through which an organization's present or past behaviour is assessed for consistency with relevant ethical principles or norms. They suggested that such auditing can bridge the gap between principles and practice by providing evidence that commitments are being met, while also promoting procedural regularity and transparency. However, both internal and ethics-based auditing assume that the organization has access to the data, models and documentation needed to conduct the audit, an assumption that may not hold when AI is purchased from external vendors.

Human oversight is another central mechanism. Shneiderman (2020) proposed a three-level governance structure for human-centred AI. The first level is reliable systems based on sound software engineering practices. The second is a safety culture built through organizational design and management strategies. The third is trustworthy certification through external reviews and independent oversight. He emphasized that meaningful human control should be preserved for consequential decisions. The EU AI Act similarly requires that high-risk systems be designed so that they can be effectively overseen by humans (European Parliament & Council of the European Union, 2024). Yet the literature says relatively little about how managers themselves understand oversight and accountability, or how they decide which AI decisions require human review.

2.6 Managers, practitioners and the research gap

Empirical research on RAI practice has concentrated mainly on technical practitioners. Madaio et al. (2020) co-designed AI fairness checklists with practitioners from technology organizations and found that the usefulness of checklists depended heavily on organizational culture, incentives and leadership

support. Practitioners worried that raising fairness concerns could slow projects or be seen as obstructive, and they wanted checklists to be supported by organizational processes rather than treated as individual responsibilities. Rakova et al. (2021) similarly identified enablers of responsible practice, such as leadership support, clear roles, accountability structures and aligned incentives, and barriers, such as short-term product pressures and a lack of accountability. Many of these enablers and barriers fall directly within managerial responsibility.

Broader reviews have also called for more attention to organizational actors. Dwivedi et al. (2021), drawing on contributions from a large number of experts, highlighted the governance, ethical and skills-related challenges that organizations face in adopting AI and called for research on how these challenges are managed in practice. Mikalef et al. (2022) argued that research on the “dark side” of AI should look beyond the technology to examine how organizations and their members make sense of, and respond to, the risks of AI. Despite these calls, few studies have examined managers directly, even though managers approve AI projects, select vendors, allocate resources and are ultimately answerable for outcomes.

The gap is even more pronounced in emerging economies. As Jobin et al. (2019) observed, the global discussion of AI ethics has been dominated by perspectives from North America, Europe and East Asia. In Southeast Asia, national guidelines are recent and voluntary (Ministry of Science, Technology and Innovation, 2024), and there is limited empirical evidence on how organizations interpret them. Local cultural and religious values may also shape how responsibility is understood, but such values are largely absent from the mainstream RAI literature.

2.7 Synthesis and conceptual lens

The literature reviewed above reveals three gaps. First, although RAI has been extensively conceptualized and codified in principles and standards (Floridi et al., 2018; High-Level Expert Group on Artificial Intelligence, 2019; National Institute of Standards and Technology, 2023), there is limited evidence on how these principles are interpreted by managers, who are often non-technical but formally accountable. Second, the principle–practice gap is well established (Mittelstadt, 2019; Morley et al., 2020; Munn, 2023), but most empirical accounts come from technical practitioners (Madaio et al., 2020; Rakova et al., 2021), leaving the managerial experience of the gap under-explored. Third, research on organizational AI governance remains largely conceptual (Birkstedt et al., 2023), and frameworks distinguishing structural, procedural and relational practices (Papagiannidis et al., 2025) have rarely been examined empirically, particularly in non-Western contexts.

To address these gaps, this study adopts an organizational AI governance lens (Mäntymäki et al., 2022; Papagiannidis et al., 2025) that connects three elements: how managers perceive RAI, the challenges that sustain the principle–practice gap, and the structural, procedural and relational practices organizations adopt in response. Drawing on Rakova et al.’s (2021) observation that practices range from reactive to aspirational, the study also remains open to the possibility that governance develops over time. This lens directly informs the three research questions presented in the next section.

Research Questions

1. How do managers perceive and interpret responsible AI in their organizations? (RQ1)
2. What challenges do managers face in implementing responsible AI? (RQ2)
3. What governance practices do managers adopt to ensure the responsible use of AI? (RQ3)

Methodology

Research design

This study adopted a qualitative exploratory design grounded in an interpretivist paradigm, suitable for understanding how participants construct meaning around an emerging phenomenon. Reflexive thematic

analysis (Braun & Clarke, 2019, 2021) was used because it allows themes to be developed through the researcher's active engagement with the data rather than being assumed to "emerge" passively.

Participants and sampling

Purposive sampling was used to recruit managers directly responsible for approving, overseeing or using AI systems. The inclusion criteria were: (a) holding a middle or senior managerial role; (b) at least five years of managerial experience; and (c) direct involvement in at least one AI initiative within the previous two years. Each participant came from a different organization. Eleven managers from six sectors took part. A sample of this size is appropriate for an exploratory study using a specific, information-rich sample and in-depth dialogue (Malterud et al., 2016).

Saturation was assessed using the approach of Guest et al. (2020). The base size was the first six interviews, and the run length was two interviews. Interviews 7 to 9 contributed four new codes, which is 8% of the final code set. Interviews 10 and 11 contributed no new codes, so data collection ended at eleven.

Table 1. Participant profile

Code	Sector	Position	Sex	Mgmt. exp. (yrs)	AI use context	Mode	Duration (min)
P1	Banking	Head of Risk Management	M	14	Credit scoring model	Face-to-face	68
P2	Banking	Compliance Manager	F	11	Anti-money-laundering screening	MS Teams	55
P3	Telecommunications	Human Resource Manager	F	10	AI-based CV screening	MS Teams	52
P4	Telecommunications	Head of Customer Analytics	M	12	Churn prediction, customer chatbot	Face-to-face	71
P5	Healthcare (private hospital)	Hospital Operations Manager	F	15	Emergency patient triage	Face-to-face	64
P6	Healthcare (specialist centre)	Clinical Informatics Lead	M	7	Diagnostic imaging support	MS Teams	58
P7	Public sector (federal agency)	Deputy Director, Digital Services	M	16	Citizen service chatbot	Face-to-face	75
P8	Public sector (state agency)	Procurement Manager	M	12	Procurement of vendor AI systems	MS Teams	49
P9	Higher education	Deputy Dean (Academic)	F	11	Generative AI in teaching and assessment	Face-to-face	62
P10	Manufacturing (SME)	Operations Manager	M	10	Predictive maintenance	MS Teams	46
P11	Manufacturing (multinational)	Chief Strategy Officer	M	20	Generative AI for reporting	Face-to-face	57

Note. Mean managerial experience = 12.5 years; mean interview length = 59.7 minutes; total audio ≈ 10.9 hours.

4.3 Data collection

Semi-structured interviews (46–75 minutes) were conducted in English and Bahasa Malaysia, face-to-face or via Microsoft Teams. The interview protocol covered four areas: understanding of RAI, current AI use, implementation challenges, and governance arrangements. Interviews were audio-recorded with consent, transcribed verbatim, and Malay-language segments were translated and back-translated to preserve meaning.

4.4 Data analysis using ATLAS.ti

Transcripts were imported into ATLAS.ti 24 and analysed following Braun and Clarke's (2019) six phases, adapted to the software's workflow (Frieze, 2019): (1) familiarization through repeated reading and memo-writing; (2) open coding of quotations; (3) grouping codes into code groups; (4) developing candidate themes using the Network function; (5) reviewing themes with the Code-Document Table and Code Co-occurrence Table; and (6) defining themes and reporting. Open coding produced 286 quotations and 52 initial codes, which were merged into 21 final codes within 11 sub-themes and 4 themes. Figure 1 shows this analytical workflow.

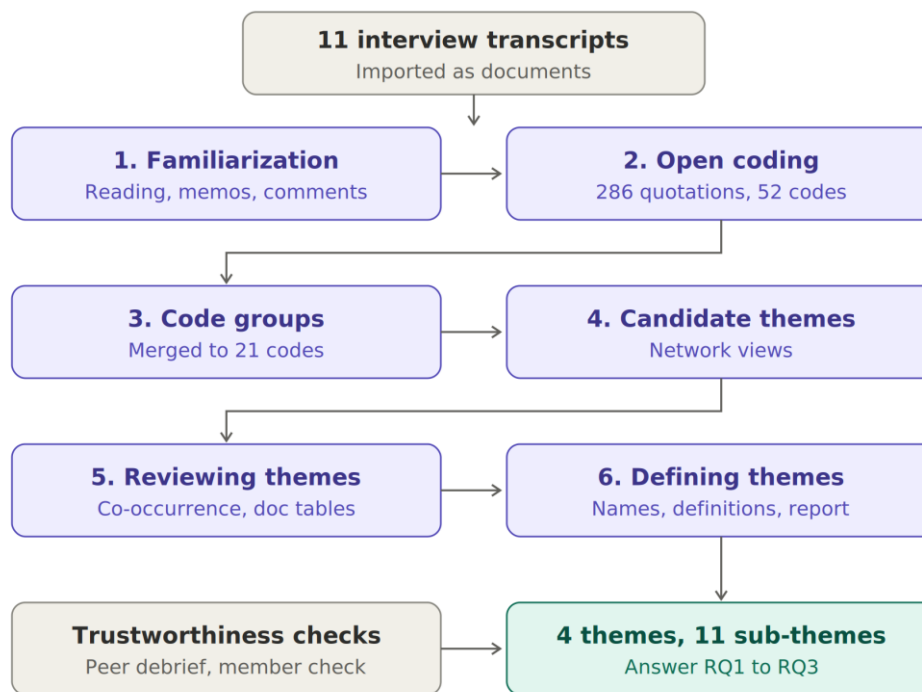


Figure 1. Thematic analysis workflow

Trustworthiness

Credibility, dependability, confirmability and transferability were addressed following Nowell et al. (2017). Strategies included an audit trail of memos and code-book versions within ATLAS.ti, peer debriefing with two experienced qualitative researchers, member checking of theme summaries with four participants, and thick description of context. Rather than calculating inter-coder agreement, which is inconsistent with reflexive thematic analysis (Braun & Clarke, 2021), a second researcher reviewed a sample of coded transcripts to challenge and refine interpretations.

4.6 Ethical considerations

Ethical approval was obtained from the university research ethics committee. Participants gave written informed consent, were assured of anonymity and confidentiality, and could withdraw at any time. Identifying details of organizations were removed, and participants are referred to by codes (P1–P11).

Findings

The analysis produced 286 coded quotations across 21 final codes. These were organized into four themes and eleven sub-themes. Table 2 shows the thematic structure. Table 3 shows how the quotations are distributed across participants (ATLAS.ti Code-Document Table), and Table 4 shows the strongest code co-occurrences.

Table 2. Themes, sub-themes and groundedness

Theme	Sub-theme	Quotations (G)	Participants (n = 11)
T1. RAI as accountable and trustworthy decision-making (RQ1)	1.1 Fairness and non-discrimination	26	9
	1.2 Transparency and explainability	29	10
	1.3 Human accountability and oversight	33	11
	1.4 Values-based framing (<i>amanah</i> , integrity)	15	6
T2. The principle–practice gap (RQ2)	2.1 Limited AI literacy among managers	25	9
	2.2 Vendor opacity and “black-box” systems	22	7
	2.3 Regulatory ambiguity	19	7
	2.4 Resource constraints and unclear ownership	23	8
T3. Emergent governance mechanisms (RQ3)	3.1 Governance structures (committees 15; role assignment 12)	27	8
	3.2 Governance processes (risk-tiered approval 16; human-in-the-loop 12; vendor due diligence 9)	37	9
T4. Governance maturity as a continuum (cross-cutting)	4.1 Reactive → structured → embedded	30	11
Total		286	

Table 3. Code-Document Table: quotations per participant by theme

Participant	T1 Perceptions	T2 Challenges	T3 Governance	T4 Maturity	Total	Maturity stage
P1 Banking	12	6	10	4	32	Embedded
P2 Banking	10	8	7	3	28	Structured
P3 Telco	9	10	3	2	24	Reactive
P4 Telco	9	9	6	3	27	Structured
P5 Healthcare	11	7	8	3	29	Structured
P6 Healthcare	10	6	9	4	29	Embedded
P7 Public	11	9	7	3	30	Structured
P8 Public	8	11	6	2	27	Structured
P9 Higher ed.	9	7	4	2	22	Reactive
P10 Manufacturing	6	8	1	2	17	Reactive
P11 Manufacturing	8	8	3	2	21	Reactive
Total	103	89	64	30	286	

Table 3 shows a clear pattern. Participants in *embedded* organizations (P1, P6) contributed proportionally more governance quotations (T3). Participants in *reactive* organizations (P3, P9, P10, P11) spoke mostly about challenges (T2).

Table 4. Strongest code co-occurrences (ATLAS.ti c-coefficient)

Code pair	Co-occurrences	c-coefficient
Risk-tiered approval ↔ Human-in-the-loop	11	0.42
Human oversight ↔ Transparency	17	0.38
AI literacy gap ↔ Vendor opacity	12	0.31
Unclear ownership ↔ Regulatory ambiguity	10	0.29
Amanah/integrity ↔ Fairness	7	0.21

Figure 2 presents the ATLAS.ti network view linking the four themes and their constituent codes. Relationship types follow ATLAS.ti conventions: *is part of*, *is associated with*, *contradicts*, and *is cause of*.

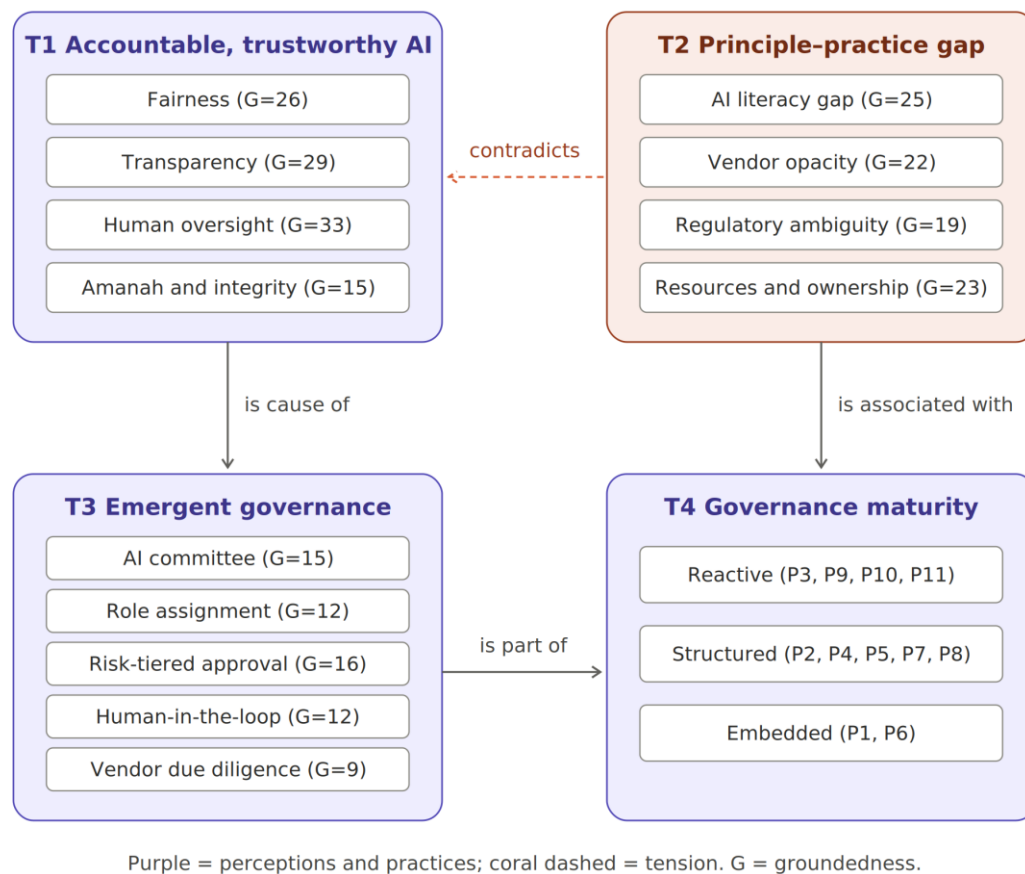


Figure 2. ATLAS.ti network view of themes, codes and relations

5.1 Theme 1: Responsible AI as accountable and trustworthy decision-making (RQ1)

All eleven managers defined responsible AI through its consequences for decisions rather than through technical properties. Their accounts converged on one question: *who answers for the outcome when AI is involved?*

Sub-theme 1.3: Human accountability and oversight (G = 33; 11/11)

This was the only sub-theme raised by every participant. Managers described accountability as something that cannot be delegated to a system, and several did so using regulatory or professional language.

“Our credit model scores about four thousand applications a month. The model recommends, but the approval authority is still human, and that is written in our credit policy. When the internal auditors or Bank Negara ask why a customer was rejected, I cannot say ‘the algorithm decided’. For me, responsible AI starts there: the machine can assist, but the accountability never moves to the machine.” (P1, Head of Risk Management, Banking)

“In the emergency department, the triage tool gives a priority score. Our rule is simple: the nurse must look at the patient, not just the screen. Early on, a junior nurse followed the score for a patient who looked stable on paper but was clearly deteriorating. We caught it, but that incident shaped how I understand this. AI supports clinical judgment. It is not clinical judgment.” (P5, Hospital Operations Manager, Healthcare)

Managers from less mature organizations also stressed oversight, but admitted it was more of an intention than a practice:

Interviewer: *Who checks the output of the predictive maintenance system?*

P10: *Honestly? The technician on shift. If the system says replace the bearing, usually we replace it. Nobody really questions it, because it’s usually right. But if it’s wrong one day, I’m not sure who is responsible: me, the vendor or the technician.* (P10, Operations Manager, Manufacturing)

Sub-theme 1.2: Transparency and explainability (G = 29; 10/11)

Transparency co-occurred strongly with human oversight ($c = 0.38$). Managers wanted *decision-level* explanations they could pass on to others, not algorithmic detail.

“I’m not a data scientist and I don’t need the mathematics. What I need is, when a patient’s family asks why their father was flagged as lower priority, the doctor can explain the main reasons in plain language. If we cannot explain it to the family, we should not be using it on the family.” (P6, Clinical Informatics Lead, Healthcare)

“Our chatbot answers thousands of public queries every week. We decided that it must always tell the citizen, ‘You are talking to an automated assistant,’ and give an option to reach an officer. People have the right to know whether a human or a machine is answering them, especially in government.” (P7, Deputy Director, Digital Services, Public sector)

Sub-theme 1.1: Fairness and non-discrimination (G = 26; 9/11)

Fairness concerns centred on recruitment, credit and eligibility decisions. The most detailed account came from human resources:

“We piloted the CV screening tool on about two thousand applications for graduate positions. When my team compared the shortlist with our manual process, the tool consistently ranked graduates from certain universities, mostly smaller regional ones, much lower. Nobody programmed that. It learned it from ten years of our own hiring decisions. That was a wake-up call. The bias was ours; the AI just made it faster and harder to see.” (P3, Human Resource Manager, Telecommunications)

“In anti-money-laundering, false positives are not neutral. If the model flags certain names or nationalities more often, those customers get their accounts frozen and have to

explain themselves. It's a fairness issue, even if technically the model is 'accurate'." (P2, Compliance Manager, Banking)

Sub-theme 1.4: Values-based framing — *amanah* and integrity (G = 15; 6/11)

Six participants explicitly grounded RAI in personal, organizational or religious values. This framing co-occurred with fairness ($c = 0.21$) and gave responsibility a moral dimension beyond compliance.

"Bagi saya, data rakyat itu amanah. Kalau kita guna AI dan data itu disalahgunakan atau keputusan itu tidak adil, itu bukan sekadar isu undang-undang, itu isu integriti dan kita akan dipersoalkan, bukan sahaja oleh audit, tetapi di akhirat nanti." [For me, citizens' data is a trust (*amanah*). If we use AI and that data is misused or the decision is unfair, it is not just a legal issue; it is an integrity issue, and we will be held accountable, not only by the auditors but in the hereafter.] (P7, Public sector)

"When I brief lecturers about generative AI, I don't start with the policy. I start with amanah in assessment. Marks are a trust between us and the student. If AI is used to grade or to produce work, that trust must be protected." (P9, Deputy Dean (Academic), Higher education)

5.2 Theme 2: The principle–practice gap (RQ2)

Every participant endorsed RAI principles, but most described an organization that was not yet able to act on them. In the network view (Figure 2), this theme *contradicts* Theme 1.

Sub-theme 2.1: Limited AI literacy among managers (G = 25; 9/11)

Managers described being formally accountable for systems they could not independently evaluate. Literacy co-occurred with vendor opacity ($c = 0.31$), which suggests the two problems reinforce each other.

"Every quarter I sign off on AI-related proposals: forecasting tools, generative AI for board reports, now a quality inspection system. To be honest, I rely on what the vendor and IT tell me. I don't know the right questions to ask. What is a 'good' accuracy? What is 'bias testing'? I'm accountable, but I'm not equipped." (P11, Chief Strategy Officer, Manufacturing)

Interviewer: *Has your organization provided any training on AI for managers?*

P8: *One webinar on ChatGPT, about one hour. Nothing on how to evaluate an AI system we are buying. Procurement officers are trained on price, specification and compliance, but not on algorithms.* (P8, Procurement Manager, Public sector)

Sub-theme 2.2: Vendor opacity and "black-box" systems (G = 22; 7/11)

This barrier was most acute in the public sector and in organizations buying off-the-shelf solutions.

"We asked the vendor for information about the training data and how they tested for bias. The answer was, 'It's proprietary; we can't disclose.' So how do we check? We are a government agency; we are supposed to be accountable to the public. In reality, we are trusting a brochure and a demo." (P8, Procurement Manager, Public sector)

"The chatbot runs on a large language model from an overseas provider. When it gives a wrong answer about a customer's bill, we can adjust the prompts, but we can't really see why it answered that way. We own the customer relationship, but we don't own the model." (P4, Head of Customer Analytics, Telecommunications)

Sub-theme 2.3: Regulatory ambiguity (G = 19; 7/11)

Participants were aware of the national AI governance and ethics guidelines but were uncertain about their status and how they relate to sector regulation and personal data protection law.

“The national guidelines are good principles, but they are voluntary. The first question from our legal team was, ‘Is this binding?’ When the answer is no, RAI competes with every other priority. When budgets are tight, voluntary things go to the bottom.” (P7, Public sector)

“Banking is heavily regulated, so we are used to model risk management. But for AI specifically, the expectations are still evolving. We look at the EU AI Act and at the NIST framework to guess where regulators here will go. It’s like preparing for an exam without the syllabus.” (P2, Compliance Manager, Banking)

Sub-theme 2.4: Resource constraints and unclear ownership (G = 23; 8/11)

Ownership problems co-occurred with regulatory ambiguity ($c = 0.29$). Without external clarity, responsibility tended to fall between departments.

“Everybody agrees responsible AI is important, but nobody owns it. IT says it’s a compliance matter. Compliance says it’s a technology matter. The business just wants the tool live next month. So it ends up in a meeting where everyone agrees and nobody acts.” (P4, Telecommunications)

“We are an SME. We don’t have a data science team, a legal team or an ethics committee. It’s me, the IT executive and the vendor. Frameworks from big organizations assume resources we simply don’t have.” (P10, Manufacturing)

5.3 Theme 3: Emergent governance mechanisms (RQ3)

Despite these barriers, most organizations had begun building governance. The mechanisms fell into **structures** and **processes**, and they were most developed in regulated sectors.

Sub-theme 3.1: Governance structures (G = 27; 8/11)

Seven participants described a cross-functional AI committee ($G = 15$), and six had named model or system owners ($G = 12$).

“We set up an AI and model governance committee two years ago. It includes risk, IT, legal, compliance, data protection, and one business head on rotation. Every new AI use case must pass through it before procurement or development. Each approved model has a named owner at senior manager level. If something goes wrong, we know whose phone to call.” (P1, Banking)

“Our AI working group is chaired by the Deputy Director General, not by IT. That was deliberate. When the chair is from IT, it becomes a technology project. When the chair is from the top management, it becomes an organizational responsibility.” (P7, Public sector)

Sub-theme 3.2: Governance processes (G = 37; 9/11)

The most common process was **risk-tiered approval** ($G = 16$), which co-occurred most strongly with **human-in-the-loop controls** ($c = 0.42$). Participants linked the level of risk directly to the level of human review.

“We classify every use case as low, medium or high risk. A chatbot answering general FAQs is low; we just need a basic review. Anything affecting a customer’s money, credit or legal status is high. High risk means a documented impact assessment, bias testing, a

human reviewer for adverse decisions, and re-validation every six months.” (P2, Banking)

“For imaging support, the rule is that the AI can prioritize the worklist but never issue a report. A radiologist must review every flagged case. We also track the disagreement rate between the AI and our radiologists every month. If it goes above our threshold, the system is paused for review.” (P6, Healthcare)

Vendor due diligence (G = 9) was an emerging practice, prompted by earlier negative experiences:

“After that experience with the vendor refusing to disclose information, we revised our tender template. Now there is a section on AI: data sources, bias testing, data residency, explainability, and the right for us to audit. Two vendors withdrew from the tender. I think that tells you something.” (P8, Public sector)

In higher education, governance focused on generative AI and academic integrity:

“We issued faculty guidelines on generative AI: when students may use it, how they must declare it, and that lecturers remain fully responsible for assessment decisions. AI detection tools are not used as the sole evidence of misconduct, because they are not reliable enough to punish a student.” (P9, Higher education)

5.4 Theme 4: Governance maturity as a continuum (cross-cutting)

When cases were compared in the Code-Document Table (Table 3), the organizations fell into three stages of maturity.

Reactive (P3, P9, P10, P11). These organizations had no formal RAI policy and acted in response to incidents or complaints.

“We only started talking about AI rules after students were caught submitting AI-generated assignments. Before that, there was nothing. We were reacting, not planning.” (P9, Higher education)

Structured (P2, P4, P5, P7, P8). These organizations had committees, risk classification and written policies, but applied them inconsistently.

“We have the policy and the committee, but, honestly, enforcement depends on the project. Big projects go through properly. Small pilots in business units sometimes bypass the process because people don’t know it applies to them.” (P4, Telecommunications)

Embedded (P1, P6). These organizations integrated RAI into the project lifecycle, procurement, training and monitoring.

“At first we were reactive; we only acted after a complaint. Now RAI review is a formal gate in our project lifecycle, the same as cybersecurity and data protection. No gate approval, no go-live. It’s not an extra step anymore; it’s simply how we work.” (P6, Healthcare)

Across cases, movement toward embedded governance was associated with three conditions: sector regulation (banking, healthcare), visible sponsorship from top management, and at least a basic level of AI literacy among approving managers. Both embedded organizations reported all three, while none of the reactive organizations reported more than one.

Discussion

This study set out to explore how managers perceive responsible AI, what challenges they face in implementing it, and what governance practices they adopt. The findings show that managers approach RAI mainly as a problem of accountable decision-making. They experience a clear gap between the principles they endorse and what their organizations can actually deliver. They respond to that gap with governance mechanisms that develop unevenly along a maturity continuum. This section discusses each research question in turn and then considers the study's theoretical and practical implications.

6.1 Managers' perceptions of responsible AI (RQ1)

The first research question asked how managers perceive and interpret responsible AI. Managers described RAI mainly through four ideas: accountability, transparency, fairness and human oversight. This broadly matches the convergence around transparency, justice and fairness, responsibility and non-maleficence that Jobin et al. (2019) found in their review of 84 guidelines, and the principles of the AI4People framework (Floridi et al., 2018). The emphasis, however, was different. Principles such as beneficence and the promotion of well-being, which feature prominently in the ethics literature, were rarely mentioned. Instead, managers concentrated on the principles that directly affect their own position: who signs off on a decision, who must explain it, and who answers for it when it goes wrong. This suggests that managers read RAI through the lens of their organizational role. It becomes less a set of moral aspirations and more a question of how to defend a decision to customers, regulators and auditors.

The centrality of human accountability and oversight is especially significant. Every participant raised it, and it co-occurred strongly with transparency. Managers did not ask for technical interpretability of models. They asked for explanations good enough to justify a decision to the people affected by it, such as a patient's family, a rejected loan applicant or a member of the public using a chatbot. This supports Arrieta et al.'s (2020) argument that explainability must be tailored to its audience, and it gives empirical weight to Shneiderman's (2020) claim that meaningful human control is a cornerstone of trustworthy AI. At the same time, the account of the predictive maintenance system in a reactive organization shows that oversight can exist in name only. When a system is usually right, people stop questioning it, and responsibility quietly moves from the human to the machine. This reinforces the concern of Mikalef et al. (2022) that the "dark side" of AI often arises from how AI is used in practice, not only from flaws in the technology.

A distinctive finding is the values-based framing of RAI through concepts such as *amanah*. Six managers, mainly in the public sector and higher education, described data and AI-supported decisions as a trust for which they are morally, and in some cases religiously, accountable. This framing rarely appears in the largely Western-centred RAI literature, which tends to ground ethics in secular principles and legal compliance (Hagendorff, 2020; Jobin et al., 2019). The finding matters for two reasons. First, it shows that RAI principles do not have to be "imported" as external rules. They can be connected to ethical traditions that employees already hold. Second, it offers a possible answer to the motivational weakness identified by Mittelstadt (2019) and Munn (2023), who argue that non-binding principles lack force. When responsibility is framed as a personal moral trust rather than a compliance requirement, it may carry more weight in everyday decisions. This is especially relevant in Malaysia, where the national guidelines are voluntary (Ministry of Science, Technology and Innovation, 2024).

6.2 Challenges in implementing responsible AI (RQ2)

The second research question addressed the challenges managers face. The findings strongly confirm the principle–practice gap that is well documented in the literature (Mittelstadt, 2019; Morley et al., 2020; Munn, 2023). Every participant endorsed RAI principles, yet most described organizations that lacked the skills, structures or incentives to act on them. The contribution of this study is to show how the gap is experienced from the managerial position. Earlier empirical studies focused mainly on data scientists and engineers (Madaio et al., 2020; Rakova et al., 2021). This study shows that managers face

a different version of the problem: they are formally accountable for AI systems, but they are not able to evaluate those systems themselves.

Limited AI literacy was the most widely shared barrier, and it co-occurred with vendor opacity. Together, these create what can be described as *accountability without understanding*. Managers sign off on systems whose data, logic and limitations they cannot assess. They therefore depend on the claims of vendors and technical staff, the same parties whose work they are supposed to oversee. This is consistent with Berente et al.'s (2021) argument that the autonomy, learning and inscrutability of AI create new challenges for management. It also extends Rakova et al.'s (2021) observation that unclear roles and weak incentives undermine responsible practice, showing that the same problem appears at the level of decision-makers, not only practitioners. The single one-hour webinar described by the procurement manager (P8) illustrates how far current training is from what managers actually need.

Vendor opacity reveals that, for many organizations, RAI is increasingly a procurement problem rather than a development problem. Public sector bodies and SMEs rarely build their own models. They buy them. When vendors refuse to disclose training data or bias testing on the grounds that the model is proprietary, the buying organization cannot carry out the kind of internal auditing proposed by Raji et al. (2020) or the ethics-based auditing described by Mökander and Floridi (2021). This creates an accountability gap in which the organization remains answerable to citizens and customers for a system it cannot inspect. The finding suggests that RAI frameworks designed around in-house development need to be adapted to the more common situation of organizations that acquire AI from third parties.

Regulatory ambiguity and unclear ownership completed the picture, and the two were closely linked in the co-occurrence analysis. When the legal status of RAI guidelines is uncertain, no single department sees RAI as its responsibility. It then falls between IT, compliance, legal and business units, a pattern captured in P4's description of "a meeting where everyone agrees and nobody acts". Managers in the banking sector coped with this uncertainty by using international frameworks such as the EU AI Act and the NIST AI RMF as a guide to future local expectations. This is an early sign that international instruments shape practice even outside their jurisdiction. For SMEs such as P10's organization, however, the challenge was more basic. Many existing frameworks assume a scale of resources, such as dedicated ethics committees and data science teams, that smaller organizations simply do not have. This points to the need for proportionate, lightweight governance models for smaller organizations.

6.3 Governance practices (RQ3)

The third research question examined the governance practices managers adopt. Despite the barriers described above, most organizations had begun to build governance. The practices fell into two groups: structural practices, such as cross-functional AI committees and named model owners, and procedural practices, such as risk-tiered approval, human-in-the-loop controls and vendor due diligence. This distinction aligns closely with the structural and procedural dimensions of the RAI governance framework proposed by Papagiannidis et al. (2025). It also matches several of the organization-level patterns catalogued by Lu et al. (2024). The findings therefore provide empirical support, from a non-Western setting, for frameworks that until now have been mostly conceptual.

Risk-tiered approval was the most common procedural practice, and it co-occurred most strongly with human-in-the-loop controls. Organizations did not apply the same level of scrutiny to every AI system. Instead, they matched the intensity of review to the potential impact on people, from a basic check for an FAQ chatbot to impact assessments, bias testing and periodic re-validation for credit decisions. This mirrors the risk-based logic of the EU AI Act (European Parliament & Council of the European Union, 2024) and the NIST AI RMF (National Institute of Standards and Technology, 2023). It also shows how managers turn abstract principles into concrete rules. The imaging example in healthcare, where the AI may prioritize a worklist but never issue a report and where disagreement rates are monitored monthly,

is a clear illustration of human oversight being designed into a process rather than left to individual discretion.

The findings also show important gaps in governance. First, the choice of who chairs AI committees matters. When the committee was led by top management rather than IT, RAI was treated as an organizational responsibility rather than a technical project. This supports Rakova et al.'s (2021) finding that leadership support is a key enabler. Second, relational governance was largely absent. This refers to practices that involve affected stakeholders, such as customers, patients, students and communities, in how AI is designed, evaluated or challenged. Papagiannidis et al. (2025) identify relational practices as a core dimension of RAI governance. Their near-absence here suggests that current practice remains inward-looking, focused on protecting the organization rather than giving voice to those affected by its AI systems. Third, vendor due diligence was still emerging. Its effect was nonetheless visible: when one agency added AI disclosure requirements to its tender template, two vendors withdrew. This suggests that procurement can act as a powerful lever for RAI.

6.4 Governance maturity as a continuum

The maturity continuum is the study's main theoretical contribution. Comparing cases in the Code-Document Table showed that organizations moved from *reactive* governance, triggered by incidents and complaints, to *structured* governance, with committees and policies that were applied inconsistently, and finally to *embedded* governance, in which RAI review is a standard stage gate in the project lifecycle. This responds to the call by Birkstedt et al. (2023) for empirical research on how AI governance is enacted inside organizations, a field they describe as fragmented and largely conceptual. It also adds a temporal dimension to the definition of organizational AI governance offered by Mäntymäki et al. (2022). Governance is not simply present or absent. It develops over time and can be assessed in stages.

The cross-case analysis identified three conditions associated with movement along the continuum: sector regulation, visible sponsorship from top management, and a basic level of AI literacy among approving managers. The two embedded organizations reported all three, while no reactive organization reported more than one. These conditions help explain why organizations that endorse the same principles can produce very different outcomes. Principles are necessary but not sufficient. What makes the difference is the combination of external pressure, internal leadership and managerial capability. This finding also links the four themes. Managers' perceptions (Theme 1) motivate governance (Theme 3), the principle–practice gap (Theme 2) slows it, and the balance between the two determines an organization's position on the maturity continuum (Theme 4).

6.5 Theoretical and practical implications

Theoretically, the study makes three contributions. First, it moves the focus of empirical RAI research from technical practitioners to managers. It shows that managers interpret RAI through the lens of accountability and that their main difficulty is being accountable without the means to understand. Second, it demonstrates that culturally grounded values such as *amanah* can serve as a bridge between abstract principles and everyday commitment, adding a non-Western perspective to a literature dominated by Western frameworks. Third, it proposes a governance maturity continuum that links perceptions, challenges and practices, offering a framework that future research can test and refine.

Practically, the study shows organizations where to focus their efforts. Improving managerial AI literacy, clarifying ownership and strengthening procurement are likely to have more effect than issuing further statements of principle. The maturity continuum also gives managers a simple diagnostic tool: they can locate their organization at a stage and identify which of the three enabling conditions is missing. These implications are developed further in the recommendations below.

7. Conclusion and Recommendations

This study shows that managers view responsible AI as a matter of accountable and trustworthy decision-making, but face a persistent principle–practice gap driven by limited AI literacy, vendor opacity, regulatory ambiguity and unclear ownership. Organizations are responding with emergent governance mechanisms, and their practices can be understood as a maturity continuum from reactive to embedded. Managers are therefore not merely recipients of AI ethics policy but key agents in translating principles into practice. Based on these findings, the following recommendations are offered for organizations and for policymakers.

7.1 Recommendations for organizations

First, organizations should invest in AI literacy for managers, not only for technical staff. The findings show that managers are formally accountable for AI systems they cannot evaluate. Training should therefore go beyond general awareness of tools such as ChatGPT. It should equip managers to ask critical questions before approving an AI system: what data it was trained on, how it was tested for bias, how accurate it is for different groups, how its decisions can be explained, and what happens when it fails. Short, role-specific programmes for approving managers, procurement officers and heads of department are likely to be more effective than one-off webinars. Organizations could also provide a simple checklist of questions that managers must answer before signing off on any AI proposal.

Second, organizations should assign clear ownership of responsible AI. The finding that RAI often falls between IT, compliance, legal and business units shows that shared responsibility can easily become no responsibility. Each organization should establish a cross-functional AI governance committee with a written mandate and, where possible, chaired by a member of top management rather than by IT, so that RAI is treated as an organizational matter. Every approved AI system should also have a named owner at senior manager level who is accountable for its performance, monitoring and any harm it causes. This makes it clear, in P1's words, "whose phone to call" when something goes wrong.

Third, organizations should adopt risk-tiered governance so that scrutiny is proportionate to impact. Not every AI system needs the same level of control, and applying heavy procedures to low-risk tools can make governance seem like a burden. Organizations should classify AI use cases into risk levels based on their potential effect on people's rights, money, health, education or legal status. Low-risk uses may need only a basic review, while high-risk uses should require a documented impact assessment, bias testing, a human reviewer for adverse decisions and regular re-validation. This approach is consistent with the EU AI Act and the NIST AI RMF, and it helps smaller organizations focus their limited resources where the risks are greatest.

Fourth, organizations should use procurement as a lever for responsible AI. Since many organizations buy rather than build AI, their greatest influence over RAI often comes at the point of purchase. Tender documents and contracts should require vendors to disclose training data sources, bias testing methods and results, data residency arrangements and the limits of their systems. Contracts should also give the buying organization the right to audit the system and to receive notice of significant model changes. The experience of the state agency in this study, where two vendors withdrew after AI disclosure requirements were introduced, shows that such requirements can filter out suppliers who are unwilling to be transparent.

Fifth, organizations should embed responsible AI into the full project lifecycle rather than treating it as a one-time approval. RAI review should become a formal stage gate, alongside cybersecurity and data protection, that every AI project must pass before going live. After deployment, organizations should monitor AI systems continuously, for example by tracking disagreement rates between AI outputs and human experts, complaint patterns and differences in outcomes across groups, and should be willing to pause a system when agreed thresholds are exceeded. Moving from reactive to embedded governance in this way reduces the risk that problems are discovered only after harm has occurred.

Sixth, organizations should connect responsible AI to their own values and culture. The findings show that some managers understand RAI as *amanah*, a trust for which they are personally accountable. Organizations can build on this by framing RAI in terms of integrity, trust and service in their policies, training and leadership communication, rather than presenting it only as a compliance requirement. Where appropriate, this may include drawing on religious and cultural values that employees already hold. Framing RAI in this way is likely to strengthen commitment and encourage staff to raise concerns, even when formal rules are unclear. Organizations should also begin to involve affected stakeholders, such as customers, patients and students, in how AI systems are evaluated and challenged, so that governance serves the people affected and not only the organization itself.

7.2 Recommendations for policymakers

For policymakers, the most pressing need is to reduce regulatory ambiguity. Managers in this study were uncertain whether the national guidelines on AI governance and ethics were binding and how they related to sector regulations and personal data protection law. Regulators should clarify this relationship, for example through official guidance that explains which obligations are mandatory, which are recommended, and how existing laws apply to AI. Clear signals from regulators would help organizations justify investment in RAI and reduce the tendency for voluntary measures to fall to the bottom of the priority list when budgets are tight.

Policymakers should also develop sector-specific guidance and practical tools, particularly for the public sector and for SMEs. General principles are difficult to apply without examples. Standard procurement templates with AI disclosure clauses, simple risk classification tools and model impact assessment forms would help organizations that lack specialist staff to put principles into practice. Government agencies, as major buyers of AI, can lead by example by adopting such templates in their own procurement, which would in turn encourage vendors to improve transparency across the market. Support for training programmes aimed at managers, rather than only at technical staff, would also address the literacy gap identified in this study.

Finally, policymakers should consider aligning national frameworks with international standards such as ISO/IEC 42001 and the risk-based approaches of the EU AI Act and the NIST AI RMF. Alignment would reduce uncertainty for organizations that operate across borders or depend on foreign vendors, and it would make it easier for Malaysian organizations to demonstrate trustworthy AI practice to international partners. A phased approach could be adopted, starting with voluntary certification and sector-specific requirements for high-risk uses, and moving towards clearer obligations as organizational capacity grows.

7.3 Limitations and future research

The study involved 11 managers in one national context, so findings are not statistically generalizable. Future research could test the maturity continuum quantitatively across a larger sample, conduct longitudinal case studies of organizations moving between maturity stages, and include the perspectives of employees and customers affected by AI decisions.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Acknowledgement

Finally, sincere appreciation is extended to everyone who contributed directly or indirectly to the successful completion of this study.

References

- Adibah, L. A., Hafizah, Z. (2021). The creativity practice of islamic education teachers in 21st century learning. *ASEAN Comparative Education Research Journal on Islam and Civilization*, 4(1), 40–54. <https://spaj.ukm.my/acerj/index.php/acer-j/article/view/67>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Birkstedt, T., Minkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7), 133–167. <https://doi.org/10.1108/INTR-01-2022-0042>
- Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597. <https://doi.org/10.1080/2159676X.2019.1628806>
- Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3), 328–352. <https://doi.org/10.1080/14780887.2020.1769238>
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Friese, S. (2019). *Qualitative data analysis with ATLAS.ti* (3rd ed.). SAGE.
- Guest, G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PLOS ONE*, 15(5), e0232076. <https://doi.org/10.1371/journal.pone.0232076>
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission.
- International Organization for Standardization. (2023). *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system*. ISO.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys*, 56(7), Article 173. <https://doi.org/10.1145/3626234>

- Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. <https://doi.org/10.1145/3313831.3376445>
- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, 26(13), 1753–1760. <https://doi.org/10.1177/1049732315617444>
- Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics*, 2(4), 603–609. <https://doi.org/10.1007/s43681-022-00143-x>
- Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and “the dark side” of AI. *European Journal of Information Systems*, 31(3), 257–268. <https://doi.org/10.1080/0960085X.2022.2026621>
- Ministry of Science, Technology and Innovation. (2024). *The national guidelines on AI governance and ethics*. MOSTI Malaysia.
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, 31(2), 323–327. <https://doi.org/10.1007/s11023-021-09557-8>
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, 3(3), 869–877. <https://doi.org/10.1007/s43681-022-00209-w>
- 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>
- Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1–13. <https://doi.org/10.1177/1609406917733847>
- Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), 101885. <https://doi.org/10.1016/j.jsis.2024.101885>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). ACM. <https://doi.org/10.1145/3351095.3372873>
- Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 7. <https://doi.org/10.1145/3449081>
- Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4), Article 26. <https://doi.org/10.1145/3419764>

Publisher: CLM Publishing Resources Malaysia



Open Access: This article is licensed under a [Creative Commons Attribution 4.0](#) International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third-party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Data Availability Statement: All relevant data are within the manuscript and its [Supporting Information](#) files.