Bridging the Gap: Effective Frameworks for Ethical AI Governance

A Systematic Literature Review

John Lee

A research component submitted to Auckland University of Technology

in partial fulfilment of the requirements for the degree of Master of Business (MBus)

2026

Faculty of Business, Economics and Law

Abstract

Artificial intelligence (AI) is increasingly embedded in organisational decision-making, reshaping how work is performed, how services are delivered, and how risks are identified and managed. Alongside these opportunities, however, AI introduces material ethical, legal, and reputational risks, particularly in high-impact contexts where automated or model-assisted judgements can affect people's rights, access to services, and life chances. In response, AI ethics principles and guidelines have proliferated globally; yet comparative reviews highlight substantial variation in scope, terminology, and enforceability, contributing to uneven translation from high-level commitments into operational organisational practice. This persistent "principles-to-practice" gap is widely recognised as a core limitation of contemporary AI ethics discourse, because principles alone rarely provide the procedural specificity, organisational incentives, and accountability mechanisms required to reliably shape decisions under real constraints.

This dissertation addresses that gap by synthesising evidence on organisational AI governance through a systematic literature review (SLR), reported in alignment with PRISMA 2020 guidance and informed by evidence-based approaches to management knowledge development. The study examines (1) how organisations implement and adapt governance frameworks to support ethical AI use, and (2) which governance approaches appear most effective for ensuring responsible AI use across organisational contexts. The synthesis identifies recurring patterns through which organisations operationalise ethical aspirations into structures, such as defined roles and oversight forums, processes, for example, risk and impact assessments across the AI lifecycle, and socio-technical practices, such as documentation, auditability, and monitoring. It further shows that governance effectiveness is context-dependent and increasingly shaped by regulatory momentum, including risk-based regimes that formalise differentiated obligations by system risk and use case. Overall, the dissertation concludes that robust AI governance is best understood as an organisational capability rather than a static framework: it requires enforceable accountability, fit-for-purpose operational controls, and continuous adaptation as technologies, organisational incentives, and societal expectations evolve.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

John Lee

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Associate Professor Marcus Ho. I am deeply grateful to him. His guidance was more than academic: it shaped not only the research itself, but how I learned to think as a scholar. Through moments of uncertainty and breakthrough alike, his rigour challenged me, his critique sharpened my understanding, and his encouragement arrived when I needed it most. I have grown immeasurably under his supervision.

Ko te manu e kai ana i te miro, nōna te ngahere;

ko te manu e kai ana i te mātauranga, nōna te ao.

This whakataukī has taken on deeper meaning for me through this work. I have come to understand that pursuing knowledge is not merely an individual achievement. It is a gift that expands one's place in the world, and one's capacity to contribute to it.

And finally, I think of my wife. Postgraduate study asks much: time, focus, resilience, and she gave me the space and support to offer all of that and more. Her love was not loud, but it was constant. In the late nights and difficult stretches, it was her quiet belief in me that reminded me why I began this work, and why it mattered to see it through. I carry her strength with me, and I am profoundly grateful.

In preparing this dissertation, I used Grammarly as an editing aid to support proofreading and to identify issues related to sentence structure, clarity, and grammar. I also used Microsoft Copilot in a limited technical capacity to assist with reformatting bibliographic details between referencing styles (for example, converting Vancouver-style entries to APA 7 format). EndNote was used as the primary reference management tool, and all citations and reference lists were checked and finalised by the author.

Ultimately, this work is dedicated to the belief that true governance is an act of stewardship. In seeking to bridge the gap between technological capability and ethical responsibility, we mirror a higher call to order, truth, and the service of the common good. S.D.G.

Chapter 1: Introduction

1.1 Background and Problem Statement

Artificial intelligence (AI) has moved rapidly from experimental capability to a routine component of organisational operations, influencing how decisions are made, how services are delivered, and how performance is measured. Contemporary evidence suggests that both analytical and generative AI are increasingly used across business functions, reflecting competitive pressure to improve productivity and accelerate innovation (Maslej et al., 2025; Singla et al., 2025). Yet the very features that make AI valuable at scale, such as automation, pattern recognition, and the capacity to influence decisions in complex environments, also mean that failures can be serious. When AI systems are used to support, shape, or automate decisions that affect people, the risks include discriminatory outcomes, opacity in judgement, privacy and security harms, and unclear accountability for downstream impacts (OECD, 2019a, 2019b; U.S. Department of Commerce, 2023). These risks are not merely technical; they intersect with organisational legitimacy, public trust, and social licence, particularly in high-impact domains.

AI-enabled automation creates not only external risks (for example, bias, privacy harms) but also internal governance risks, because employees' perceptions of technological change can directly shape organisational stability (Brougham & Haar, 2018, 2020). In Brougham and Haar (2018) STARA literature, employees' awareness that smart technologies, such as AI, robotics, algorithms may substitute for aspects of their work is associated with lower organisational commitment and career satisfaction, alongside higher cynicism, depressive symptoms, and turnover intentions (Brougham & Haar, 2018). These findings imply that ethical AI governance must extend beyond "responsible use" principles to include credible workforce assurance and change practices that protect trust, wellbeing, and retention during AI adoption (Brougham & Haar, 2018).

This dynamic is reinforced in multi-country evidence showing that perceived technological disruption increases job insecurity, which in turn predicts turnover intentions, even when job satisfaction is controlled indicating a distinct mechanism through which AI-related change can accelerate staff churn (Brougham & Haar, 2020). Perceived job mobility further conditions these relationships, suggesting that high demand labour markets may amplify the retention risks of poorly governed AI transitions (Brougham & Haar, 2020).

Viewed together, these studies position AI governance as a socio-technical change programme: effectiveness should be judged not only by compliance and harm reduction, but also by whether organisations can maintain social licence, capability, and workforce confidence as AI reshapes work (Brougham & Haar, 2017, 2018, 2020).

Accordingly, organisations face a governance challenge alongside the technical challenge of adoption. The central issue is not only whether AI systems can be implemented, but how they should be governed to ensure responsible, credible use over time. This governance imperative is sharpened by policy and regulatory momentum. Risk-based approaches to AI oversight are becoming embedded in statutory frameworks, exemplified by the European Union's Artificial Intelligence Act, which differentiates obligations according to system risk and intended use (European Union, 2024). In parallel, public-sector expectations around transparency and responsible use of algorithmic decision-making have been set out in initiatives such as Aotearoa New Zealand's Algorithm Charter, signalling that trustworthy use is increasingly framed as a governance responsibility rather than a voluntary aspiration (Statistics New Zealand, 2020). Taken together, these developments position ethical AI governance as a practical, organisational concern that spans compliance, risk management, corporate accountability, and organisational culture.

1.2 From Principles to Practice: Why Governance is Difficult

In response to AI-related harms and growing scrutiny, ethical principles, guidelines, and "responsible AI" frameworks have proliferated across governments, industry bodies, and academic communities. Comparative syntheses show partial convergence around high-level values, such as fairness, transparency, accountability, and respect for human rights, while also demonstrating divergence in terminology, priority-setting, and operational interpretation (Fjeld et al., 2020; Jobin et al., 2019). This proliferation has undoubtedly created a shared ethical vocabulary; however, it has also produced a practical organisational problem. The presence of principles does not guarantee that incentives, workflows, and decision rights will reliably align with those principles in day-to-day development and deployment, particularly where speed-to-market pressures, performance metrics, and commercial imperatives dominate.

A sustained critique in the literature is that principles alone are insufficient for ensuring ethical outcomes in organisational contexts. Mittelstadt (2019) argues that principles frequently lack the specificity and procedural "hooks" required to guide action under conditions of uncertainty, competing priorities, and organisational constraint. This critique helps explain why scholarship repeatedly returns to a familiar tension: organisations may publicly endorse ethical principles while struggling to build governance arrangements that meaningfully change behaviour, allocate responsibility, and prevent harm. The risk, therefore, is not only technical failure but governance failure, where ethical intent remains symbolic, ambiguous, or inconsistently implemented, thereby widening the gap between what organisations claim and what their AI systems actually do in practice.

1.3 Defining AI Governance in Organisational Settings

In this dissertation, AI governance refers to the institutional arrangements through which organisations direct, control, and monitor AI systems across their lifecycle. It encompasses the allocation of roles and responsibilities, the structuring of decision rights and oversight, and the establishment of processes for assurance, review, and continuous monitoring. Importantly, this definition treats governance as socio-technical: governance is not reducible to policy statements or compliance checklists, but is embedded in organisational routines, incentives, accountability structures, and technical practices across design, development, deployment, and post-deployment management. This framing aligns with contemporary guidance that positions trustworthy AI as a lifecycle challenge requiring sustained organisational capability. For example, the NIST AI Risk Management Framework emphasises that trustworthy AI outcomes depend on governance structures, contextual risk mapping, measurement and evaluation practices, and ongoing management responses (U.S. Department of Commerce, 2023). In organisational terms, this implies that effective governance is unlikely to be achieved through a single committee, a one-off risk review, or the publication of principles alone. Rather, governance requires an aligned system of organisational decision-making and operational controls that can scale with AI use, adapt to new model behaviours, and respond credibly to incidents and stakeholder expectations (OECD, 2019a, 2019b; U.S. Department of Commerce, 2023).

1.4 Research Purpose and Research Questions

Despite broad agreement that ethical AI governance is necessary, the evidence base remains fragmented regarding what organisations actually do to implement governance frameworks and which approaches appear to be most effective in practice. Mittelstadt (2019) alludes to a growing number of toolkits, maturity models, and prescriptive recommendations, yet practical insights remain uneven regarding implementation patterns, the mechanisms through which governance gains traction, and the criteria by which effectiveness should be assessed under real organisational constraints. This creates a practical dilemma: organisations must design governance under conditions of uncertainty, balancing ethical commitments against innovation incentives, capability constraints, and shifting regulatory expectations.

This dissertation addresses that gap by investigating organisational governance for ethical AI through two research questions:

> **RQ1:** How do organisations implement and adapt governance frameworks to ensure the ethical use of AI technologies?
>
> **RQ2:** What are the most effective governance frameworks for ensuring responsible AI use in organisations?

Together, these questions illuminate the organisational "how" of governance implementation and the evaluative question of "what works", while recognising that effectiveness is likely contingent on sectoral risk exposure, organisational maturity, and institutional context (European Union, 2024; OECD, 2019b).

1.5 Methodological Overview and Rationale

To address these research questions, the dissertation adopts a systematic literature review (SLR) methodology. An SLR is well-suited to an interdisciplinary and rapidly evolving topic because it provides a transparent, replicable approach to identifying, screening, and synthesising evidence across diverse empirical and conceptual contributions. The review is reported in accordance with PRISMA 2020 guidance, strengthening methodological transparency through clear reporting of search processes, inclusion criteria, and synthesis logic (Page et al., 2021). The rationale also aligns with evidence-informed approaches in

management scholarship, which argue that systematic synthesis can consolidate fragmented knowledge and provide a stronger basis for theory development and practice guidance in complex organisational domains (Tranfield et al., 2003).

This methodological choice reflects the dissertation's substantive intent: rather than describing one organisational setting in depth, the study synthesises patterns of governance across contexts to identify recurring mechanisms, tensions, and effectiveness claims. By integrating evidence across the literature, the dissertation is positioned to clarify how organisations operationalise ethical commitments, where implementation frays under organisational pressure, and which governance features appear most consistently associated with credible accountability and sustained responsible practice.

1.6 Significance and Expected Contributions

This dissertation contributes to the ethical AI governance literature in three ways. First, it consolidates evidence on how organisations translate ethical commitments into concrete governance mechanisms, making visible the organisational "plumbing" that is often under-specified in principles-based discourse (Mittelstadt, 2019). Second, it advances an evaluative lens on governance effectiveness by synthesising the features that the literature associates with operational traction, accountability, and sustained oversight over the AI lifecycle (OECD, 2019a; U.S. Department of Commerce, 2023). Third, it clarifies the role of context, particularly sectoral risk profiles and regulatory momentum, in shaping governance design choices, thereby cautioning against uncritical transfer of "best practice" templates across materially different environments (European Union, 2024; OECD, 2019b).

The practical importance of these contributions is amplified by the pace of AI diffusion and by rising expectations for demonstrably responsible practice. As legal obligations, stakeholder scrutiny, and reputational risk intensify, organisations increasingly require governance that is not only ethically well-intentioned but operationally credible, auditable, and resilient. Similarly, policy makers benefit from understanding how organisations interpret and implement governance obligations in practice, and where additional clarity, enforcement, standards, or assurance regimes may be needed to close recurring gaps between principles and outcomes.

1.7 Structure of the Dissertation

The dissertation is organised to move from conceptual foundations to systematic synthesis and interpretive discussion. Chapter 2 reviews the literature on AI ethics and organisational governance, including the proliferation of principles and the recurrent problem of translation into practice (Fjeld et al., 2020; Jobin et al., 2019; Mittelstadt, 2019). Chapter 3 outlines the SLR methodology, detailing the search strategy, screening process, and synthesis approach in alignment with PRISMA 2020 reporting guidance (Page et al., 2021). Chapter 4 presents the findings thematically, highlighting the mechanisms through which organisations implement and adapt governance across contexts. Chapter 5 then interprets these findings in relation to the literature and the research questions, drawing out implications for organisational governance design under evolving regulatory expectations and risk-based oversight regimes (European Union, 2024; U.S. Department of Commerce, 2023). Lastly, Chapter 6 wraps up by putting together the answers to the two research questions. RQ1: How do organisations implement and adapt governance frameworks to ensure the ethical use of AI technologies, and RQ2: What are the most effective governance frameworks for ensuring responsible AI use in organisations, and outlining implications, recommendations, and future directions for research and practice.

Chapter 2: Literature Review

2.1 Introduction

Artificial intelligence (AI) has emerged as a critical determinant of organisational innovation, productivity, and competitiveness, particularly within high-stakes sectors such as healthcare, finance, logistics, and education (Hyiamang & Liu, 2025). However, as algorithmic capabilities accelerate, the prevailing organisational imperative has shifted: the central challenge is no longer the technical feasibility of adoption, but the establishment of robust governance frameworks to ensure its responsible deployment (Jobin et al., 2019; OECD, 2019a). This imperative for governance arises from a growing body of ethical risks, including bias and discrimination (Ryan & Stahl, 2020), algorithmic opacity (Fjeld et al., 2020), privacy violations (Jobin et al., 2019), and persistent questions of accountability and responsibility (Khan et al., 2022). As AI systems increasingly influence decisions that materially affect human lives, organisations face mounting pressure to design, implement, and adapt robust ethical governance frameworks.

Governance, in the context of organisational artificial intelligence (AI), can be understood as the constellation of structures, processes, and shared norms through which organisations steer the responsible design, deployment, and oversight of AI systems (Butcher & Beridze, 2019). This definition foregrounds governance as more than compliance activity or technical risk management alone: it encompasses how decision rights are allocated, how accountability is enacted, and how ethical commitments are operationalised across the AI lifecycle and within organisational contexts. Yet, despite the rapid proliferation of frameworks, guidelines, and "responsible AI" programmes, the literature indicates that governance remains a contested and unevenly stabilised concept in practice, with significant variation in what counts as governance, who is responsible for it, and how it is evidenced as effective (Mittelstadt, 2019; Munn, 2023).

Three recurring critiques in the literature provide the core rationale for the research questions that guide this review. First, there is limited consistency in definitions and scope across the field. Although many sources use similar language (for example, ethics, accountability, transparency, and fairness), the underlying constructs are not uniformly specified, and governance is alternately treated as a set of principles, a compliance function, an organisational capability, or a lifecycle management system (Butcher & Beridze,

2019; Mittelstadt, 2019). This definitional instability makes it difficult to compare studies, accumulate evidence, or determine whether organisations are implementing substantively similar governance frameworks. Second, the literature repeatedly identifies a persistent principles-to-practices gap. Systematic reviews show that organisations commonly endorse high-level ethical principles yet struggle to translate these commitments into operational mechanisms that shape day-to-day design and deployment decisions (Fjeld et al., 2020; Jobin et al., 2019; Morley et al., 2020). Third, even where governance mechanisms are described, their practical effectiveness is often contested because implementation is highly context-dependent and evidence of sustained outcomes is reported unevenly, with limited agreement on what evaluative criteria should be used to judge success (Mittelstadt, 2019; Munn, 2023).

Against this backdrop, this chapter critically examines the literature on the ethical governance of AI in organisations to address two guiding research questions:

> **RQ1**: How do organisations implement and adapt governance frameworks to ensure the ethical use of AI technologies?
>
> **RQ2**: What are the most effective governance frameworks for ensuring responsible AI use in organisations?

The review draws upon conceptual work, systematic literature reviews, empirical studies, and sector-specific frameworks to provide a holistic account of current knowledge. Consistent with the dissertation's systematic literature review design, the chapter synthesises published evidence rather than reporting primary organisational data. It also highlights the importance of practice and empirical findings in bridging the gap between principles and practices.

This chapter consolidates conceptual scholarship, systematic literature reviews, empirical research, and sector-specific frameworks to provide an integrated account of ethical AI governance. In keeping with the systematic literature review approach adopted in this dissertation, the discussion synthesises peer-reviewed evidence and authoritative governance frameworks rather than presenting primary organisational data. Importantly, it foregrounds the role of practice-oriented insights and empirical findings in clarifying how high-level ethical commitments can be translated into operational governance arrangements. The chapter is structured to align with the following sections: it first outlines the conceptual foundations of ethical AI

governance (Section 2.2) and clarifies what is meant by AI governance (Section 2.3), before examining ethical principles as foundational reference points (Section 2.4). It then compares the respective functions and limitations of soft and hard law approaches (Section 2.5), and synthesises the major ethical challenges associated with AI (Section 2.6). Building from these foundations, the review analyses organisational responses and emergent practices (Section 2.7) and evaluates prominent governance frameworks for ethical AI (Section 2.8). Lastly, it looks at real-world evidence on implementation (Section 2.9) to find the tensions, gaps, and implications that led to this study.

2.2 Conceptual Foundations of Ethical AI Governance

Several strands of literature highlight why ethical AI governance has become an important and time-sensitive organisational concern, and these strands point to distinct (and sometimes competing) diagnoses of the governance problem. First, global reviews of AI ethics guidelines document the rapid proliferation of ethics principles and guidance, alongside substantial variation in scope, terminology, and enforceability (Farooq et al., 2021; Jobin et al., 2019). In practice, this variation is not merely semantic: it reflects a deeper tension between aspirational ethics statements that articulate broad values and action-guiding governance instruments that specify concrete responsibilities, processes, or compliance expectations. As a result, the growth in guidelines signals both heightened recognition of AI-related risks and continuing uncertainty about how organisations should translate normative commitments into operational controls.

Second, scholarly work that analyse the “principles-to-practice” gap argues that principles alone rarely deliver reliable governance outcomes without accompanying organisational mechanisms (Mittelstadt, 2019; Munn, 2023). This literature highlights a core debate: whether ethics programmes function primarily as internal capability-building efforts that meaningfully reshape decision-making, or whether they risk becoming symbolic “commitments” that sit alongside unchanged incentives, workflows, and accountability arrangements (Attard-Frost et al., 2023; Munn, 2023). Third, empirical research on organisational adoption further indicates that governance is not simply a conceptual aspiration but a practical challenge with real consequences for accountability across the AI lifecycle. For example, Raji et al. (2020) illustrate how the absence of systematic documentation, review, and escalation pathways can erode traceability and weaken accountability once systems move from development into deployment and maintenance. Relatedly, Morley

et al. (2020) show that many governance proposals remain difficult to operationalise when organisations lack concrete processes, role clarity, and routine monitoring practices.

Practitioner-oriented evidence reinforces these concerns while adding a managerial lens on organisational readiness and maturity. Capgemini Research Institute (2020) for instance, reports a mismatch between high levels of stated commitment to ethical AI and relatively limited uptake of routine controls such as audits or formal governance structures. Likewise, practitioner analyses of corporate governance arrangements, such as Microsoft's AETHER Committee and SAP's use of external AI ethics advisors illustrate that organisations are experimenting with oversight models that combine technical, legal, and ethical perspectives (International Institute for Management Development, 2023). Yet, consistent with critical scholarship, the presence of committees or advisors does not, by itself, resolve questions of authority, enforcement, or integration into everyday delivery practices; where oversight bodies are weakly empowered or poorly embedded, they may struggle to influence product timelines, risk decisions, or mitigation resourcing (Attard-Frost et al., 2023; Munn, 2023).

Ultimately, the research underscores that external forces progressively transform ethical enquiries into governance mandates. Risk-based regulatory approaches, most prominently the European Union (2024) AI Act raise the cost of inadequate governance by introducing obligations and scrutiny that can translate into legal exposure and reputational harm (Butcher & Beridze, 2019; Qureshi et al., 2024; Veale & Zuiderveen Borgesius, 2021). In Aotearoa New Zealand, the public sector's Algorithm Charter similarly frames algorithm use through expectations of transparency, accountability, bias mitigation, and Te Tiriti o Waitangi obligations, reinforcing that governance is also a matter of social licence and public trust (Statistics New Zealand, 2020). Taken together, these strands converge on a clear implication: ethical AI governance matters because it mediates organisational exposure to ethical, legal, and reputational risks; shapes whether AI systems reinforce or mitigate social inequities; and determines whether AI deployment remains aligned with societal values and evolving regulatory expectations. In this dissertation, ethical AI governance is as a result positioned as a central organisational capability, requiring durable structures, enforceable processes, and continuous monitoring rather than a peripheral ethical add-on.

2.3 Defining AI Governance

AI governance is widely recognised as an evolving and contested concept, with no single definition that commands universal agreement across disciplines, sectors, or policy arenas (Butcher & Beridze, 2019; Mäntymäki et al., 2022). Across this literature, scholars emphasise different but complementary dimensions of how emerging technologies such as AI are governed, including (i) the allocation of decision rights and accountability, (ii) the institutional design of oversight and assurance mechanisms across the AI lifecycle, and (iii) the organisational translation of ethical principles into routine practices, processes, and tools (Birkstedt et al., 2023; Mäntymäki et al., 2022). In response to this definitional plurality, this dissertation adopts a working definition that is fit for purpose: it treats organisational AI governance as a coherent system of interlinked mechanisms through which organisations align AI use with strategy and values, meet legal obligations, and operationalise the ethical principles they claim to uphold (Mäntymäki et al., 2022). At the same time, contemporary scholarship suggests an emerging shift in emphasis away from governance framings centred primarily on technical compliance or risk minimisation, a dominant orientation in many organisational control and risk-management approaches and towards broader accounts that foreground social legitimacy, public value, and sustained ethical oversight (European Commission, 2019; Floridi et al., 2018; U.S. Department of Commerce, 2023). This broader framing is reinforced by critiques that warn against treating high-level principles as sufficient evidence of effective governance, noting that principles often mask deep normative disagreement and routinely fail to translate into practice without robust accountability arrangements and professional norms (Mittelstadt, 2019). Within this tradition, sociotechnical perspectives explicitly position AI governance as inseparable from the social, political, and cultural environments in which AI is developed and deployed and argue that governance should be anchored in substantive ethical aims, such as promoting human flourishing rather than reduced to narrow procedural compliance (Stahl, 2021).

Complementing this ethical and normative orientation, Butcher and Beridze (2019) describe AI governance as the constellation of global initiatives, principles, policies, standards, and regulatory proposals designed to steer AI development. Their analysis highlights the interaction between soft-law and hard-law instruments and depicts governance as an evolving global ecosystem of formal and informal rules. This literature

highlights the increasing institutionalisation of AI governance within international policymaking, regulatory experimentation, and transnational standard-setting efforts.

Taken together, these perspectives frame AI governance as a macro-level construct. They articulate universal, institutional, and general principles that operate at societal or global scales, shaping the broader normative and regulatory environment in which organisations must function. In this sense, they provide a conceptual foundation for understanding *what* AI governance seeks to achieve and *why* governance is necessary, positioning responsible AI as a collective societal project rather than an organisation-specific technical procedure. This distinction is analytically significant for this dissertation, as the research questions concern how organisations implement and adapt governance frameworks within these external normative environments.

However, translating these high-level principles into organisational practice requires more granular conceptualisations. As a result, a second stream of scholarship defines AI governance as a subset of corporate and IT governance. Drawing on these traditions, authors emphasise decision rights, accountability structures, and control mechanisms specific to AI systems, nested within broader organisational governance arrangements (Capgemini Research Institute, 2020; International Institute for Management Development, 2023). In this view, AI governance concerns who decides which systems are developed or deployed, how risks are assessed and mitigated, and how oversight is enacted over time. This meso- and micro-level orientation focuses on how organisations operationalise ethical and regulatory expectations within their internal structures, cultures, and processes.

A further technical orientation defines AI governance in terms of procedures and tools that ensure AI systems are fair, transparent, robust, and auditable. For example, Raji et al. (2020) foreground algorithmic auditing as a core governance mechanism, while Morley et al. (2020) catalogue methods are designed to translate high-level ethical principles into concrete organisational and technical practices. These perspectives highlight the practical challenges of embedding responsible AI principles into system design, development, and evaluation.

Together, these organisationally focused perspectives illustrate the shift from broad normative aspirations to the operational realities of governance, where principles must be interpreted, prioritised, and enacted within

the specific constraints and incentives of organisational life. This conceptual progression from macro-level institutional visions to micro-level organisational mechanisms provides the analytical scaffolding for examining how organisations implement and adapt AI governance frameworks, as explored in the subsequent chapters.

As such, the literature increasingly supports the idea that AI governance is multi-dimensional (Cath, 2018; Gasser & Almeida, 2017), spanning (1) institutional and normative arrangements, such as ethical frameworks and broad regulatory norms (Dafoe, 2018; Floridi et al., 2018). (2) Organisational structures and processes, including internal oversight committees and risk management workflows (International Organization for Standardization, 2022; Mökander et al., 2021) and (3) technical mechanisms, which operationalise these norms through audits, documentation, and explainability tools (Mitchell et al., 2019; Raji et al., 2020).

For the purposes of this dissertation, and in direct support of RQ1 and RQ2, AI governance is defined as "the set of institutional, organisational, and technical arrangements through which organisations direct, control, and oversee the design, deployment, and ongoing use of AI systems so that they align with ethical principles, legal requirements, and organisational values". The above definition is adopted because it recognises the multi-level nature of AI governance (institutional, organisational, and technical) (Cath, 2018; Dafoe, 2018), foregrounds organisational decision-making and responsibility, specifically the mandate to direct and oversee (International Organization for Standardization, 2022) which is central to the research questions, and explicitly links governance to both normative goals (ethical principles, societal values) and formal constraints (law, regulation) (European Commission, 2019; Floridi et al., 2018).

2.4 Ethical Principles as Foundations

A consistent theme in the literature is convergence around a core set of ethical principles. Fjeld et al. (2020) analysed 84 ethical guidelines and identified repeated emphasis on fairness, accountability, transparency, privacy, and human rights. Jobin et al. (2019) similarly document global patterns of convergence, though with regional variation. For example, stronger rights-based discussion in European guidelines and greater emphasis on innovation and economic competitiveness in US-based documents. This divergence reflects different regulatory traditions and institutional priorities, indicating that ethical AI governance is shaped by

societal context as well as technical risk (Veale & Zuiderveen Borgesius, 2021). Farooq et al. (2021) and Attard-Frost et al. (2023) synthesise these guidelines and show that, despite differences in terminology and emphasis, most frameworks converge on a small set of recurring principles, often framed around transparency, justice and fairness, non-maleficence, responsibility, and privacy. These principles are often treated as a normative anchor for AI governance, providing high-level criteria against which organisational practices can be evaluated by internal governance functions (e.g., ethics committees and risk teams), external auditors and regulators, and affected stakeholders.

However, the literature is clear that principles alone are insufficient. First, key concepts such as "fairness" are contested: they can be put in place as demographic parity, equal opportunity, calibration, or other formal criteria, each with different technical and ethical implications (Khan et al., 2022). Second, transparency is constrained by the inherent complexity and opacity of many AI systems, sometimes referred to as the "black box problem" (Burrell, 2016). Third, tensions arise between principles. For example, enhancing transparency may conflict with privacy or intellectual property concerns. Therefore, Mittelstadt (2019) argues that, due to these issues, principles cannot guarantee ethical AI in themselves. They need to be supported by governance structures, processes, and tools that guide trade-offs and make explicit how principles are interpreted and applied in specific organisational contexts. This insight directly motivates the focus of this dissertation on governance frameworks as mechanisms for implementing and adapting principles in practice.

2.5 Comparisons between Soft and Hard Laws

Scholars often distinguish between soft law voluntary codes of conduct, industry guidelines, standards, and best-practice frameworks and hard law binding regulations backed by enforcement mechanisms (Butcher & Beridze, 2019). Historically, AI governance has been dominated by soft law instruments, reflecting the rapid pace of technological change and the relative slowness of formal regulatory processes.

Critics question whether soft law is sufficient. Mittelstadt (2019) and Munn (2023) argue that many guidelines are "toothless", easily co-opted for publicity purposes without leading to meaningful change in organisational behaviour. This reflects broader concerns about ethics washing (Attard-Frost et al., 2023; Farooq et al., 2021). Recent regulatory developments, particularly the European Union's AI Act, signals a shift towards hard law. The AI Act introduces a risk-based approach, categorising AI systems into

unacceptable, high, limited, and minimal risk, and imposing binding obligations on high-risk systems around documentation, monitoring, transparency, and human oversight (European Union, 2024; Veale & Zuiderveen Borgesius, 2021). However, commentators also warn that overly burdensome regulation may stifle innovation and disproportionately affect small and medium enterprises (SMEs) (Veale & Zuiderveen Borgesius, 2021).

In the New Zealand context, the Manage My Health cyber incident illustrates how *hard-law* mechanisms, including urgent injunctive relief can operate alongside privacy-regulatory guidance and executive oversight to constrain downstream harms and stabilise public trust following a high-impact data breach ("Manage My Health Ltd v Unknown Defendants [2026] NZHC 2," 2026; Office of the Privacy Commissioner, 2026b).

The literature increasingly points towards hybrid models that combine soft-law flexibility with hard-law enforceability, seeking to capture the benefits of both approaches (Butcher & Beridze, 2019; Cath, 2018; Veale & Zuiderveen Borgesius, 2021). This hybrid logic underpins many emerging governance frameworks and is directly relevant to assessing which frameworks are “most effective” for responsible AI use (RQ2).

2.6 Ethical Challenges in AI

2.6.1 Algorithmic bias and discrimination

Algorithmic bias is one of the most persistent ethical challenges associated with AI. Bias arises when training data reflect historical inequalities or when model design embeds normative assumptions, resulting in discriminatory outcomes against vulnerable groups (Ryan & Stahl, 2020). A widely cited study by Buolamwini and Gebru (2018) exposed significantly higher error rates for darker skinned women in commercial facial recognition systems, illustrating how model performance can vary systematically across demographic groups. In Aotearoa New Zealand, the Privacy Commissioner’s inquiry into Foodstuffs North Island’s live facial recognition trial explicitly noted known bias and accuracy issues for people of colour and flagged potential impacts for Māori, Pasifika, Indian, and Asian shoppers (Office of the Privacy Commissioner, 2025b). Media reporting on such deployments further illustrates that biometric AI in public-facing settings is socially contested and places governance decisions under heightened public scrutiny (Radio New Zealand, 2024). Collectively, these findings demonstrate that AI can reproduce and amplify existing structural inequities if not carefully governed.

2.6.2 Transparency and explainability

The opacity of many AI models makes it difficult for stakeholders including affected individuals, regulators, and even developers to understand how decisions are made, to meaningfully contest outcomes, or to assess whether a system meets legal and ethical requirements (Burrell, 2016; Pasquale, 2015). The lack of explainability undermines accountability, as it becomes unclear who is responsible when AI systems cause harm or generate contested decisions. Sargiotis (2024) point out the challenge of providing explanations that are both technically meaningful and understandable to non-expert stakeholders.

2.6.3 Privacy and data integrity

AI systems depend on large volumes of data, often including personal information, raising concerns about privacy, consent, surveillance, and potential misuse (Jobin et al., 2019). Weak data governance can lead to compromised data integrity and quality, which in turn undermines model performance and fairness. The rise of generative AI intensifies these issues, as models may inadvertently memorise and reproduce sensitive data, blurring boundaries between training data and outputs. Effective governance must therefore integrate AI ethics with existing data protection regimes, ensuring that data collection, processing, and retention practices are aligned with both legal requirements and ethical expectations. In Aotearoa New Zealand, the Biometric Processing Privacy Code 2025 illustrates the growing specificity of privacy expectations for biometric and AI-enabled identification practices (Office of the Privacy Commissioner, 2025a).

2.6.4 Accountability deficits

Responsibility for AI outcomes is often distributed across developers, managers, vendors, and end-users, creating what Matthias (2004) calls a “responsibility gap”. Organisations have experimented with advisory boards and ethics committees as mechanisms to address this, but empirical studies suggest that robust accountability mechanisms remain rare and unevenly implemented (Attard-Frost et al., 2023; Munn, 2023). The challenge for governance frameworks is thus to clarify who is accountable for which decisions across the AI life cycle, and to ensure there are meaningful consequences and remediation pathways when harms occur.

2.7 Organisational Responses and Practices

Many organisations respond to ethical AI risks by adopting high-level codes of conduct aligned with widely cited principles frameworks and by establishing ethics boards or advisory groups to review higher-risk applications (Farooq et al., 2021; Fjeld et al., 2020). While these structures indicate growing recognition that ethical AI requires formal oversight, their practical influence varies substantially with mandate clarity, independence, resourcing, and proximity to decision-making power (Attard-Frost et al., 2023). In addition, such arrangements are more feasible in large, well-resourced firms, whereas smaller organisations often need proportionate governance mechanisms that fit their capacity and risk profile rather than replicating committee-heavy models (Capgemini Research Institute, 2020; Morley et al., 2020).

2.7.1 Codes of conduct and advisory boards

Many organisations have adopted high-level AI ethics principles or codes of conduct, often aligned with widely cited frameworks (Farooq et al., 2021; Fjeld et al., 2020). Some have also implemented ethics boards and advisory groups. For example, Microsoft's AETHER Committee (Accountability, Ethics, and Transparency in AI) provides internal review of high-risk AI projects, while SAP engages external experts to advise on AI ethics (International Institute for Management Development, 2023).

These initiatives signal recognition of the need for oversight and cross-functional deliberation. However, empirical analyses suggest that the influence of such bodies varies widely depending on their mandate, resourcing, and organisational position (Attard-Frost et al., 2023). Moreover, these structures are most readily established in large, well-resourced organisations; smaller firms and many resource-constrained organisations may instead require proportionate governance mechanisms that fit their capacity and risk profile (Capgemini Research Institute, 2020; Morley et al., 2020). This suggests that merely establishing committees is insufficient; rather, how they are integrated into governance structures is crucial.

2.7.2 Audits and risk assessments

Another set of organisational responses involves audits and risk assessments. Raji et al. (2020) propose an end-to-end framework for internal algorithmic auditing, emphasising documentation, process traceability, and stakeholder engagement. However, evidence suggests uneven adoption in practice: in a multi-country

executive survey, only 46% of organisations reported that the ethical implications of current AI systems were independently audited (Capgemini Research Institute, 2020). Regulatory experiments illustrate both the promise and limitations of mandated audits. New York City's Local Law 144 requires employers to obtain an annual independent "bias audit" for automated employment decision tools and publicly post audit-related information alongside candidate notices (New York City Department of Consumer and Worker Protection, 2021). Yet an empirical assessment of employer facing transparency found very low observable posting of audit reports and notices, and argues that "null compliance" can arise because employers retain substantial discretion over whether a system is considered in scope, while the regime relies heavily on transparency and end-user accountability rather than robust verification (Wright et al., 2024). Concerns about scope-narrowing are also visible in the rulemaking record: Hickok (2023) argues that definitional qualifiers can create loopholes that "significantly narrow the scope and intent" of the law. Taken together, this case illustrates that even when audit requirements exist, governance outcomes depend on precise scoping, independent verification, and credible oversight mechanisms (Hickok, 2023; Wright et al., 2024).

In healthcare settings, the rapid uptake of AI "scribe" tools provides a concrete illustration of how privacy and governance risks emerge alongside operational benefits. In a New Zealand primary care survey, Ballantyne et al. (2025) found that only 66% of respondents had read the scribe tool's terms and conditions and only 59% reported seeking patient consent, while concerns included data security and the possibility of data leaving New Zealand. In parallel, professional expectations are being formalised: the Medical Council of New Zealand's draft statement explicitly includes scribing tools within scope and emphasises that AI must be used responsibly in ways that prioritise patient safety and privacy, and that clinicians remain accountable for decisions and actions (Medical Council of New Zealand, 2025). From a legal and organisational governance perspective, the Office of the Privacy Commissioner (2024) clarifies that when personal information is stored or processed by a 3rd party provider, the originating organisation remains responsible, and that heightened risks arise where providers may use information for their own purposes (including as AI training data) or where cross-border disclosures may occur. These obligations sit within sector-specific privacy rules: the Health Information Privacy Code 2020 governs how health agencies collect, use, hold, and disclose identifiable health information and includes explicit constraints relevant to disclosure and offshore

disclosure, reinforcing the need for privacy impact assessment and vendor governance as part of AI tool adoption (Office of the Privacy Commissioner, 2020).

2.7.3 Ethics washing and symbolic governance

A critical strand of the literature warns that some corporate AI ethics initiatives are largely symbolic. Attard-Frost et al. (2023) and Munn (2023) document how organisations may adopt ethics statements and publish guidelines while continuing to engage in questionable practices, a phenomenon frequently labelled "ethics washing". Farooq et al. (2021) similarly note that many guidelines lack mechanisms for enforcement or monitoring. These findings reinforce the importance of moving from principles and public commitments towards robust governance arrangements that shape everyday decision-making.

2.7.4 Incident response and breach management

From an organisational practice perspective, the organisation's public incident updates, considered alongside communications from the National Cyber Security Centre and the Office of the Privacy Commissioner, provide a practice-facing account of how governance expectations are enacted through containment, notification, coordination, and remediation during an evolving cyber event (Manage My Health, 2026; National Cyber Security Centre, 2026; Office of the Privacy Commissioner, 2026d).

2.8 Governance Frameworks for Ethical AI

This section reviews key governance frameworks proposed in the literature, focusing on their underlying logics and implications for organisational practice.

2.8.1 Principles-based frameworks

Early governance work is dominated by principles-based frameworks. Butcher and Beridze (2019) survey global initiatives and find heavy reliance on voluntary principles such as fairness, transparency, accountability, and privacy. Floridi et al. (2018) propose the AI4People framework, which articulates a set of principles and recommendations aimed at promoting a "good AI society".

Principles-based frameworks provide flexibility and are relatively easy to adopt, making them attractive to organisations and policymakers. However, their lack of enforceability and operational detail limits their

capacity to drive substantive change (Mittelstadt, 2019; Munn, 2023). For the purposes of this dissertation, these frameworks are best understood as normative reference points rather than complete governance solutions.

2.8.2 Risk-based approaches

More recent frameworks adopt risk-based approaches (European Union, 2024). The European Union's AI Act is the most prominent example, classifying AI systems into unacceptable, high, limited, and minimal risk categories and imposing corresponding obligations (Veale & Zuiderveen Borgesius, 2021). High-risk systems must undergo conformity assessments, maintain detailed documentation, and implement ongoing monitoring and human oversight (Schuett, 2023).

Risk-based frameworks have important implications for organisations. They require systems for identifying and classifying AI applications, assessing potential harms, and ensuring compliance with risk-specific requirements (Schuett, 2023). They also create incentives to avoid or carefully justify high-risk applications. However, these frameworks may pose resource and capability challenges, particularly for SMEs (Kop, 2021; Renda et al., 2021).

2.8.3 Sector specific models

Sectoral frameworks highlight how ethical AI governance is applied in specific contexts. For example, in healthcare industries, Morley et al. (2020) emphasise the importance of equity, privacy, and interpretability in diagnostic and decision-support systems, noting that failures in these areas can exacerbate existing health inequities. In the financial sector, Qureshi et al. (2024) discuss ethical considerations in AI-driven financial services, including privacy, bias mitigation, and algorithmic transparency, and link these directly to regulatory compliance and consumer protection. In the public sector, Whittlestone et al. (2019) and related work highlight concerns around surveillance, citizen trust, and patchy adoption of governance mechanisms. This is particularly evident where AI is used in policing (Richardson et al., 2019), welfare (Eubanks, 2018), or immigration, often leading to a "digital welfare state" characterised by rigid automated decision-making and reduced accountability (Alston, 2019).

These sector-specific models emphasise that "effective" governance is context-dependent: what is adequate in one domain may be insufficient in another, depending on stakes, power asymmetries, and regulatory environments.

2.8.4 Hybrid models and human flourishing

Emerging hybrid models seek to combine the flexibility of principles-based approaches with the enforceability of regulation and the specificity of technical safeguards. Stahl et al. (2021) for instance, propose governance that integrates ethical principles, legal compliance, and organisational accountability with a view to supporting human flourishing. Hybrid models are particularly instrumental in addressing RQ2, which seeks to identify the most effective governance frameworks for responsible AI. By moving beyond purely symbolic ethics, these models offer tangible frameworks that can be embedded directly within organisational governance systems, allowing for a rigorous evaluation of their practical effectiveness.

2.9 Empirical Findings on Implementation

Systematic reviews and empirical studies provide critical insight into how governance frameworks are applied in practice and where they fall short. Fjeld et al. (2020) show convergence on ethical principles but divergence in how they are interpreted and implemented, pointing to the principles-to-practices gap. Jobin et al. (2019) document the global spread of AI ethics guidelines and highlight cultural and political differences for example, variations in emphasis on human rights, innovation, or national security. Capgemini Research Institute (2020) find that while most organisations recognise the importance of ethical AI, relatively few have robust governance processes such as formal audits, impact assessments, or dedicated governance roles. Stahl et al. (2021) identify selective adoption of mitigation strategies, with organisations more likely to implement low-cost measures (e.g., ethics training) than resource-intensive practices (e.g., independent audits or redesigning business models). Taken together, these empirical studies suggest that external pressures including regulatory expectations, industry norms, and public scrutiny do encourage organisations to adopt some governance practices. However, they also reveal persistent barriers such as privacy concerns, lack of governance frameworks, scepticism toward algorithmic outputs, capacity constraints, and competing organisational priorities. Nonetheless, the empirical evidence base remains relatively limited and is often

skewed towards analyses of principles and policy proposals rather than longitudinal evaluations of governance interventions and their outcomes (Attard-Frost et al., 2023; Munn, 2023).

2.9.1 Key challenges and gaps in practice

Based on the literature reviewed, several persistent challenges can be identified. The Principles-to-practices gap, where ethical principles remain abstract and difficult to operationalise (Mittelstadt, 2019). Organisations frequently endorse principles but lack clear processes, metrics, or tools for implementation (Morley et al., 2020). Furthermore, lack of standardisation and variation in fairness definitions, auditing methods, and documentation practices undermines comparability and consistency across organisations and sectors (Fjeld et al., 2020; Raji et al., 2020). Transparency limitations means proprietary models and trade secrets constrain transparency (Pasquale, 2015), as do technical limitations of explainability for complex models (Burrell, 2016). Weak accountability mechanisms inadvertently means responsibility gaps persist, and there are few enforceable methods to assign responsibility and provide remedies when harms occur (Matthias, 2004; Munn, 2023; Raji et al., 2020). Cultural and political differences brings light to divergent governance philosophies, such as rights based approaches in Europe versus innovation-focused approaches elsewhere which complicate international harmonisation (Jobin et al., 2019; Roberts et al., 2021). These challenges emphasise that effective AI governance is not simply a matter of selecting the “right” principles or frameworks; it requires sustained organisational effort to design, implement, and evaluate governance arrangements in context.

2.9.2 Importance of practice and empirical evidence

A critical theme across the literature is the necessity of grounding ethical governance in practice and empirical evidence. Given the recency of the incident, peer-reviewed evaluations of implementation effectiveness are not yet available; accordingly, the Manage My Health case is treated as contextual grey literature that motivates the research problem and illustrates governance dynamics, rather than being included as empirical evidence within the systematic review dataset ("Manage My Health Ltd v Unknown Defendants [2026] NZHC 2," 2026; Office of the Privacy Commissioner, 2026a, 2026c). Conceptual principles and frameworks provide direction, but without empirical validation they risk remaining ineffective.

The example of New York City's bias audit regime illustrates how poorly designed regulation can fail to improve fairness outcomes if audits are narrow in scope, weakly enforced, or detached from broader organisational change. In healthcare, empirical studies show that without attention to model interpretability, clinical workflow integration, and data quality, AI systems can exacerbate, rather than reduce, health inequities (Morley et al., 2020). In finance, Qureshi et al. (2024) highlight how empirical testing of algorithms for bias and transparency directly affects regulatory compliance and consumer outcomes.

For this dissertation, the implication is clear: evaluating governance frameworks requires attention not only to their conceptual coherence but also to how they are enacted in real organisational settings and what effects they have. This provides the rationale for focusing the systematic literature review on organisational experiences of implementing and adapting AI governance frameworks (RQ1) and on reported evidence of their effectiveness (RQ2).

2.9.3 Research gaps and implications for this study

Overall, this chapter demonstrates that the literature on ethical AI governance is vibrant yet fragmented. Reviews of ethical guidance indicate broad convergence around recurring principles, but they also reveal persistent disagreement over definitions, prioritisation, and the degree to which principles can be translated into actionable organisational controls (Jobin et al., 2019; Morley et al., 2020). This is reflected in a central debate in the field: whether ethical governance is best advanced through voluntary, principles-based "soft law" approaches that encourage organisational learning and flexibility, or through enforceable "hard law" regimes that standardise obligations and create stronger accountability (Butcher & Beridze, 2019; European Union, 2024). Even within regulatory approaches, scholarship highlights tensions between innovation and compliance, with concerns that risk-based models may impose disproportionate burdens or encourage box-ticking compliance, particularly for smaller organisations, unless supported by clear guidance, monitoring, and meaningful enforcement practices (European Union, 2024; Veale & Zuiderveen Borgesius, 2021).

At the organisational level, the literature similarly contrasts symbolic adoption with substantive implementation. While governance frameworks and organisational mechanisms are increasingly promoted, including practices like audits and broader assurance activities, evidence of sustained effectiveness remains limited, and many studies continue to emphasise a "principles-to-practices" gap (Morley et al., 2020; Raji et

al., 2020). Moreover, what counts as “effective” governance is shaped by the ethical problems governance is meant to mitigate, most consistently, bias and unfairness, opacity, privacy risks, and unclear accountability, and by contextual factors such as organisational capability, sectoral constraints, and the surrounding regulatory environment (Capgemini Research Institute, 2020). These gaps justify the systematic literature review undertaken in this dissertation to synthesise how organisations implement and adapt AI governance frameworks and to assess which approaches appear most effective in supporting responsible AI use. Chapter 3 therefore sets out the methodology for the review, including the search strategy, inclusion and exclusion criteria, and the analytical approach used to interpret and synthesise the evidence.

Chapter 3: Methodology

3.1 Introduction

This chapter sets out the methodology used to investigate ethical governance of artificial intelligence (AI) in organisational contexts. The dissertation is guided by two research questions: RQ1: How do organisations implement and adapt governance frameworks to ensure the ethical use of AI technologies? RQ2: What are the most effective governance frameworks for ensuring responsible AI use in organisations? A methodological approach capable of producing a transparent and reproducible synthesis is required because the literature is interdisciplinary, fast-moving, and characterised by uneven translation of ethical principles into organisational practice. Indeed, prior research shows that ethical AI guidance has proliferated globally, yet governance often remains aspirational, fragmented, and inconsistently operationalised-an enduring "principles-to-practice" gap (Fjeld et al., 2020; Jobin et al., 2019; Mittelstadt, 2019; Morley et al., 2020). Building on the conceptual framing developed in Chapter 2, this chapter explains how the study systematically identified, screened, appraised, extracted, and synthesised peer-reviewed and scholarly literature to generate the evidence base for Chapter 4 (Findings) and Chapter 5 (Discussion).

3.2 Research Approach and Logic of Inquiry

This dissertation is positioned within an interpretivist sensibility that treats ethical AI governance as an organisational accomplishment rather than a purely technical artefact. Governance is enacted through structures, processes, accountabilities, and socio-technical practices that are interpreted and negotiated within specific institutional and organisational conditions (Mäntymäki et al., 2022; Raji et al., 2020). This stance is appropriate because ethical governance is most clearly expressed in operational decisions and routines, such as how organisations set risk tolerances and escalation thresholds, define oversight roles and decision rights, specify documentation and transparency standards, implement audit and monitoring practices, and allocate accountability for downstream impacts (Mökander et al., 2021; Raji et al., 2020).

The analysis is guided by an abductive logic of inquiry. Abduction is particularly suitable for a qualitative synthesis in a contested and evolving domain because it allows the researcher to move iteratively between empirical material and theory, refining explanatory interpretations as patterns emerge (Timmermans &

Tavory, 2012). In this study, the abductive approach supports two goals: first, to remain anchored in the empirical claims and evidence presented in the included studies; and second, to interpret those claims using the conceptual vocabulary established in Chapter 2 (including governance mechanisms such as audits, impact assessments, oversight bodies, standards, and accountability processes). This approach is well aligned with the dissertation's applied focus: the intent is not merely to catalogue ethical principles but to explain how governance is operationalised and what conditions appear to support effectiveness.

3.3 Research Design: Systematic Literature Review

To address the research questions with rigour and transparency, the dissertation adopts a systematic literature review (SLR) design. In management and organisational research, SLRs are valued for producing evidence-informed syntheses that are less vulnerable to selective citation and post hoc inclusion decisions than narrative reviews (Denyer & Tranfield, 2009; Tranfield et al., 2003). The SLR approach is appropriate here because research on ethical AI governance is distributed across business ethics, management, information systems, public administration, law and policy, and computing-oriented outlets. An SLR provides a structured method for locating relevant studies across these boundaries and for documenting selection decisions in a traceable manner. Contemporary incident reporting and other non-peer-reviewed materials were excluded from the systematic review dataset; where cited, such sources are used only for contextual framing and are not treated as empirical evidence for answering the review questions.

The review employs a qualitative synthesis logic. Given the variation of research designs and outcomes in this field, ranging from conceptual frameworks and organisational case studies to surveys and evaluative audits, statistical meta-analysis is not feasible or conceptually appropriate. Instead, the review employs thematic synthesis to integrate findings into analytically coherent themes that directly address both research questions (Braun & Clarke, 2006; Thomas & Harden, 2008)

3.4 Review Protocol and Reporting Standard

The review is reported in alignment with the PRISMA 2020 statement (Page et al., 2021). PRISMA is used as a transparency and audit-trail framework, documenting each stage from identification of records through screening, eligibility assessment, and final inclusion. In addition, the design follows established guidance on systematic reviewing in management research (Denyer & Tranfield, 2009; Tranfield et al., 2003) and draws

on contemporary methodological discussions about rigour and transparency in review work (Fan et al., 2022; Snyder, 2019). Together, these sources inform the protocol's emphasis on replicability, explicit inclusion rules, and careful documentation of selection and synthesis procedures.

3.5 Scoping of Prior Reviews and Protocol Refinement

Before finalising the protocol, an initial scoping exercise was undertaken to understand how the field has been previously mapped and where gaps in evidence remain. This scoping did not form part of the final SLR included studies; rather, it was used to refine the search terminology and to ensure that the review design responded to recognised weaknesses in the literature, particularly the tendency for ethics discourse to emphasise high-level principles without corresponding detail about organisational implementation and evaluation (Mittelstadt, 2019; Morley et al., 2020). Key insights from global mappings of AI ethics guidelines were used to inform keyword development and to ensure that the search strategy captured both principle-based and practice-oriented scholarship (Fjeld et al., 2020; Jobin et al., 2019). The scoping phase, therefore, strengthened construct clarity and improved search sensitivity to governance mechanisms that may not be labelled uniformly across disciplines.

3.6 Search Strategy

3.6.1 Information sources

The primary search was conducted in July 2025 using the EBSCOhost platform (via AUT Library access). EBSCOhost was selected because it provides integrated searching across multiple disciplinary collections and supports the construction of systematic queries, the application of consistent limiters, and export to reference management software. Given the interdisciplinary nature of ethical AI governance, the review sought to capture organisational and managerial scholarship alongside material from computing and the applied sciences. The core collections searched included Business Source Premier and Business Source Complete (management and organisational research), Academic Search Complete (multidisciplinary coverage), and Computers & Applied Sciences Complete (technology-oriented scholarship), alongside additional AUT-licensed EBSCO collections accessible at the time of search. To support replicability, the full list of collections searched is provided in Table 7 in Appendix A.

3.6.2 Temporal scope

The review covered publications from 2015 to 2025. This ten-year window is methodologically chosen for three reasons. First, it captures the acceleration of organisational AI adoption associated with contemporary machine learning capabilities and increasing deployment of automated decision systems in operational contexts. Second, it reflects the period in which “ethical AI” shifted from a predominantly normative discussion to a more applied governance agenda inside organisations, including the emergence of governance structures such as internal standards, oversight bodies, and audit routines. Third, the period encompasses major global regulatory and policy activity that has shaped organisational governance expectations and practice, for example, the OECD (2019a) principles and the European Commission (2019) Trustworthy AI guidance.

3.6.3 Search fields and query construction

Given the interdisciplinary nature of this topic, relevant studies often employ varying terminology. To address this, the search utilised EBSCOhost’s "TX" (All Text) field code. This approach maximised sensitivity by scanning the full text of articles rather than limiting the search to titles or abstracts, ensuring that significant studies using non-standard terminology were not overlooked. Following this broad retrieval, a manual screening of Titles (TI) and Abstracts (AB) was conducted to filter for precision and relevance.

To ensure this study can be replicated by future researchers, the exact Boolean search string employed is detailed in Figure 1 below. The query utilises truncation (marked by an asterisk *) to capture grammatical variations of key terms. For example, organis searches for organisation, organisations, and organisational.

Figure 1. The final Boolean search string used in EBSCOhost

```
TX ("artificial intelligence" OR AI OR "machine learning" OR "deep
learning" OR "generative AI" OR "large language model*" OR LLM* OR
"algorithmic decision*" OR "automated decision*")
AND TX (governance OR "AI governance" OR "governance framework*" OR policy
OR policies OR oversight OR "risk management" OR audit* OR "algorithmic
audit*" OR "impact assessment*" OR "model monitoring" OR compliance OR
regulation OR regulatory OR standard* OR guideline*)
AND TX (ethic* OR "responsible AI" OR "trustworthy AI" OR fairness OR bias
OR discrimination OR accountability OR transparency OR explainab* OR
privacy OR "data protection" OR "human rights" OR justice OR moral*)
AND TX (organisation* OR organization* OR business OR firm* OR compan* OR
enterprise* OR workplace* OR "public sector" OR government* OR agency)
```

The query was limited to English-language publications within the 2015-2025 date range. Where EBSCOhost limiters were available, scholarly publication types were prioritised. The review included peer-reviewed journal articles and refereed conference proceedings, as well as scholarly book chapters and dissertations that met relevance and quality criteria, reflecting the interdisciplinary nature of the publication landscape in this domain.

3.6.4 Search refinement and documentation

The search string was iteratively refined through pilot searches to balance recall (capturing the breadth of relevant governance and ethics terminology) and precision (reducing retrieval of purely technical optimisation studies with no governance or ethical content). Refinement was informed by the scoping work and by the conceptual framing developed in Chapter 2.

3.7 Eligibility Criteria

Eligibility criteria were established prior to screening to reduce ad hoc judgement and to ensure consistent selection decisions. Studies were eligible for inclusion if they: (a) addressed AI or closely related systems (including machine learning and automated decision systems); (b) engaged substantively with governance mechanisms (for example, oversight arrangements, policies, standards, audits, impact assessments, compliance mechanisms, or regulatory governance); (c) treated ethical considerations as central rather than incidental (for example, fairness/bias, accountability, transparency/explainability, privacy, rights-based concerns, or responsible AI framing); and (d) situated these issues within organisational contexts (such as firms, public agencies, workplaces, or institutional settings where governance structures and processes operate). Publications were required to be in English and to fall within 2015-2025.

Studies were excluded if they were primarily technical contributions focused on performance improvements without ethical and governance engagement; if they addressed individual attitudes or consumer behaviour without organisational governance relevance; or if they were non-scholarly grey literature (for example, blogs, news articles, and unreviewed commentary) where academic rigour could not be assured. Where seminal pre-2015 works were needed to support conceptual framing (for example, foundational discussions of accountability gaps), these were treated as part of Chapter 2 rather than included in the SLR included studies, consistent with the review's "implementation-forward" evidentiary focus.

3.8 Screening and Study Selection Procedure

3.8.1 Record management

All retrieved records were exported to EndNote for reference management, deduplication, and screening organisation. EndNote provided an auditable basis for tracking records through the review stages and ensured consistent handling of citations during extraction and write-up.

3.8.2 Screening stages

Screening proceeded in three stages. First, titles and abstracts were screened against the eligibility criteria to remove records that were clearly irrelevant. Second, full texts were assessed against the inclusion criteria and the study's evidentiary requirements for addressing implementation/adaptation and effectiveness. Third, a snowballing reference list was undertaken from included studies and selected high-relevance papers to identify additional relevant literature that may not have been captured through keyword searching alone. Snowballing is particularly important in interdisciplinary fields where relevant work may be indexed inconsistently or described using non-standard terminology.

3.8.3 Two-stage screening rubric: gatekeepers and evidence weighting

To align the final included studies with the applied nature of the research questions, this review employed a two-stage rubric at title/abstract screening. Table 1 below shows Stage 1 applied mandatory "gatekeeper" criteria: studies had to explicitly address AI, governance mechanisms, and ethical orientation. Records were retained only if titles and abstracts simultaneously referenced artificial intelligence, governance mechanisms,

and ethical concepts. These filters ensured conceptual alignment with the review's scope. A record stays in the pool only if all three gatekeepers match the title and abstract.

Table 1. Stage 1 Gatekeeper (mandatory) filters

Gatekeeper	Why it is mandatory	Required keywords (substring OR list)
Artificial intelligence focus	Keeps only work about AI systems (not generic IT).	*"artificial intelligence"*, AI (as standalone token), *"machine learning"*, *"deep learning"*, *"generative AI"*, *"algorithmic decision"*
Governance focus	Ensures structural / oversight discussion (compared to technical design alone)	*governance*, *framework*, *oversight*, *policy/ies*, *standard/s*, *audit*, *compliance*, *regulation*, *regulatory*
Ethical orientation	Anchors review in "responsible AI", not performance optimisation	any stem of *ethic*, *responsible*, *trustworthy*, *bias*, *fairness*, *accountability*, *transparency*, *justice*, *moral*

Table 2 below shows Stage 2 used a structured scoring rubric to privilege studies able to speak directly to organisational implementation/adaptation and evaluative evidence. Specifically, studies were scored across four dimensions: organisational context (+1), implementation/adaptation detail (+2), evaluative evidence about governance effectiveness (+2), and empirical basis (+1), for a maximum of six points. Only studies achieving the maximum score (6/6) were progressed to full-text review. Papers passing the gatekeepers were scored across four dimensions totalling six possible points. Simple substring search, case-insensitive; stemming/truncation (*) used.

Only papers achieving 6/6 were shortlisted for full-text analysis, producing a final included study of 38 studies. This scoring approach allowed transparent, reproducible filtering of literature to ensure both relevance and methodological robustness.

Table 2. Stage 2: Point-scored criteria

Code	Construct	Points	Operational keywords	Link to RQs
O	*Organisational context*	**+1**	organisation/organisation, firm, company, corporate, enterprise, business, industry, management, hospital, bank, agency, government, public sector, institution	Needed for *"...in organisations"* scope
I	*Implementation / adaptation detail*	**+2**	implement*, adoption/roll-out, adapt*, embed*, operational*, maturity, change-management, integration, deployment	Direct evidence for **RQ 1**
E	*Evidence of effectiveness / evaluation*	**+2**	evaluate*/evaluation, compare/ison, outcome, effectiveness, audit, impact, assessment, metric, result, evidence	Direct evidence for **RQ 2**
M	*Empirical basis*	**+1**	data, survey, interview, **case study**, experiment, empirical, qualitative, quantitative, mixed-methods, field study, dataset, analysis	Favours studies grounded in data rather than opinion

This strict threshold is a deliberate methodological choice designed to operationalise the dissertation's focus on the principles-to-practice gap. Prior scholarship has demonstrated that ethics guidance is abundant, yet evidence about implementation mechanisms and effectiveness is comparatively limited (Mittelstadt, 2019; Morley et al., 2020). A 6/6 threshold, therefore, functions as an evidence-density strategy: it concentrates the included studies on studies that provide organisational enactment detail and evaluative grounding, thereby improving the capacity of the synthesis to answer both research questions. The trade-off is that some high-quality conceptual work may be excluded from the SLR included studies; however, this is structurally mitigated by the dissertation design, as conceptual foundations and normative debates are addressed comprehensively in Chapter 2, while Chapter 3 curates a included studies suited to implementation and effectiveness analysis.

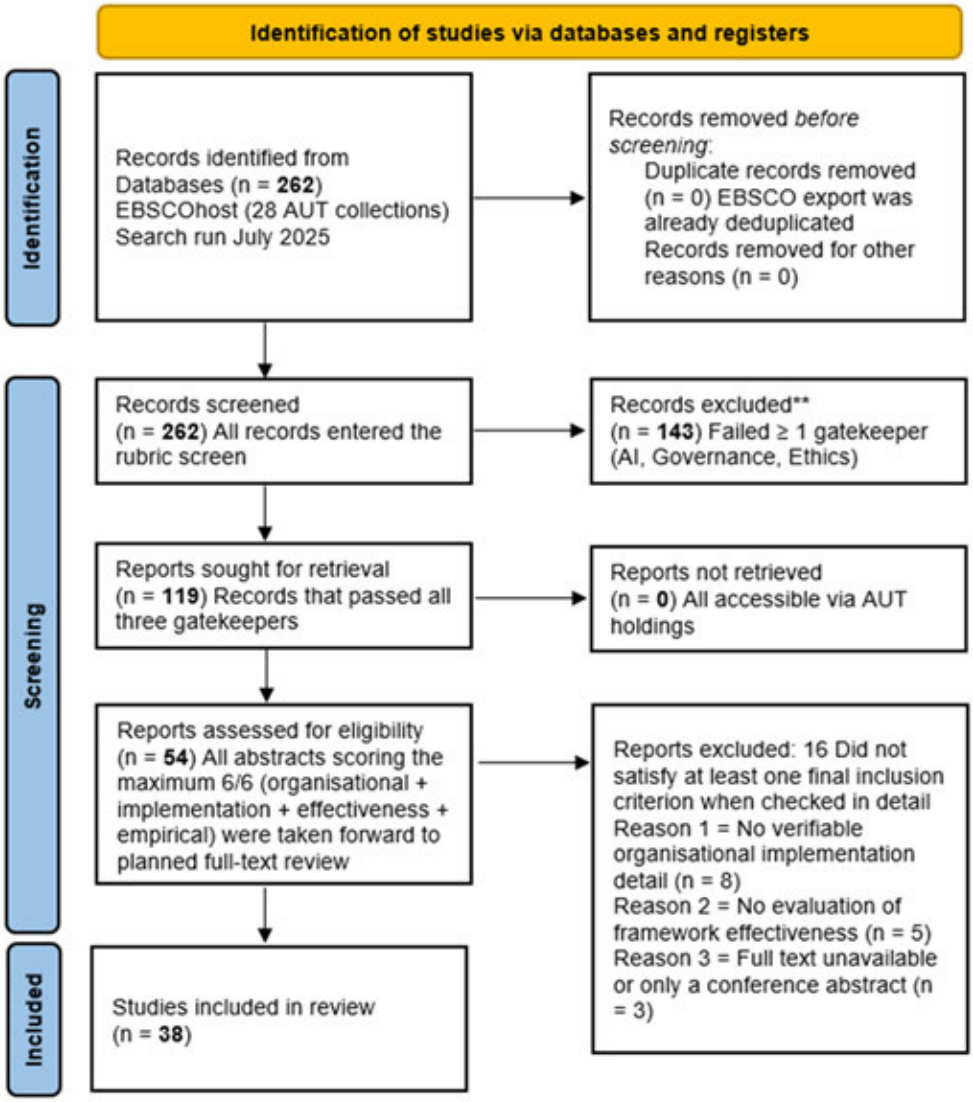
To address methodological sensitivity, namely, the risk that a strict threshold could systematically exclude important but less operationalised contributions, the review incorporated two mitigation strategies. First, citation chaining was employed to ensure that seminal governance contributions connected to the included studies were not inadvertently omitted due to indexing or terminology variations. Second, findings were interpreted with explicit acknowledgement of evidentiary boundaries: where a governance claim was

supported primarily by conceptual argument rather than evaluation, this was treated as suggestive rather than conclusive and was framed accordingly in Chapter 4 and interrogated in Chapter 5.

3.8.6 PRISMA aligned flow of studies

The PRISMA aligned process provides a transparent summary of how the final included studies was produced. The July 2025 search identified 262 records. After Stage 1 screening, 143 records were excluded because they failed at least one gatekeeper criterion (AI focus, governance focus, or ethical orientation). The remaining 119 records proceeded to Stage 2 scoring. Of these, 65 records were excluded because they scored below 6/6. This yielded 54 studies for full-text eligibility assessment. At full-text screening, 16 studies were excluded due to insufficient organisational implementation details, the absence of evaluative evidence relevant to governance effectiveness, or a lack of accessible full text. The final included studies comprised 38 studies included in the qualitative synthesis. These counts are reported in the PRISMA flow diagram in Figure 2 below.

Figure 2. PRISMA 2020 Flowchart adapted from Page et al. (2020)



3.9 Quality Appraisal

In addition to relevance screening, the included studies were critically appraised to support quality-aware synthesis. The appraisal approach was informed by the Critical Appraisal Skills Programme (CASP) checklists and adapted to accommodate the diversity of study designs in the included studies (Critical Appraisal Skills Programme, n.d.). Rather than treating appraisal as a binary inclusion decision, the study used appraisal to inform evidentiary weighting during synthesis. That is, studies judged to have stronger methodological transparency, clearer data provenance, and more defensible evaluation logic were treated as higher-confidence sources when interpreting effectiveness claims, while weaker studies were used more cautiously as illustrative or hypothesis-generating material.

Across designs, the appraisal focused on the clarity of aims and governance constructs, the appropriateness of the method to research questions, the transparency of data sources and sampling, the credibility of the analysis, and the plausibility of claims about governance implementation or outcomes. For evaluative studies, particular attention was given to how "effectiveness" was defined, what indicators were used (for example, audit outcomes, compliance measures, documented improvements in governance practices, or stakeholder outcomes), and whether claims were supported by evidence rather than assertion.

3.10 Data Extraction and Management

Data extraction was conducted using a structured extraction matrix to support systematic comparison across studies and to maintain an auditable chain of evidence from primary sources to thematic claims. Matrix-based extraction is widely used in qualitative evidence synthesis as a way of structuring varied materials for consistent analysis (Gale et al., 2013). The extracted data in Table 10: comparative analysis of AI governance frameworks and implementation strategies within Appendix C were arranged to correspond with the extraction matrix headings in. For each article, bibliographic information was recorded, along with the appropriate evidence mapped to implementation details (RQ1) delineate how organisations execute and modify governance frameworks; governance focus (RQ2) encapsulates the type and orientation of the governance framework discussed; challenges reported; and framework components specify the concrete elements of the governance framework

The extraction matrix served not only as a repository but also as an analytical instrument. It enabled cross-study comparison of governance mechanisms, identification of recurring implementation pathways and barriers, and systematic assessment of how different studies conceptualised and evidenced governance "effectiveness". Throughout the extraction, memos were used to record interpretive judgements, borderline inclusion decisions, and emerging conceptual connections to Chapter 2, thereby supporting transparency and reflexive awareness of the researcher's interpretive role.

3.11 Data Synthesis and Analysis

Given the variation of study designs and outcomes, the dissertation synthesised the evidence using a thematic synthesis approach. The synthesis followed an iterative process consistent with established thematic analysis and thematic synthesis approaches (Braun & Clarke, 2006; Thomas & Harden, 2008). Codes were initially generated to capture governance mechanisms, implementation dynamics, organisational enablers and constraints, and evaluative claims (for example, “use of lifecycle gates, monitoring, and governance checkpoints in deployment”; “committees/boards and explicit roles for accountability and oversight”). Codes were then clustered into descriptive themes that represent recurring patterns across the included studies (e.g., Theme 1 translating principles into practice; Theme 4 accountability and multi-level governance). Table 3 below provides an audit trail demonstrating how raw text extracts were first coded descriptively and then aggregated into more abstract subthemes and final themes reported in Chapter 4.

Table 3. Theme development audit trail (illustrative examples)

Raw extracted code (examples)	Category	Subtheme	Theme
Ethics principles are high-level and unenforced; risk of 'ethics washing'	Principles to practice gap / symbolic compliance	T1c Principles-to-practice failure	Theme 1: Translating principles into practice
Use of lifecycle gates, monitoring, and governance checkpoints in deployment	Workflow embedding mechanisms	T1b Lifecycle embedding	Theme 1: Translating principles into practice
Sectoral constraints and domain-specific accountability (e.g., clinical governance)	Contextual governance variation	T2a Healthcare / T2b Public sector / T2c Finance	Theme 2: Sector-specific governance models
Capability gaps; need ethics literacy, training, and governance maturity	Organisational readiness	T3b Capability & culture	Theme 3: Organisational capacity and culture
Data governance, privacy, and security as preconditions for responsible AI	Data foundations	T3c Data governance & privacy/security foundations	Theme 3: Organisational capacity and culture

Committees/boards and explicit roles for accountability and oversight	Internal decision rights	T4a Internal accountability structures	Theme 4: Accountability and multi-level governance
Bias audits, evaluation, assurance and ongoing monitoring	Assurance and control	T4b Assurance mechanisms	Theme 4: Accountability and multi-level governance
Alignment with external regulation and standards (e.g., GDPR, emerging AI law)	External legitimacy and compliance	T4c External regulation/standards alignment	Theme 4: Accountability and multi-level governance
Risk-tiering and proportional governance to allocate controls by severity	Risk-based governance design	T5a Risk-based/proportional governance	Theme 5: Characteristics of effective governance frameworks
Fairness/equity as design and governance criteria; bias/discrimination mitigation	Equity-centred governance	T5b Fairness/equity & discrimination	Theme 5: Characteristics of effective governance frameworks
GenAI introduces new risks (scale, opacity, misuse) requiring adapted governance	Emerging technology adaptations	T5c GenAI governance challenges	Theme 5: Characteristics of effective governance frameworks
Governance linked to trust, organisational performance, and value creation	Effectiveness outcomes	T5d Effectiveness outcomes	Theme 5: Characteristics of effective governance frameworks

To strengthen theoretical coherence while avoiding a biased, predetermined governance model, inductive themes were subsequently cross-checked against widely recognised normative frameworks such as OECD AI principles (OECD, 2019a) and the European Commission (2019) guidance on trustworthy AI. This step functions as analytic triangulation to support alignment between observed organisational practices (for example, audit routines, accountability structures, lifecycle management processes) and established governance principles (for example, accountability, transparency, fairness), while also enabling the synthesis to identify where organisational practice departs from, narrows, or selectively adopts normative frameworks.

Following the coding process detailed in Table 3 above, five overarching themes emerged. Table 4 below presents a high-level overview of these themes, sorted by coverage showing the evidence base concentrated

most heavily on the organisational and accountability conditions of governance (Themes 3 and 4), and on characteristics associated with effective frameworks (Theme 5). By contrast, detailed descriptions of concrete operationalisation artefacts (for example, specific checklists, impact assessment templates, or model documentation protocols) appeared in only a small subset of studies, underscoring a recurring gap between high-level principles and implementable practice. This thematic framework is then used to organise the reporting of findings in Chapter 4.

Table 4. Overview of themes and coverage across the included studies

Theme	Analytic focus	n studies	% of studies
Theme 1: Translating principles into practice	How ethics principles are operationalised through instruments, embedded in lifecycle workflows, and undermined by symbolic compliance.	15	39%
Theme 2: Sector-specific governance models	How governance design varies by sector and domain constraints (for example, healthcare, public sector, finance, labour, justice).	23	61%
Theme 3: Organisational capacity and culture	Capabilities, cultural conditions, and data foundations that enable or constrain responsible AI governance.	31	82%
Theme 4: Accountability and multi-level governance	Internal accountability structures, assurance mechanisms, and alignment with external standards and regulation.	32	84%
Theme 5: Characteristics of effective governance frameworks	Risk-based and fairness-focused approaches, GenAI-specific challenges, and reported effectiveness outcomes.	36	95%

3.12 Rigour, Trustworthiness and Reflexivity

Rigour in systematic review research is established through transparency, procedural consistency, and a defensible chain of evidence. This dissertation strengthens trustworthiness in four ways. First, the PRISMA-aligned reporting provides an explicit audit trail for identification, screening, and inclusion decisions (Page et al., 2021). Second, the two-stage rubric establishes a clear inclusion logic aligned with the dissertation's applied focus, thereby reducing subjectivity at the early screening stages. Third, the structured extraction matrix ensures consistent capture of key constructs and facilitates traceable movement from evidence to themes. Fourth, appraisal-informed weighting supports cautious interpretation of effectiveness claims, particularly where governance outcomes are asserted rather than evaluated.

Because synthesis is interpretive, reflexivity is also necessary. The researcher's prior engagement with the ethical AI governance literature (as developed in Chapter 2) provides analytic sensitivity but may also shape attention towards particular constructs (for example, audits, accountability, transparency). To manage this risk, reflexive memos (Appendix D and E) were maintained throughout screening, extraction, coding, and theme refinement to document borderline judgements, decision rationales, and code/theme changes, thereby strengthening transparency and auditability. These memos are complemented by the updated comparative synthesis table (Appendix C), which maps each included study to its RQ1/RQ2 contributions and reported outcomes, providing a clear evidence trail from study-level data to the thematic claims presented in Chapters 4 and 5.

3.13 Ethical Considerations

The dissertation is based exclusively on secondary data from published sources and does not involve human participants, interviews, or collection of personal data. Formal ethics approval was therefore not required. Nevertheless, ethical scholarship obligations remain central. These include the accurate representation of primary studies, attribution and citation, and the avoidance of overclaiming where the evidence base is limited or methodologically weak.

3.14 Chapter Summary

This chapter has described the methodology used to construct a rigorous evidence base for analysing ethical AI governance in organisations. Using a PRISMA-aligned systematic literature review design (Page et al., 2021; Tranfield et al., 2003), the study searched EBSCOhost in July 2025, applied explicit eligibility criteria, and used a two-stage rubric to prioritise studies offering organisational implementation detail and evaluative evidence. Records were managed in EndNote, extracted through a structured matrix, appraised for quality using a CASP-informed approach, and synthesised via thematic synthesis with subsequent triangulation against established normative frameworks (Braun & Clarke, 2006; European Commission, 2019; OECD, 2019a; Thomas & Harden, 2008).

Chapter 4 presents the findings from the 38 included studies. It is organised to address the research questions in sequence: first, the governance mechanisms and adaptation patterns organisations use to operationalise ethical AI governance (RQ1), and second, the forms of evaluative evidence and effectiveness claims associated with governance approaches (RQ2). Discussion in Chapter 5 makes sense of the findings by linking them back to the key debates and missing pieces identified in the literature review in Chapter 2. It brings the evidence together to explain what we can be confident about, what is still unclear, and what organisations and researchers should focus on next.

Chapter 4: Findings

4.1 Overview of Thematic Findings

This chapter reports the findings of the systematic literature review (SLR) examining governance frameworks that support responsible artificial intelligence (AI) use in organisations. The evidence base comprises 38 included studies, which were treated as the primary dataset for synthesis. Findings were developed through a transparent coding and thematic synthesis process. Specifically, extracted study-level data were coded and then iteratively grouped into subthemes, before being consolidated into five higher-order themes: Translating principles into practice (theme 1), sector-specific governance models (theme 2), organisational capacity and culture (theme 3), accountability and multi-level governance (theme 4) and characteristics of effective governance frameworks (theme 5). To make the analytical pathway explicit, Table 3 in Chapter 3 shows examples of theme development audit trail, and summarises how salient codes were aggregated into categories, subthemes, and ultimately themes, and coverage of the themes across studies is summarised in Table 4 in Chapter 3.

4.1.1 Governance frameworks

Across the included studies, governance frameworks were most often described as combination of mechanisms, such as formal structures, repeatable processes, and supporting artefacts that operate across the AI lifecycle and often span organisational boundaries (Mäntymäki et al., 2022). Table 5 below summarises these mechanisms and how frequently they were reported in the evidence base.

In the table, “governance mechanism type” refers to the broad category of mechanism described in the studies, “n studies” indicates how many of the included studies mentioned that mechanism, and the “representative contributing studies” column provides illustrative examples underpinning each category.

Importantly, the mechanism types are non-mutually exclusive, meaning a single study could contribute to multiple categories (for example, a paper might discuss both lifecycle controls and committee oversight).

Table 5. Frequently reported governance mechanisms in the included studies (non-mutually exclusive)

Governance mechanism type	n studies	Representative contributing studies
Ethics principles, values statements, or guidelines (normative layer)	18	Khan et al. (2022); Nalbandian (2022); Monyela and Tella (2024); Okun et al. (2023); Ridzuan et al. (2024); Bensa and Fattore (2024)
Operationalisation tools (impact assessments, checklists, documentation, audits)	8	Munn (2023); T. Zhou et al. (2025); Luo et al. (2025); Mujtaba (2025); D. Zhou et al. (2025); Yang et al. (2024)
Lifecycle governance (deployment gates, monitoring, MLOps)	11	Mäntymäki et al. (2022); T. Zhou et al. (2025); Luo et al. (2025); Bensa and Fattore (2024); Mujtaba (2025); (D. Zhou et al. (2025)
Dedicated roles or bodies (ethics officers, committees, boards)	18	Attard-Frost et al. (2023); Munn (2023); Khan et al. (2022); Mäntymäki et al. (2022); Monyela and Tella (2024); Okun et al. (2023)
Capability building (ethics literacy, training, organisational maturity)	14	Khan et al. (2022); Mäntymäki et al. (2022); Khadka and Ullah (2025); Monyela and Tella (2024); Luo et al. (2025); Mujtaba (2025)
Data foundations (privacy, security, data governance)	19	Khan et al. (2022); Batool et al. (2025); Nalbandian (2022); Khadka and Ullah (2025); Olawade et al. (2025); Luo et al. (2025)
External alignment (regulation, standards, certification)	25	Munn (2023); Mäntymäki et al. (2022); Batool et al. (2025); Khadka and Ullah (2025); Luo et al. (2025); Okun et al. (2023)

As we can see in Table 5 above, the most frequently reported mechanism type was external alignment, such as references to regulation, standards, and certification that were reported in 25 studies, indicating that many governance approaches are framed in relation to external expectations and compliance pressures. Data foundations (privacy, security, and data governance) were also prominent (19 studies), reflecting the view that trustworthy data handling is a precondition for responsible AI governance.

At the level of organisational design, ethics principles and guidelines (the "normative layer") appeared in 18 studies, but this is widely understood as insufficient on its own without operational controls and accountability structures (Mittelstadt, 2019; Munn, 2023). Complementing this, dedicated roles or bodies

(such as ethics officers, committees, or boards) were also reported in 18 studies, signalling the importance of formal oversight and decision rights in making governance actionable. More practice-oriented mechanisms included lifecycle governance (deployment gates, monitoring, and MLOps) reported in 11 studies, and operationalisation tools (impact assessments, checklists, documentation, and audits) reported in 8 studies, such as mechanisms that translate high-level principles into day-to-day governance routines (Morley et al., 2020; Raji et al., 2020). In the end, 14 studies showed that competence building (ethical literacy, training, and organisational maturity) is important for good governance. This shows that good governance depends not only on rules and institutions, but also on the right skills and ability to follow them consistently.

4.2 Theme 1: Translating Principles into Practice

Theme 1 synthesises how organisations attempt to translate responsible AI principles into operational practice, drawing on 15 included studies (Alzebda & Matar, 2025; Attard-Frost et al., 2023; Bensa & Fattore, 2024; Birkstedt et al., 2023; Ismail & Goh, 2024; Lacmanovic & Skare, 2025; Luo et al., 2025; Mac, 2024; Mäntymäki et al., 2022; Mujtaba, 2025; Munn, 2023; Palladino, 2023; Tian et al., 2025; D. Zhou et al., 2025; T. Zhou et al., 2025). Across these studies, the central challenge is not the absence of ethical aspiration, but the difficulty of specifying, embedding, and sustaining governance controls that meaningfully shape day-to-day design and deployment decisions.

4.2.1 Operationalisation of measurement instruments

Only two studies provided comparatively concrete accounts of operationalisation instruments such as checklists, impact assessments, or standardised documentation (Lacmanovic & Skare, 2025; Munn, 2023). These studies emphasised that instrument design matters: the artefacts must be usable by practitioners, integrated with existing risk and compliance workflows, and supported by clear accountability for completion and review. In this evidence base, operationalisation was framed less as a one-off ‘ethics checklist’ exercise and more as a structured decision record that enables contestability and auditability over time (Lacmanovic & Skare, 2025; Munn, 2023).

4.2.2 Lifecycle embedding and continuous governance

A broader set of 11 studies argued that effective governance depends on embedding controls across the AI lifecycle, particularly at points of model development, deployment approval, and post-deployment monitoring (Alzebda & Matar, 2025; Bensa & Fattore, 2024; Birkstedt et al., 2023; Ismail & Goh, 2024; Luo et al., 2025; Mac, 2024; Mäntymäki et al., 2022; Mujtaba, 2025; Tian et al., 2025; D. Zhou et al., 2025; T. Zhou et al., 2025). In these accounts, governance is enacted through gates and feedback loops: models are assessed before release, monitored for drift and unintended consequences, and re-evaluated when context changes. Studies that frame AI governance as an organisational capability highlight the importance of integrating responsible AI requirements into machine learning operations (MLOps) pipelines and quality assurance routines (Alzebda & Matar, 2025; Birkstedt et al., 2023; D. Zhou et al., 2025). Public-sector oriented analysis similarly stress that administrative AI systems require ongoing oversight because downstream effects on citizens can evolve as data, policies, and operational contexts shift (Bensa & Fattore, 2024; Mac, 2024). Embedding governance into each stage of the AI lifecycle also connects to Theme 2: what "good" lifecycle governance looks like varies by sector because legal duties, risk exposure, and the severity of potential harm differ across domains.

4.2.3 Principles-to-practice failure and symbolic compliance

Three studies critically examine the risk of symbolic compliance, where ethical principles are adopted rhetorically but remain weakly coupled to organisational decision-making (Attard-Frost et al., 2023; Munn, 2023; Palladino, 2023). These studies described "ethics washing" dynamics, in which voluntary guidelines and high-level commitments are used to signal responsibility without materially constraining behaviour. The findings suggest that this failure mode is most likely when governance remains at the level of aspirational principles without enforceable accountability or independent scrutiny (Attard-Frost et al., 2023; Munn, 2023; Palladino, 2023).

4.3 Theme 2: Sector-Specific Governance Models

Theme 2 highlights that "effective" governance is context-sensitive: the included studies consistently located governance design within domain-specific constraints, including regulatory regimes, professional standards, data sensitivity, and the nature of potential harms. This theme drew on 23 studies (Albalawee & Fahoum,

2024; AlTawil & Rahhal, 2025; Alzebda & Matar, 2025; Bensa & Fattore, 2024; Diaz-Asper et al., 2024; Emah & Bennett, 2025; Lacmanovic & Skare, 2025; Loufek et al., 2024; Luo et al., 2025; Mac, 2024; Mujtaba, 2025; Munn, 2023; Nalbandian, 2022; Okun et al., 2023; Olawade et al., 2025; Paul & Rena, 2025; Rana et al., 2024; Ridzuan et al., 2024; Sandeep et al., 2025; Saw & Ng, 2022; Zada et al., 2024; Zhao et al., 2021; T. Zhou et al., 2025) with the most sustained sectoral attention given to healthcare, public-sector and governmental contexts, and finance and fintech.

4.3.1 Healthcare and clinical-adjacent contexts

Ten studies addressed responsible AI governance in healthcare and clinical-adjacent settings (Bensa & Fattore, 2024; Diaz-Asper et al., 2024; Emah & Bennett, 2025; Loufek et al., 2024; Luo et al., 2025; Okun et al., 2023; Olawade et al., 2025; Paul & Rena, 2025; Saw & Ng, 2022; Zhao et al., 2021). Collectively, these studies treat governance as inseparable from patient safety, data stewardship, and the ethics of clinical decision support, while also spanning a wide range of healthcare AI use cases and governance concerns. Some contributions are anchored in specific clinical and operational settings (for example, ophthalmic AI and diagnostic support, and the implementation challenges of AI in medical imaging), while others foreground system-level governance questions such as ethical sourcing and reuse of patient data, and the governance implications of surveillance systems for infectious diseases. The set also includes work that extends beyond “clinical AI” narrowly defined, highlighting administrative uses of AI in public healthcare, the ethical challenges of language technologies in clinical and behavioural applications, and the newer governance demands introduced by generative AI and large language models in healthcare contexts. Across these different emphases, the overall message is consistent: performance metrics and technical validation matter, but they are not enough on their own. Credible healthcare governance also requires clear purpose statements, explicit clinical accountability, robust consent and privacy practices, and ongoing attention to bias and representativeness in real-world deployment.

4.3.2 Public sector and administrative AI

Six studies examined governance in public-sector or governmental settings (Alzebda & Matar, 2025; Bensa & Fattore, 2024; Mac, 2024; Paul & Rena, 2025; Rana et al., 2024; T. Zhou et al., 2025). Collectively, they foreground how public-sector AI governance must be anchored in administrative law and democratic

accountability, including contestability, due process, and proportionality, because algorithmic decisions can directly affect citizens' rights and access to services (Bensa & Fattore, 2024; Mac, 2024; Rana et al., 2024).

Within this set, Bensa and Fattore (2024) highlights these concerns through the administrative realities of public healthcare, where governance needs to make oversight and responsibility visible in day-to-day administration, while Mac (2024) emphasises public legitimacy and the need for procedures that allow challenge and review when automated or AI-supported decisions are used. Alzebda and Matar (2025) adds a complementary lens by showing why government regulation matters for public uptake and acceptance of AI, reinforcing that governance in the public sector is not only about internal controls but also about public trust and regulatory signals. T. Zhou et al. (2025) draws attention to the operational side of public-sector governance, including cross-agency coordination and procurement clarity when vendors, external models, or external data sources are involved. Finally, Paul and Rena (2025) and Rana et al. (2024) extend the discussion to the governance implications of newer AI capabilities, including generative AI (GenAI), linking ethical safeguards and policy expectations to how organisations adopt these tools and how downstream impacts are managed in high-accountability environments.

4.3.3 Finance and fintech

Eight studies engaged with finance and fintech contexts (Albalawee & Fahoum, 2024; AlTawil & Rahhal, 2025; Lacmanovic & Skare, 2025; Luo et al., 2025; Munn, 2023; Okun et al., 2023; Sandeep et al., 2025; Zada et al., 2024). Across this set, governance is shaped by long-standing financial traditions of model risk management, auditability, and regulatory oversight, but the papers approach these concerns from different angles. Munn (2023) foregrounds governance as something that must be made operational through concrete documentation and control practices rather than remaining aspirational, while Luo et al. (2025) reinforces how data governance and ongoing monitoring underpin trustworthy decision-making in data-intensive environments.

Okun et al. (2023) highlights the governance implications of bias and representativeness (and the organisational responsibility to detect and respond to disparate impacts), whereas Albalawee and Fahoum (2024) contributes a legal and governance lens that emphasises the importance of statutory and corporate governance settings when AI is introduced into regulated organisational environments. Lacmanovic and

Skare (2025) and Zada et al. (2024) both emphasise explainability and documentation, with Zada et al. (2024) additionally bringing GenAI-specific considerations and ethical dilemmas into focus for fintech use cases. Sandeep et al. (2025) links effectiveness to organisational capability and “foundational” controls, including clear roles and strong data practices, while AlTawil and Rahhal (2025) adds a complementary compliance perspective by examining how legal and corporate governance frameworks can be aligned with broader responsibility expectations in business practice.

4.3.4 Workforce, HR, and recruitment

Only two studies focused explicitly on workforce and recruitment contexts (Mujtaba, 2025; Sandeep et al., 2025), yet their findings were salient because hiring and workforce systems concentrate risks of discrimination and opacity. These studies emphasised the need for demonstrable fairness testing, clear responsibility for adverse impact review, and careful boundary-setting on what model outputs may be used for decision-making (Mujtaba, 2025; Sandeep et al., 2025).

4.3.5 Migration and justice-related domains

Three studies that focused on migration and justice-related domains (Bensa & Fattore, 2024; Emah & Bennett, 2025; Munn, 2023) framed governance as having high stakes because of the possibility of rights violations and compounding disadvantage. The evidence emphasised that contestability and transparency are essential, but they must be operationalised through independent oversight and procedural safeguards, especially when AI outputs influence state decisions that are coercive or exclusive (Emah & Bennett, 2025; Nalbandian, 2022). When considered collectively, the sectoral evidence demonstrates that "one-size-fits-all" governance mechanisms are not possible. This conclusion serves as the inspiration for Theme 3, which shifts the focus from sectoral variation to the organisational capabilities that allow governance mechanisms to operate consistently in real-world scenarios.

4.4 Theme 3: Organisational Capacity and Culture

Theme 3 consolidates the evidence on organisational conditions that enable responsible AI governance, drawing on 31 studies. Across this evidence base, governance is treated as a socio-technical capability rather than a standalone policy: studies show that adoption and institutionalisation depend on leadership mandate,

resourcing, and integration into existing risk and decision structures, including clear decision rights and organisational accountability (Alzebda & Matar, 2025; Bensa & Fattore, 2024; Khan et al., 2022; Nalbandian, 2022; Rana et al., 2024; Tian et al., 2025; Wamba et al., 2025; Yang et al., 2024; Yue et al., 2024). They also show that capability building, literacy, and culture are not "nice-to-haves" but practical preconditions for governance to work day to day, spanning technical competence (for example evaluation and monitoring), ethical and legal literacy, cross-functional collaboration, and a speak-up culture that supports escalation and learning (Birkstedt et al., 2023; Ismail & Goh, 2024; Loufek et al., 2024; Mäntymäki et al., 2022; Monyela & Tella, 2024; Mujtaba, 2025; Sandeep et al., 2025; Zada et al., 2024). A third cluster makes clear that credible governance rests on data governance, privacy, and security foundations, such as privacy, security, provenance, consent, and data quality, particularly in high-stakes health and citizen-facing contexts where downstream harms and trust impacts are acute (Albalawee & Fahoum, 2024; Batool et al., 2025; Diaz-Asper et al., 2024; Luo et al., 2025; Okun et al., 2023; Olawade et al., 2025; Paul & Rena, 2025; Rugiubei & Stoica, 2025; Saw & Ng, 2022; Zhao et al., 2021). Several studies further highlight that these organisational conditions are shaped by the wider governance environment, including the practical work of translating external expectations (regulation, standards, cross-jurisdiction requirements) into internal routines and evidence practices (Khadka & Ullah, 2025; Ridzuan et al., 2024; D. Zhou et al., 2025). Finally, the evidence base includes a cautionary note that governance can become symbolic when incentives favour appearance over substance, reinforcing why organisational capacity and culture ultimately determine whether governance mechanisms genuinely shape real decisions (Palladino, 2023).

4.4.1 Institutionalisation and organisational adoption

Eighteen studies examined how responsible AI governance is adopted and institutionalised. Across this set, institutionalisation is consistently portrayed as a dual process: "top-down" direction, such as leadership mandate, funding, decision rights, and integration with enterprise risk and strategy, must be matched by "bottom-up" translation into routines that teams can actually carry out (Khan et al., 2022; Tian et al., 2025; Yang et al., 2024; Yue et al., 2024; D. Zhou et al., 2025). Several studies show that adoption is also shaped by context: public-sector and rights-oriented accounts emphasise legitimacy, due process, procurement, and cross-agency coordination as practical conditions for making governance stick (Alzebda & Matar, 2025; Bensa & Fattore, 2024; Nalbandian, 2022; Rana et al., 2024), while sectoral exemplars illustrate how

institutionalisation often “piggy-backs” on existing regimes, clinical governance and safety norms in healthcare (Okun et al., 2023; Saw & Ng, 2022) and model risk management and regulatory scrutiny in finance/fintech (Ridzuan et al., 2024; Zada et al., 2024). Organisational learning and continuous improvement are also recurring motifs: governance matures through iteration, documentation, incident response, and the stabilisation of repeatable practices, including data stewardship and security routines that keep governance claims credible over time (Monyela & Tella, 2024; Rugiubei & Stoica, 2025; Sandeep et al., 2025; Tian et al., 2025; D. Zhou et al., 2025). Importantly, the evidence base includes a cautionary counterpoint: governance can appear institutionalised on paper while remaining weakly coupled to real decisions, particularly when it becomes symbolic or performative (Palladino, 2023). Finally, several studies position institutionalisation within broader organisational transformation, arguing that governance becomes durable when it is embedded into everyday capability building and innovation practices rather than treated as a standalone ethics exercise (Khan et al., 2022; Wamba et al., 2025).

4.4.2 Capability building, literacy, and culture

Fourteen studies foregrounded capability building and organisational culture as preconditions for effective governance. Across this set, “capability” is treated as more than technical skill: it includes staff competence in evaluation and monitoring, ethical and legal literacy, and the ability to work across disciplines so that risks are identified early and escalated appropriately (Albalawee & Fahoum, 2024; Khadka & Ullah, 2025; Luo et al., 2025; D. Zhou et al., 2025). Four studies highlight the organisational side of capability, how governance becomes credible when it is institutionalised through leadership commitment, clear expectations, and a culture that rewards openness about limitations rather than performance theatre (Khan et al., 2022; Mäntymäki et al., 2022; Monyela & Tella, 2024; Yue et al., 2024). Others make capability more concrete by proposing staged maturity or roadmaps, suggesting organisations should build governance progressively and avoid over-claiming responsibility before the necessary people, processes, and evidence practices are in place (Birkstedt et al., 2023; Ismail & Goh, 2024; Loufek et al., 2024). The set also shows how capability needs shift by application: workforce decision-making and hiring concentrate fairness and accountability demands (Mujtaba, 2025; Sandeep et al., 2025), while newer generative AI use cases intensify training needs around acceptable use, misuse risk, and governance literacy beyond traditional model management (Zada et al., 2024).

4.4.3 Data governance, privacy, and security foundations

Nineteen studies emphasised data governance, privacy, and security as foundational to responsible AI governance. Collectively, these studies link data stewardship to both performance and legitimacy, arguing that AI outputs are only as reliable and fair as the upstream data-generating process, data collection and labelling decisions, consent and secondary-use controls, and access management (Batool et al., 2025; Ismail & Goh, 2024; Khan et al., 2022; Yue et al., 2024). Several studies illustrate this "foundational" claim in domain settings where poor data governance directly undermines safety and trust: in healthcare and clinical-adjacent contexts, data protection and consent are treated as preconditions for ethical clinical decision support and patient-facing legitimacy (Okun et al., 2023; Olawade et al., 2025; Paul & Rena, 2025; Saw & Ng, 2022; Zhao et al., 2021), while finance and fintech studies highlight that data governance enables auditability, compliance, and fairness in high-volume consequential decisions (Albalawee & Fahoum, 2024; Luo et al., 2025; Ridzuan et al., 2024; D. Zhou et al., 2025). A number of contributions also draw attention to the operational weak points that cause governance to fail in practice, especially gaps in consent management and secondary data use (Okun et al., 2023; Saw & Ng, 2022; Zhao et al., 2021), limited provenance and traceability (Diaz-Asper et al., 2024; Rugiubei & Stoica, 2025), and data quality issues that compromise monitoring and accountability over time (Luo et al., 2025; Olawade et al., 2025). Public-sector and cross-jurisdiction perspectives further stress that data governance is inseparable from procurement and inter-organisational arrangements, because vendor models and cross-agency data flows intensify accountability requirements and the consequences of misuse (Alzebda & Matar, 2025; Khadka & Ullah, 2025; Nalbandian, 2022). Finally, Palladino (2023) offers a cautionary note that data governance can be performed rhetorically yet remain weakly coupled to practice, reinforcing the wider Theme 3 conclusion that organisational capacity is the enabling substrate for governance: without mature data foundations, privacy/security controls, and evidence practices, responsible AI commitments are difficult to sustain or demonstrate credibly.

4.5 Theme 4: Accountability and Multi-Level Governance

Theme 4 synthesises how accountability is allocated, exercised, and evidenced within organisations and across external governance environments, drawing on 32 studies in total (Albalawee & Fahoum, 2024; AlTawil & Rahhal, 2025; Alzebda & Matar, 2025; Attard-Frost et al., 2023; Batool et al., 2025; Bensa & Fattore, 2024; Birkstedt et al., 2023; Diaz-Asper et al., 2024; Ismail & Goh, 2024; Khadka & Ullah, 2025;

Khan et al., 2022; Lacmanovic & Skare, 2025; Loufek et al., 2024; Luo et al., 2025; Mac, 2024; Mäntymäki et al., 2022; Monyela & Tella, 2024; Mujtaba, 2025; Munn, 2023; Okun et al., 2023; Palladino, 2023; Paul & Rena, 2025; Rana et al., 2024; Ridzuan et al., 2024; Rugiubei & Stoica, 2025; Sandeep et al., 2025; Wamba et al., 2025; Yang et al., 2024; Zada et al., 2024; Zhao et al., 2021; D. Zhou et al., 2025; T. Zhou et al., 2025). Taken together, these studies show that "accountability" is not a single mechanism but a multi-level arrangement, internal decision structures plus assurance evidence plus external alignment. Several papers warn that, without real authority and scrutiny, accountability can slip into symbolic compliance rather than shaping decisions in practice (Attard-Frost et al., 2023; Munn, 2023).

4.5.1 Internal accountability structures

Eighteen studies in this dataset described internal accountability structures, such as named roles (for example, responsible AI leads), ethics or governance committees, and formal boards, as the practical "machinery" that turns abstract principles into enforceable decision-making (AlTawil & Rahhal, 2025; Attard-Frost et al., 2023; Birkstedt et al., 2023; Diaz-Asper et al., 2024; Ismail & Goh, 2024; Khan et al., 2022; Lacmanovic & Skare, 2025; Loufek et al., 2024; Mäntymäki et al., 2022; Monyela & Tella, 2024; Mujtaba, 2025; Munn, 2023; Okun et al., 2023; Rana et al., 2024; Ridzuan et al., 2024; Sandeep et al., 2025; Wamba et al., 2025; D. Zhou et al., 2025). Across the set, these structures are consistently framed as mechanisms that allocate decision rights and escalation pathways: they specify who can approve deployment, who can pause or remediate a system, and how disagreements (for example, between innovation goals and harm reduction) are adjudicated (Attard-Frost et al., 2023; Khan et al., 2022; Mäntymäki et al., 2022). Several studies treat these bodies as safeguards against symbolic compliance, arguing that principles are unlikely to constrain behaviour unless there is a designated group with authority to ask for evidence, challenge assumptions, and require follow-up actions (Attard-Frost et al., 2023; Munn, 2023). Others emphasise how accountability is made workable: committees need access to technical information and operational artefacts (for example, documentation and evaluation results), and they must be embedded in routine governance rhythms rather than operating as an ad hoc "ethics add-on" (Birkstedt et al., 2023; Ismail & Goh, 2024; D. Zhou et al., 2025). Domain-focused papers illustrate why these structures matter in high-stakes settings: in healthcare, internal accountability is tied to clinical governance and patient safety (Loufek et al., 2024; Okun et al., 2023), while in finance and audit-oriented contexts, committees and accountable

roles are positioned as essential for defensible explainability, model risk sign-off, and auditability (AlTawil & Rahhal, 2025; Lacmanovic & Skare, 2025; Ridzuan et al., 2024). Workforce and HR oriented governance highlights accountability for adverse impact review and escalation when discrimination risks arise (Mujtaba, 2025; Sandeep et al., 2025). Public-sector and cross-context studies further stress that internal structures must align with democratic accountability expectations and, in practice, often require coordination across organisational boundaries (Diaz-Asper et al., 2024; Rana et al., 2024). Finally, capability and culture oriented contributions underline that formal structures only function as intended when staff are willing and able to surface concerns, so accountability is as much about organisational "speak-up" conditions as it is about the existence of committees on paper (Monyela & Tella, 2024; Wamba et al., 2025).

4.5.2 Assurance mechanisms and evidence of compliance

Eight studies discussed assurance mechanisms such as audits, monitoring programmes, and formal evaluation processes. Collectively, they treat assurance as a credibility device: governance claims carry more weight when organisations can show evidence of testing, documentation, monitoring, and remediation rather than relying on stated principles alone. Within this set, Munn (2023) positions assurance as a safeguard against "paper governance", stressing the importance of demonstrable scrutiny; Lacmanovic and Skare (2025) anchors assurance in finance-style expectations for explainability, documentation, and review that can withstand regulatory and audit scrutiny; and Mujtaba (2025) highlights assurance in workforce settings through fairness testing and ongoing checks for adverse impact. Several studies emphasise continuous monitoring as the practical core of assurance: Luo et al. (2025) and D. Zhou et al. (2025) foreground monitoring and evaluation over time (including post-deployment oversight) so performance, drift, and unintended effects can be detected and addressed, while Yang et al. (2024) explicitly links monitoring outputs to organisational learning, where findings feed back into updated policies, controls, and governance routines. Public-sector oriented work adds an additional emphasis on assurance as accountability to external stakeholders: T. Zhou et al. (2025) and Alzebda and Matar (2025) highlight the need for assurance practices that can demonstrate responsible operation in environments shaped by procurement, vendor dependencies, and regulatory or public expectations, reinforcing the common conclusion across the eight studies that assurance is both evidence for compliance and a mechanism for continuous improvement.

4.5.3 External regulation and standards alignment

A substantial majority of the evidence base (25 studies) engaged with external regulation or standards alignment, indicating that accountability for responsible AI is frequently shaped "from the outside in", rather than being purely an internal design choice. Across these studies, external governance is repeatedly treated as a structuring context that sets the language, minimum expectations, and evidence requirements that organisations then translate into internal policies, documentation, and decision rights (for example, through defined governance architectures and role clarity, or through operational controls and reporting routines) (Birkstedt et al., 2023; Lacmanovic & Skare, 2025; Mäntymäki et al., 2022; Munn, 2023; Wamba et al., 2025). A consistent practical message is that alignment work is not automatic: organisations must interpret external requirements and make them workable through local controls, particularly when operating across jurisdictions or when aligning multiple standards, such as by mapping regulatory expectations into templates, audit-ready documentation, and measurable controls (Diaz-Asper et al., 2024; Ismail & Goh, 2024; Khadka & Ullah, 2025; Loufek et al., 2024; Rugiubei & Stoica, 2025). Sector-focused papers show why this matters: in healthcare and other high-stakes contexts, alignment is tied to privacy, consent, safety, and clinical accountability (Bensa & Fattore, 2024; Luo et al., 2025; Okun et al., 2023; Paul & Rena, 2025; Zhao et al., 2021), while finance and fintech studies emphasise documentation, explainability, model risk management traditions, and supervisory scrutiny (Albalawee & Fahoum, 2024; AlTawil & Rahhal, 2025; Lacmanovic & Skare, 2025; Ridzuan et al., 2024; Sandeep et al., 2025; Zada et al., 2024). Public-sector and citizen-facing accounts further underline that alignment often involves procurement realities and vendor dependencies, increasing the need for clear accountability chains and defensible evidence of compliance (Alzebda & Matar, 2025; Mac, 2024; Rana et al., 2024). Finally, several studies explicitly caution that "alignment" can degrade into tick-box compliance if external standards are treated as reputational cover rather than coupled with meaningful oversight, independent challenge, and consequences for non-compliance (Mac, 2024; Munn, 2023; Palladino, 2023; Rana et al., 2024).

4.6 Theme 5: Characteristics of Effective Governance Frameworks

Theme 5 brings together the included studies proposed design features of effective responsible AI governance and the outcomes used to judge whether governance "works" in practice, and it is supported by 36 of the 38 included studies. Within this broad set, four studies make risk-based/proportional governance

explicit, arguing that governance strength should match the severity and likelihood of harm (Loufek et al., 2024; Palladino, 2023; Ridzuan et al., 2024; Tian et al., 2025). Fairness, equity, and discrimination are far more pervasive (24 studies), with the collective message that fairness is not only a technical testing task but also an organisational accountability issue that depends on choices about data, decision rights, and escalation when harms appear (Khan et al., 2022; Mac, 2024; Mujtaba, 2025; Rana et al., 2024; T. Zhou et al., 2025). A smaller subset focuses on generative AI-specific challenges (four studies), highlighting risks linked to scale, opacity, misuse, and shifting acceptable-use boundaries (Paul & Rena, 2025; Rana et al., 2024; Wamba et al., 2025; Zada et al., 2024). A large proportion of the theme (26 studies) connects governance design to effectiveness outcomes, especially trust, legitimacy, performance, and public value, and stresses that effectiveness claims are strongest when organisations can show evidence of monitoring, response, and improvement rather than relying on policy statements alone (Luo et al., 2025; Munn, 2023; Yang et al., 2024).

4.6.1 Risk-based and proportional governance

Four studies in the dataset treated risk-based or proportional governance as an explicit design principle. Collectively, they argue that governance should be scaled to the level of risk and potential harm, so higher-stakes AI systems trigger tighter requirements, for example, deeper testing, stronger sign-off, and closer monitoring, while lower-risk uses do not absorb the same level of compliance effort. In finance, Ridzuan et al. (2024) frames proportionality as a practical way to balance innovation with regulatory and ethical obligations in a highly supervised environment, where the "right" level of control must be defensible to both internal audit and external regulators. Tian et al. (2025) similarly positions proportionality as a resource-sensitive logic for governance design, matching oversight intensity to application criticality rather than applying one uniform standard to all AI use cases. Loufek et al. (2024) contributes a concrete healthcare case example, showing how proportional governance can be operationalised through institution-level accountability structures so that higher-risk clinical uses receive stronger scrutiny and clearer decision rights. Finally, Palladino (2023) provides a more cautionary lens: risk-based framing can drift into "checkbox" compliance if risk categories are vague or politically negotiated, which makes proportionality credible only when the risk thresholds, evidence requirements, and escalation rules are transparent and contestable.

4.6.2 Fairness, equity, and discrimination

Fairness and discrimination concerns were among the most pervasive issues in the evidence base (n = 24; see Table 6 below for the full list of contributing studies). Across these studies, fairness was framed as both a technical and an institutional problem: while bias metrics and testing are necessary, governance must also address upstream data generation, organisational decision-making practices, and accountability for disparate impacts. Several studies emphasised that fairness work is inherently normative and contestable, requiring stakeholder engagement and explicit value judgements rather than assuming a single universally "correct" metric (Khan et al., 2022; Mujtaba, 2025; T. Zhou et al., 2025). Studies in public and justice-adjacent settings further highlighted the risk of compounding structural inequities when AI is deployed in contexts already shaped by historical disadvantage (Emah & Bennett, 2025; Mac, 2024; Nalbandian, 2022).

4.6.3 Governance challenges specific to generative AI

Four studies raised governance issues specific to generative AI (Paul & Rena, 2025; Rana et al., 2024; Wamba et al., 2025; Zada et al., 2024). These studies highlighted challenges associated with rapid capability diffusion, unpredictable output behaviour, intellectual property and data leakage risks, and the difficulty of specifying acceptable use in contexts where models are increasingly general-purpose. The evidence therefore suggests that GenAI heightens the need for clear usage boundaries, staff training, and monitoring regimes that detect both technical failures and misuse (Paul & Rena, 2025; Rana et al., 2024; Wamba et al., 2025; Zada et al., 2024)

4.6.4 Effectiveness outcomes and evaluative criteria

Twenty six studies in the review treat governance "effectiveness" as something that must be demonstrated in practice, not simply claimed in policy: they link effectiveness to whether an organisation can show that its AI is safe, fair, and accountable in the real setting where it is used (Albalawee & Fahoum, 2024; AlTawil & Rahhal, 2025; Alzebda & Matar, 2025; Batool et al., 2025; Bensa & Fattore, 2024; Khadka & Ullah, 2025; Lacmanovic & Skare, 2025; Loufek et al., 2024; Luo et al., 2025; Mäntymäki et al., 2022; Monyela & Tella, 2024; Mujtaba, 2025; Munn, 2023; Okun et al., 2023; Olawade et al., 2025; Paul & Rena, 2025; Rana et al., 2024; Rugiubei & Stoica, 2025; Sandeep et al., 2025; Saw & Ng, 2022; Wamba et al., 2025; Yang et al., 2024; Yue et al., 2024; Zada et al., 2024; Zhao et al., 2021; D. Zhou et al., 2025). Collectively, these studies

suggest that outcomes such as stakeholder trust, legitimacy, organisational performance, and public value tend to improve when governance is visible and reviewable (for example, through documentation, monitoring, audit trails, and clear decision rights), particularly in high-stakes settings like healthcare and public administration (Bensa & Fattore, 2024; Loufek et al., 2024; Okun et al., 2023; Olawade et al., 2025; Rana et al., 2024; Saw & Ng, 2022; Zhao et al., 2021) and in regulated sectors such as finance where auditability and explainability are central to credibility (Albalawee & Fahoum, 2024; AlTawil & Rahhal, 2025; Lacmanovic & Skare, 2025; Luo et al., 2025; Zada et al., 2024). At the same time, several papers warn that effectiveness can be overstated when evaluation relies on proxy indicators (such as the mere existence of policies or standards alignment) rather than evidence of changed practice, learning from incidents, and measurable improvements in outcomes (Batool et al., 2025; Munn, 2023; Wamba et al., 2025; Yang et al., 2024; Yue et al., 2024; D. Zhou et al., 2025).

4.7 Conclusion

Across Themes 1 thru 5, the evidence portrays responsible AI governance as a layered socio-technical system, that responsible AI governance is less about declaring ethical intentions and more about making them workable in everyday decisions. Theme 1 shows how organisations try to turn principles into practice through operational tools, lifecycle controls, and routines, but what is striking is how few studies describe concrete instruments such as checklists, impact assessments, and standardised documentation in any detail, suggesting a persistent “how-to” gap. Theme 2 demonstrates that governance is highly context-dependent across sectors (such as healthcare, public services, finance, workforce, justice), with the interesting pattern that “good” governance is defined by the specific risks, duties, and harms of each domain rather than by a universal model. Theme 3 highlights that governance only becomes credible when organisational capacity exists, for example leadership commitment, staff capability, culture, and strong data governance, yet it is how often frameworks assume these conditions rather than showing how they are built and sustained. Theme 4 consolidates accountability as the backbone of governance through internal decision rights, assurance evidence, and external standards alignment, but it is notable that assurance mechanisms (audits and monitoring) are less consistently specified than committees, roles, and compliance language, which can leave accountability more formal than demonstrable. Theme 5 brings together what “effective” governance looks like, such as (risk-based proportionality, fairness, GenAI-specific challenges, and outcomes such as trust, and

the most surprising finding is that explicit proportional, risk-tiered design appears relatively rarely despite being widely implied, while effectiveness is often judged using indirect proxies (like policy adoption) rather than clear evidence of changed practice or improved outcomes.

A summary of the main themes and subthemes can be seen in Table 6 below, which maps each higher-order theme to its supporting subthemes, shows how many included studies (n) contributed to each subtheme, and lists the contributing studies to make the evidence trail transparent. The next chapter builds on these findings to develop a discussion of what constitutes effective AI governance in organisations, and to articulate how the reviewed evidence informs the dissertation's conceptual framing and practical recommendations.

Table 6. Subtheme-to-study mapping

Theme	Subtheme	n studies	Contributing studies
Theme 1: Translating principles into practice	§4.2.1 Operationalisation of measurement instruments (e.g. checklists, impact assessments, model documentation, audits)	2	Munn (2023); Lacmanovic and Skare (2025)
Theme 1: Translating principles into practice	§4.2.2 Lifecycle embedding and continuous governance (e.g. MLOps, deployment gates, monitoring)	11	Mäntymäki et al. (2022); T. Zhou et al. (2025); Luo et al. (2025); Bensa and Fattore (2024); Mujtaba (2025); D. Zhou et al. (2025); Tian et al. (2025); Ismail and Goh (2024); Birkstedt et al. (2023); Alzebda and Matar (2025); Mac (2024)
Theme 1: Translating principles into practice	§4.2.3 Principles-to-practice failure and symbolic compliance (e.g. ethics washing, symbolic compliance)	3	Attard-Frost et al. (2023); Munn (2023); Palladino (2023)
Theme 2: Sector-specific governance models	§4.3.1 Healthcare and clinical-adjacent contexts	10	Olawade et al. (2025); Luo et al. (2025); Okun et al. (2023); Zhao et al. (2021); Paul and Rena (2025); Emah and Bennett (2025); Bensa and Fattore (2024); Diaz-Asper et al. (2024); Saw and Ng (2022); Loufek et al. (2024)
Theme 2: Sector-specific governance models	§4.3.2 Public sector and administrative AI	6	T. Zhou et al. (2025); Paul and Rena (2025); Bensa and Fattore (2024); Alzebda and Matar (2025); Mac (2024); Rana et al. (2024)
Theme 2: Sector-specific governance models	§4.3.3 Finance and fintech	8	Munn (2023); Luo et al. (2025); Okun et al. (2023); Albalawee and Fahoum (2024); Lacmanovic and Skare (2025); AlTawil and Rahhal (2025); Sandeep et al. (2025); Zada et al. (2024)
Theme 2: Sector-specific governance models	§4.3.4 Workforce, HR, and recruitment	2	Mujtaba (2025); Sandeep et al. (2025)
Theme 2: Sector-specific governance models	§4.3.5 Migration and justice-related domains	3	Munn (2023); Nalbandian (2022); Emah and Bennett (2025)

Theme 3: Organisational capacity and culture	§4.4.1 Institutionalisation and organisational adoption	18	Khan et al. (2022); Nalbandian (2022); Monyela and Tella (2024); Okun et al. (2023); Ridzuan et al. (2024); Bensa and Fattore (2024); D. Zhou et al. (2025); Tian et al. (2025); Yang et al. (2024); Rugiubei and Stoica (2025); Sandeep et al. (2025); Alzebda and Matar (2025); Zada et al. (2024); Saw and Ng (2022); Yue et al. (2024); Palladino (2023); Wamba et al. (2025); Rana et al. (2024)
Theme 3: Organisational capacity and culture	§4.4.2 Capability building, literacy, and culture	14	Khan et al. (2022); Mäntymäki et al. (2022); Khadka and Ullah (2025); Monyela and Tella (2024); Luo et al. (2025); Mujtaba (2025); D. Zhou et al. (2025); Albalawee and Fahoum (2024); Ismail and Goh (2024); Sandeep et al. (2025); Birkstedt et al. (2023); Zada et al. (2024); Yue et al. (2024); Loufek et al. (2024)
Theme 3: Organisational capacity and culture	§4.4.3 Data governance, privacy, and security foundations	19	Khan et al. (2022); Batool et al. (2025); Nalbandian (2022); Khadka and Ullah (2025); Olawade et al. (2025); Luo et al. (2025); Okun et al. (2023); Zhao et al. (2021); Ridzuan et al. (2024); Paul and Rena (2025); D. Zhou et al. (2025); Albalawee and Fahoum (2024); Diaz-Asper et al. (2024); Ismail and Goh (2024); Rugiubei and Stoica (2025); Alzebda and Matar (2025); Saw and Ng (2022); Yue et al. (2024); Palladino (2023)
Theme 4: Accountability and multi-level governance	§4.5.1 Internal accountability structures (e.g. roles, committees, boards)	18	Attard-Frost et al. (2023); Munn (2023); Khan et al. (2022); Mäntymäki et al. (2022); Monyela and Tella (2024); Okun et al. (2023); Ridzuan et al. (2024); Mujtaba (2025); D. Zhou et al. (2025); Diaz-Asper et al. (2024); Ismail and Goh (2024); Lacmanovic and Skare (2025); AlTawil and Rahhal (2025); Sandeep et al. (2025); Birkstedt et al. (2023); Loufek et al. (2024); Wamba et al. (2025); Rana et al. (2024)
Theme 4: Accountability and multi-level governance	§4.5.2 Assurance mechanisms and evidence of compliance (e.g. audits, monitoring, evaluation)	8	Munn (2023); T. Zhou et al. (2025); Luo et al. (2025); Mujtaba (2025); D. Zhou et al. (2025); Yang et al. (2024); Lacmanovic and Skare (2025); Alzebda and Matar (2025)
Theme 4: Accountability and multi-level governance	§4.5.3 External regulation/standards alignment	25	Munn (2023); Mäntymäki et al. (2022); Batool et al. (2025); Khadka and Ullah (2025); Luo et al. (2025); Okun et al. (2023); Zhao et al. (2021); Ridzuan et al. (2024); Paul and Rena (2025); Bensa and Fattore (2024); Albalawee and Fahoum (2024); Diaz-Asper et al. (2024); Ismail and Goh (2024); Rugiubei and Stoica (2025); Lacmanovic and Skare (2025); AlTawil and Rahhal (2025); Sandeep et al. (2025); Birkstedt et al. (2023); Alzebda and Matar (2025); Zada et al. (2024); Palladino (2023); Loufek et al. (2024); Mac (2024) Wamba et al. (2025); Rana et al. (2024)

Theme 5: Characteristics of effective governance frameworks	§4.6.1 Risk-based and proportional governance	4	Ridzuan et al. (2024); Tian et al. (2025); Palladino (2023); Loufek et al. (2024);
Theme 5: Characteristics of effective governance frameworks	§4.6.1 Fairness, equity, and discrimination	24	Attard-Frost et al. (2023); Munn (2023); Khan et al. (2022); Batool et al. (2025); T. Zhou et al. (2025); Monyela and Tella (2024); Olawade et al. (2025); Okun et al. (2023); Ridzuan et al. (2024); Emah and Bennett (2025); Mujtaba (2025); D. Zhou et al. (2025); Yang et al. (2024); Diaz-Asper et al. (2024); Ismail and Goh (2024); Lacmanovic and Skare (2025); Sandeep et al. (2025); Alzebda and Matar (2025); Zada et al. (2024); Saw and Ng (2022); Palladino (2023); Mac (2024); Rana et al. (2024)
Theme 5: Characteristics of effective governance frameworks	§4.6.3 Governance challenges specific to generative AI	4	Paul and Rena (2025); Zada et al. (2024); Wamba et al. (2025); Rana et al. (2024)
Theme 5: Characteristics of effective governance frameworks	§4.6.4 Effectiveness outcomes and evaluative criteria (e.g. trust, performance, value)	26	Munn (2023); Mäntymäki et al. (2022); Batool et al. (2025); Khadka and Ullah (2025); Monyela and Tella (2024); Olawade et al. (2025); Luo et al. (2025); Okun et al. (2023); Zhao et al. (2021); Paul and Rena (2025); Bensa and Fattore (2024); Mujtaba (2025); D. Zhou et al. (2025); Yang et al. (2024); Albalawee and Fahoum (2024); Rugiubei and Stoica (2025); Lacmanovic and Skare (2025); AlTawil and Rahhal (2025); Sandeep et al. (2025); Alzebda and Matar (2025); Zada et al. (2024); Saw and Ng (2022); Yue et al. (2024); Loufek et al. (2024); Wamba et al. (2025); Rana et al. (2024)

Chapter 5: Discussion

5.0 Introduction

This chapter interprets the systematic literature review findings (Chapter 4) considering the conceptual framing established in Chapter 2 and the analytic approach described in Chapter 3. The discussion is organised around the five higher order themes generated through thematic synthesis, and it addresses the dissertation's two research questions: RQ1: How do organisations implement and adapt governance frameworks to ensure the ethical use of AI technologies? and RQ2: What are the most effective governance frameworks for ensuring responsible AI use in organisations?

Rather than treating governance frameworks as static templates, the chapter synthesises the evidence as depicting governance as a socio-technical arrangement that must be operationalised through organisational roles, processes, artefacts, and assurance routines across the AI lifecycle. This framing aligns with the recurring argument in the literature that high-level ethical principles, while widely endorsed, are insufficient unless they are translated into workable mechanisms that shape everyday design, deployment, and oversight decisions (Mittelstadt, 2019; Morley et al., 2020; Schiff et al., 2020).

Across Themes 1 to 5, the central interpretive claim is that the "effectiveness" of AI governance is less about adopting a named framework and more about whether governance arrangements become enforceable, measurable, and contextually fitted to the organisation's AI uses, risk exposure, and accountability environment. This interpretation also clarifies how the two research questions relate: RQ1 is fundamentally about implementation and adaptation (how governance is made real in practice), while RQ2 concerns the properties of frameworks that the literature portrays as reliable for responsible AI outcomes. The chapter therefore treats "effective frameworks" not as a single best model, but as configurations that consistently combine lifecycle controls, organisational capability, and multi-level accountability.

Chapter 5: Discussion

5.1 Theme 1: Translating Principles into Practice

Chapter 2 identified a persistent "principles-to-practice" gap in AI ethics and governance: organisations frequently adopt high-level commitments yet struggle to embed them into operational routines (Mittelstadt, 2019; Morley et al., 2020; Schiff et al., 2020). The review findings reinforce this diagnosis by showing that many accounts of governance remain stronger on aspirations than on implementation detail. A critical contribution of Theme 1 is that it sharpens what "translation" requires: not merely ethical intent, but operational instruments and lifecycle embedding that make governance auditable, contestable, and repeatable.

In the included studies, the most practice-oriented accounts describe governance as being enacted through tangible artefacts, such as documentation, checklists, impact assessments, audit trails, and review gates, combined with continuous monitoring after deployment. This aligns with the argument that governance becomes meaningful when it produces decision records and evidence that can be scrutinised, rather than relying on declarations of values (Raji et al., 2020; Schiff et al., 2020). Yet, an important interpretive tension emerges: while such instruments are frequently recommended, comparatively few studies provide sufficiently concrete descriptions of how organisations design these tools, integrate them into existing workflows, and ensure that completion and review have real consequences. This "how-to scarcity" matters directly for RQ1 because it implies that implementation remains uneven not only due to resistance or inertia, but also because many organisations lack usable models for operationalisation beyond generic prescriptions.

Theme 1 also clarifies the relationship between implementation and symbolic compliance. Prior scholarship has warned that voluntary ethics programmes can become performative, functioning as reputation management rather than as constraint on harmful practices (Attard-Frost et al., 2023; Mittelstadt, 2019). The synthesis supports this concern by indicating that governance is most vulnerable to "ethics washing" when principles are unaccompanied by enforceable routines, clear decision rights, and assurance mechanisms. In such cases, governance artefacts may exist, but they do not reliably change practice. This distinction is central to evaluating effectiveness under RQ2: the literature implicitly treats frameworks as "effective" only when they move beyond principal statements to become operational systems that can interrupt unsafe deployment, require redesign, and mandate corrective action.

At the same time, Theme 1 shows that translation is not impossible; it is organisationally demanding. Illustrative organisational practices include internal handbooks, training programmes, and codified guidance intended to align staff behaviour with stated ethical principles. These practices align with the view that implementation requires organisational learning and capability building, not merely policy publication (Morley et al., 2020). However, the evidence base does not support treating such examples as proof that the principles-to-practice gap has been solved. Instead, the synthesis suggests a more cautious conclusion: translation is occurring in pockets, but it remains inconsistent, particularly when governance controls are not integrated into lifecycle processes (for example, MLOps pipelines and deployment gates) and when oversight lacks authority. This conclusion motivates Theme 2: because translation is shaped by operational realities, governance mechanisms tend to differ across sectors and institutional environments.

5.2 Theme 2: Sector-Specific Governance Models

Theme 2 demonstrates that AI governance is strongly context-dependent: what counts as responsible practice is shaped by sectoral risk profiles, legal duties, professional norms, and the severity and distribution of potential harms. This finding extends Chapter 2's argument that governance cannot be evaluated against a single generic benchmark because accountability expectations differ across domains such as healthcare, finance, and public services (Morley et al., 2020; Qureshi et al., 2024; Whittlestone et al., 2019). In effect, the synthesis suggests that organisations do not simply "apply" a framework; they adapt governance arrangements to fit sectoral constraints, thereby directly answering RQ1 as an adaptation problem rather than a simple adoption problem.

In healthcare and other high-stakes settings, governance is frequently described as being intertwined with clinical governance norms, where legitimacy depends on interpretability, safety, privacy, and mechanisms for oversight that match the gravity of downstream impact. In public sector contexts, governance debates are intensified by concerns about citizen rights, transparency, surveillance, and the accountability deficits associated with automated decision-making (Alston, 2019; Eubanks, 2018; Richardson et al., 2019). These sectoral differences matter because they reshape what "implementation" looks like: governance must align not only with organisational priorities but also with external scrutiny and public accountability. Accordingly, the most defensible interpretation under RQ2 is not that one framework is universally best, but that

governance is most credible when it is tailored to domain-specific harms and duties, with evaluation criteria derived from those sectoral realities rather than from abstract compliance alone.

At the same time, Theme 2 does not imply that sectors are incomparable. Two cross-cutting regularities appear. First, across domains, governance is portrayed as more credible when it couples clear accountability (who decides, who is answerable, what remedial pathways exist) with operational mechanisms (documentation, monitoring, review gates) that make ethical claims testable (Morley et al., 2020; Schiff et al., 2020). Second, the synthesis suggests opportunities for cross-sector learning: regulated sectors may offer transferable discipline in documentation and auditability, while innovation-oriented sectors may offer operational tools and iterative practices, though such transfers require careful recalibration to the receiving sector's risk tolerance and stakeholder expectations. This argument provides a bridge to Theme 3, because sectoral adaptation is constrained by organisational capability: even well-designed governance expectations will fail if organisations cannot resource, sustain, and normalise them.

5.3 Theme 3: Organisational Capacity and Culture

Theme 3 positions governance effectiveness as contingent on organisational readiness: leadership commitment, resourcing, skills, and foundational data governance practices. This finding strengthens and specifies the concerns raised in Chapter 2 that capacity constraints and uneven adoption shape the real-world gap between ethical aspiration and governance maturity (Capgemini Research Institute, 2020; Stahl et al., 2021). Critically, the synthesis suggests that governance frameworks do not function as "plug-and-play" templates. The same nominal framework can yield very different outcomes depending on whether organisations invest in the people, infrastructure, and incentives needed to enact it.

This theme also helps reinterpret the principles-to-practice gap (Theme 1) as, in part, a capability problem. Where organisations lack technical competence in evaluation and monitoring, legal and ethical literacy, or cross-functional coordination, governance artefacts tend to become superficial: policies exist, but lifecycle controls are weak and accountability is difficult to demonstrate. Conversely, where organisations invest in capability-building, ethics and AI literacy training, role clarity, and maturity assessment, governance is more likely to become routinised rather than exceptional. This supports a key answer to RQ1: implementation and

adaptation occur through capacity-building and cultural anchoring, not only through formal framework adoption.

Culture and incentives emerge as particularly decisive. The literature on ethics programmes has long noted that governance fails when formal commitments are undermined by performance systems that reward speed, scale, or profit without regard to downstream harm (Mittelstadt, 2019). In such environments, ethical governance is structurally disadvantaged: staff receive mixed signals about what "counts" in decision-making. Theme 3 therefore implies that a governance framework's effectiveness under RQ2 cannot be judged purely by its content. Effectiveness depends on whether the organisation establishes a "tone at the top", embeds governance into everyday routines, and aligns incentives so that ethical risk management has practical authority rather than rhetorical support.

Finally, Theme 3 foregrounds data governance, privacy, and security as enabling foundations. This is consistent with longstanding arguments that transparency and accountability are constrained when data provenance, consent, and quality are weak (Burrell, 2016; Pasquale, 2015). It also provides a logical bridge to Theme 4: even when organisations build capability and culture, governance remains incomplete unless decision rights and responsibility pathways are clearly structured across internal and external accountability regimes.

5.4 Theme 4: Accountability and Multi-Level Governance

Theme 4 consolidates accountability as the backbone of responsible AI governance. The overview depicts governance as multi-level: internal structures (decision rights, committees, roles), operational assurance practices (audits, monitoring, evaluation), and alignment with external regulation and standards. This theme connects directly to Chapter 2's distinction between soft law approaches (voluntary guidelines and standards) and hard law regulation (binding obligations), and to critiques that soft law can be "toothless" without enforcement and remedy pathways (Butcher & Beridze, 2019; Mittelstadt, 2019; Munn, 2023). In effect, Theme 4 shows that accountability is not simply a principle; it is an institutional design problem.

A central contribution of Theme 4 is that it clarifies why responsibility gaps persist. The literature has long highlighted the difficulty of assigning responsibility for autonomous or unclear system outcomes (Matthias, 2004). The analysis implies that organisations endeavour to tackle this issue through defined governance

roles, cross-functional oversight entities, escalation procedures, and documentation systems that ensure auditability. However, the evidence also indicates a recurring weakness: accountability language often outpaces assurance detail. Many sources describe the presence of committees or policies more readily than they specify how audits are conducted, how monitoring thresholds trigger action, or how harms are detected and remediated over time. This matters because, as noted in Chapter 4, "effectiveness" is frequently assessed through proxies, such as policy existence or standards alignment, rather than through evidence of changed practice or improved outcomes (Batool et al., 2025; Munn, 2023; Yang et al., 2024).

Theme 4 also reinforces the importance of regulatory alignment, particularly as risk-based regulatory approaches become more prominent. Risk-tiered governance is repeatedly presented as a way to match oversight intensity to system impact, requiring stronger documentation and assurance for higher-risk AI uses. Yet the synthesis implies that regulatory fragmentation can also incentivise superficial compliance, especially where requirements are inconsistent across jurisdictions and where organisations treat governance as a checklist rather than as an accountability system. The interpretive implication for RQ2 is therefore conditional: effective frameworks tend to be those that combine principles with enforceable obligations and assurance mechanisms, but they are most credible when they also incorporate enablement, guidance, measurement infrastructure, and capacity support, so that accountability is feasible and not merely formal. This conclusion leads naturally to Theme 5, which synthesises what the included studies imply about "effectiveness" across sectors, organisational contexts, and accountability environments.

5.5 Theme 5: Characteristics of Effective Governance Frameworks

Theme 5 addresses the most evaluative question in the dissertation: what makes governance frameworks effective for responsible AI use. The synthesis does not support the claim that a single named framework is universally superior. Instead, it suggests that effectiveness is best understood as a set of recurring design characteristics, expressed differently depending on organisational context and sectoral risk.

First, effective governance is portrayed as comprehensive across the AI lifecycle. Frameworks are stronger when they extend beyond pre-deployment principles to include deployment decisions, monitoring for drift and emergent harm, incident response, and iterative revision (Schiff et al., 2020). Second, effective frameworks are depicted as enforceable: they allocate decision rights, establish authority to pause or redesign

high-risk deployments, and require auditable artefacts that make governance reviewable and contestable (Raji et al., 2020). Third, effectiveness is linked to adaptability, reflecting that AI systems evolve, contexts change, and governance must respond to new risks and knowledge. Fourth, effective governance is repeatedly tied to proportionality: risk-based design that differentiates obligations based on likely harm. This proportionality is central to contemporary governance thinking because it aims to avoid both under-regulation of high-stakes systems and over-burdening of low-risk use cases.

A further unifying criterion concerns trust. The literature on "Trustworthy AI" frames societal confidence as an ultimate test of governance credibility, but the synthesis indicates that trust is not produced by claims; it is "earned" through transparency, accountability, and demonstrable practice (European Commission, 2019; O'Neill, 2018; OECD, 2019b). On this reading, governance frameworks contribute to trust when they require openness about limitations, mandate scrutiny, and demonstrate learning from incidents. However, Theme 5 also exposes a significant constraint: the evidence base often prescribes what effective governance should include more confidently than it demonstrates, through sustained evaluation, that particular configurations reliably reduce harm or improve outcomes across contexts. This limitation is consistent with Chapter 4's warning that proxy indicators are frequently used in place of validated outcome measures.

Taken together, Themes 1 to 5 provide a coherent answer to the two research questions. Under RQ1: How do organisations implement and adapt governance frameworks to ensure the ethical use of AI technologies, this is achieved through a combination of lifecycle embedding (gates, monitoring, machine learning integration), operational artefacts (documentation, impact assessments, audits), and organisational enablement (capability-building, cultural anchoring, incentive alignment). Under RQ2: What are the most effective governance frameworks for ensuring responsible AI use in organisations, they are those that are lifecycle-spanning, enforceable, risk-proportionate, adaptable, and embedded within multi-level accountability regimes, while being supported by organisational capacity sufficient to enact them.

5.6 Conclusion

This chapter has discussed the systematic review findings through five themes, linking them to the conceptual foundations established in Chapter 2 and answering the dissertation's two research questions:

> **RQ1**: How do organisations implement and adapt governance frameworks to ensure the ethical use of AI technologies?
>
> **RQ2**: What are the most effective governance frameworks for ensuring responsible AI use in organisations?

The synthesis suggests that responsible AI governance is best understood as a socio-technical system that becomes effective only when ethical principles are translated into operational mechanisms, adapted to sector realities, supported by organisational capability and culture, and anchored in multi-level accountability and assurance. A practical implication is that responsible AI governance should be treated as a socio-technical change programme, not only a compliance architecture. Evidence from technology disruption research indicates that employees' perceptions of smart technology and AI-related change are associated with job insecurity, turnover intentions, and wellbeing outcomes (Brougham & Haar, 2018, 2020). Embedding workforce impact assessment, meaningful consultation, and reskilling pathways within governance arrangements is therefore a plausible route to strengthening responsible outcomes in practice (Brougham & Haar, 2017).

At the same time, the review highlights an important constraint on what can be claimed from the current evidence base. Across the included studies, governance is more consistently described than evaluated: the literature often documents the presence of structures and artefacts (for example, policies, committees, documentation regimes, and lifecycle controls) yet provides less independent and longitudinal evidence that these measures reliably reduce harm, improve contestability, or sustain trust in operational settings. As a result, "effectiveness" is frequently inferred from proxy indicators of adoption or alignment rather than demonstrated through outcomes (Mittelstadt, 2019; Morley et al., 2020; Munn, 2023; Raji et al., 2020). In addition, the breadth of sectors and organisational capacities represented in the evidence strengthens contextual insight but limits straightforward generalisation; the findings therefore travel best as transferable tendencies, not universal prescriptions (Morley et al., 2020).

These limitations sharpen the dissertation's forward-looking implications. Section 6.4 methodological limitations and future research below point to the need for evaluative research designs, particularly longitudinal and mixed-method studies that connect governance arrangements to independently observed technical and organisational outcomes, and to more consistent evaluative criteria that enable meaningful comparison across contexts while retaining sensitivity to sectoral constraints (Morley et al., 2020; Raji et al., 2020). Chapter 6 summary and conclusion build on this discussion to consolidate the dissertation's overall contribution and present a concise conclusion aligned with the evidence presented.

Chapter 6: Summary and Conclusion

6.1 Introduction

This final chapter synthesises the dissertation's key insights on responsible and ethical AI governance in organisations, drawing together the literature review, methodology, findings, and discussion. It provides a cohesive account of how the research questions have been answered and what contributions have been made. The chapter is organised into five sections, introduction, a synthesis of included studies summarises the main findings of this research, highlighting how organisations implement and adapt AI governance frameworks and what characteristics make these frameworks effective. Third, a discussion of the study's theoretical and practical significance examines how these findings advance scholarly understanding and offer actionable guidance for practitioners. Fourth, a section for future research critically reflects on the study's constraints, such as scope, data, and design, and suggests avenues for future inquiry to build on this work. finally, final reflections articulate the broader significance of the study and close the thesis with a confident, integrative outlook. Throughout, the emphasis is on consolidating the dissertation's contributions in a mature, critical tone, aligning with the framing of AI governance as a socio-technical challenge requiring both robust scholarship and practical wisdom.

6.2 Synthesis of Included Studies

This dissertation examined how organisations govern AI ethically and what constitutes an effective governance framework for responsible AI use. Guided by two research questions, how organisations implement and adapt governance frameworks (RQ1) and what makes such frameworks effective (RQ2), the systematic literature review and thematic synthesis of 38 studies produced five interlocking themes: translating principles into practice, sector-specific governance models, organisational capacity and culture, accountability and multi-level governance, and the defining characteristics of effective frameworks. Collectively, these themes provide an integrated account of how responsible AI governance is operationalised and how credibility is established in practice.

A central contribution is confirming, and specifying, the persistent principles-to-practice gap in AI ethics. While organisations commonly espouse fairness, transparency, and accountability, the evidence indicates

that these principles remain aspirational unless converted into operational mechanisms and decision processes (Mittelstadt, 2019; Morley et al., 2020). The synthesis shows that translation requires tangible governance instruments, such as documentation standards, impact assessments, monitoring checkpoints, and audit trails, embedded across the AI lifecycle from design through deployment and ongoing oversight (Mittelstadt, 2019; Morley et al., 2020; Raji et al., 2020). Only a minority of studies described these instruments in sufficient detail to demonstrate consistent enactment, underscoring that ethical intent is easier to declare than to institutionalise (Lacmanovic & Skare, 2025; Munn, 2023). This finding directly answers RQ1 by demonstrating that implementation is fundamentally about creating auditable and repeatable practices rather than relying on abstract commitments. It also distinguishes substantive governance from symbolic compliance: where routines are unenforceable, accountability is diffuse, and sanctions or remediation pathways are absent, ethics risks becoming performative "ethics washing" rather than governing behaviour (Attard-Frost et al., 2023; Mittelstadt, 2019).

A second contribution is showing that AI governance is necessarily context dependent. The review indicates that organisations adapt governance arrangements to sectoral risk profiles, legal obligations, and professional norms; consequently, responsible AI in a hospital or public agency is governed through different priorities and safeguards than in financial services or technology firms. Yet the synthesis also identifies cross-sector regularities: governance becomes credible when it couples accountability (clear decision rights, answerability, and remediation) with process mechanisms that generate evidence that controls are functioning as intended. This link between "substance" (alignment with ethical and legal standards) and "proof" (records, audits, monitoring outputs) bridges RQ1 and RQ2, suggesting that effective frameworks are those flexible enough to be tailored to context while remaining anchored in core principles of accountability and transparency (Morley et al., 2020).

Third, the findings foreground organisational capacity and culture as decisive conditions for ethical AI governance. The evidence consistently suggests that frameworks are not "plug-and-play"; they depend on investments in people, workflows, and incentives that enable practice. Without expertise in auditing, mitigation, and data governance, without ethical and legal literacy among decision-makers, and without a culture that legitimises responsible deliberation over speed of deployment, governance is prone to superficial enactment. In this sense, the principles-to-practice gap is partly a capacity gap: effective implementation

requires leadership commitment, workforce buy-in, and institutional support that normalises ethical scrutiny in routine decision-making (Mittelstadt, 2019). This strengthens the answer to RQ2 by underscoring that effectiveness cannot be inferred from a framework's design alone; it must be assessed in relation to an organisation's capability to sustain ethical routines and to treat data governance as a foundational enabler of transparency and accountability.

Fourth, the dissertation consolidates accountability as the backbone of responsible AI governance, operating across internal and external levels. The synthesis clarifies why "responsibility gaps" persist even when organisations establish committees or policies: oversight structures often lack assurance mechanisms that make accountability actionable. Where authority to pause deployments is unclear, where audits are absent or unheeded, and where escalation and remediation are under-specified, accountability remains rhetorical. The review also surfaces the interaction of soft law and hard law. Voluntary guidelines and standards shape aspirations, but they can be ineffective without enforcement, producing governance that is "toothless" in practice (Butcher & Beridze, 2019; Mittelstadt, 2019). Against this backdrop, emerging regulatory approaches, such as risk-based regimes exemplified by the European Union AI Act, signal an increasing conversion of ethical expectations into legal obligations, intensifying the need for organisations to integrate external requirements into internal governance in ways that strengthen, rather than merely formalise, accountability (Butcher & Beridze, 2019).

Finally, the study synthesises recurring characteristics of effective governance frameworks, shifting the discussion away from any single "best" named model and towards shared design features. Effective frameworks are lifecycle-spanning and enforceable, supported by auditable artefacts that enable oversight; they are adaptable to evolving technologies and contexts; and they are risk-proportionate, escalating controls for higher-stakes applications. A unifying element is the role of trust: "trustworthy AI" cannot be achieved through principal statements alone but is earned through demonstrable transparency, contestability, monitoring, and redress mechanisms (European Commission, 2019; O'Neill, 2018). At the same time, the review cautions that the evidence base remains uneven: many studies prescribe desirable governance practices, but rigorous empirical validation of which configurations reliably reduce harms or improve fairness is still emerging (Morley et al., 2020). This limitation tempers claims about effectiveness and

reinforces the need to treat governance as a measurable, evolving organisational capability rather than a settled template.

In sum, the dissertation's core contribution is an evidence-based account of responsible AI governance as a socio-technical system requiring alignment between principles, practices, and people. In relation to RQ1, organisations implement and adapt governance through a combination of lifecycle controls, operational artefacts, and organisational enablers calibrated to context. In relation to RQ2, the most effective frameworks are those that are enforceable and lifecycle-spanning, risk-proportionate and adaptable, and situated within multi-level accountability arrangements, supported by the organisational capacity required to make ethical governance routine rather than exceptional (Mittelstadt, 2019; Morley et al., 2020).

6.3 Theoretical and Practical Significance

6.3.1 Theoretical significance

This dissertation advances AI governance scholarship by offering an integrative synthesis across traditionally siloed fields, computer science, management, law, and ethics, thereby conceptualising governance as a multi-level construct that spans institutional expectations, organisational routines, and technical controls. It adopts a working definition of AI governance as the institutional, organisational, and technical arrangements through which AI use is aligned with ethical principles and legal obligations, providing needed conceptual clarity in a literature where terms and boundaries remain contested (Stahl, 2021). This framing strengthens a socio-technical view of governance by showing that ethical AI cannot be reduced to either abstract principles or technical optimisation alone; rather, it emerges from the coherence of interacting mechanisms across levels of analysis (Stahl, 2021).

The study also deepens theoretical understanding of the principle to practice gap. While earlier scholarship has identified the gap between ethical aspiration and organisational enactment (Mittelstadt, 2019; Morley et al., 2020), the synthesis demonstrates why it persists: principles are routinely undermined by missing tools, ambiguous processes, limited capability, and misaligned incentives. In doing so, it lends support to critiques that warn against ethics initiatives operating as symbolic compliance in the absence of enforceable safeguards (Attard-Frost et al., 2023; Munn, 2023). Conceptually, this reinforces organisational accounts of

decoupling and ethical drift, suggesting that effective AI governance requires explicit accountability structures and a culture in which ethical deliberation is institutionally normalised (Mittelstadt, 2019).

A further contribution is to re-orient the field toward evaluation. The review highlights a prevailing tendency to infer "effectiveness" from proxies such as the existence of ethics boards or published principles, rather than from demonstrable impacts such as reduced harm or improved fairness. Making this gap visible has theoretical value because it challenges the adequacy of current constructs and encourages more rigorous conceptualisations of effectiveness that can be empirically tested (Morley et al., 2020). Finally, the interpretivist, abductive thematic synthesis highlights that governance effectiveness is relational and context-dependent: it cannot be understood by isolating single elements (for example, fairness metrics or policy artefacts) because outcomes arise from the interaction of organisational, technical, and institutional forces (Stahl, 2021).

6.3.2 Practical significance

Practically, this dissertation provides an evidence-informed account of what responsible AI governance looks like when it is operationalised in organisational settings, rather than merely declared. A central implication is that governance is best understood as a socio-technical change programme, not a compliance overlay applied after the fact. Formal policies, oversight committees, and high-level principles are necessary, yet they remain insufficient without workforce engagement, capability-building, and process redesign that embeds ethical checkpoints into everyday AI development and deployment (Mittelstadt, 2019). Accordingly, the evidence base supports a practical orientation in which organisations invest in AI ethics literacy, distribute decision-making competence across technical and non-technical roles, and normalise escalation so that raising concerns about system impacts is safe and expected rather than stigmatised (Mittelstadt, 2019).

The synthesis also distils design criteria that strengthen governance frameworks in practice, namely, lifecycle coverage, enforceability, adaptability, and proportionality, and translates these into implementable controls. In operational terms, this includes risk assessment at inception, explicit monitoring and evaluation plans post-deployment and clearly assigned decision rights that empower designated oversight roles to pause, modify, or retire systems when standards are not met (Morley et al., 2020; Raji et al., 2020). The emphasis on auditable artefacts further implies that organisations should institutionalise documentation and traceability

practices, such as model documentation, decision logs, and incident reporting pathways, so that contested outcomes can be examined, explained, and remediated rather than absorbed as “black box” decisions (Raji et al., 2020). In this sense, governance maturity is demonstrated not by the existence of ethical commitments, but by the organisation’s capacity to evidence, test, and improve ethical performance over time through verifiable practice (Morley et al., 2020; Raji et al., 2020).

Importantly, the practical significance extends beyond organisations to the government’s role, both as a regulator and as a major purchaser and deployer of AI-enabled systems. As a regulator and policy steward, government can shape the conditions under which internal governance becomes substantive by coupling accountability expectations with enabling infrastructure, clear transparency requirements, guidance that clarifies what “good practice” looks like, and safe reporting channels that support disclosure and learning rather than suppressing incident visibility (MBIE, 2025; OECD, 2019a; Ubaldi & Zapata, 2024). Public guidance and risk management frameworks can also reduce fragmentation by providing shared language, evaluative criteria, and lifecycle-oriented controls that organisations can adapt to context, thereby supporting interoperability across sectors and jurisdictions (OECD, 2019a; U.S. Department of Commerce, 2023; Ubaldi & Zapata, 2024). As a market-shaping actor, government procurement can operationalise these expectations by requiring traceability, monitoring, and remediation mechanisms as conditions of contracting, particularly for high-stakes uses where public trust and legitimacy are most vulnerable (MBIE, 2025; O'Neill, 2018). For external stakeholders, including auditors, affected communities, and civil society, the framework produced by this review offers a basis for scrutiny that goes beyond ethical rhetoric, directing attention to auditability, post-deployment monitoring, and accessible pathways for redress, which are fundamental to sustaining trust and social legitimacy (O'Neill, 2018; Raji et al., 2020).

Finally, the review highlights recurrent pitfalls that practitioners and policymakers should avoid, which is equating governance form with a governance function, relying on symbolic structures without authority, and treating effectiveness as assumed rather than measured (Mittelstadt, 2019; Morley et al., 2020). The practical implication is the need for governance performance indicators, such as incident patterns, audit findings, and stakeholder perceptions of fairness and contestability, so that organisations and oversight bodies can evaluate whether governance is working in practice, not merely existing on paper (Morley et al., 2020; U.S. Department of Commerce, 2023). Overall, the practical contribution lies in clarifying how organisations,

alongside government as regulator and exemplar user, can move from ethical aspiration to routinised, accountable practice, supporting AI innovation that is productive while remaining trustworthy and socially defensible (Mittelstadt, 2019; OECD, 2019a, 2019b; Raji et al., 2020).

6.4 Methodological Limitations and Future Research

The conclusions of this systematic literature review should be read in light of methodological constraints that shape what the available evidence can legitimately substantiate. These constraints both qualify interpretation and identify where the field most urgently requires stronger research designs and datasets (Mittelstadt, 2019; Morley et al., 2020).

First, the review was limited to English-language publications from 2015 to July 2025. While this boundary improved feasibility and consistency, it likely under-represents governance practices documented in non-English jurisdictions and in regulatory traditions where key evidence is primarily reported in local languages. The July 2025 cut-off is also high impact in a fast-moving domain: post-2025 developments in generative AI governance, regulation, and standards are necessarily absent. Future syntheses should adopt multilingual searching and planned updating, including "living" review approaches, to strengthen global coverage and temporal relevance.

Second, the evidence base is likely shaped by publication and visibility effects. By prioritising peer-reviewed scholarship and excluding most grey literature, this review may over-weight aspirational frameworks and publicly reportable successes while under-capturing failures, resistance, and null findings, exactly the evidence needed to explain why governance breaks down in practice (Butcher & Beridze, 2019). A more practice-proximate evidence base would integrate high-quality grey sources (for example, standards artefacts, government reports, independent audits, and incident post-mortems), using explicit quality appraisal and transparent inclusion rules.

Third, retrieval and synthesis are constrained by database coverage, indexing conventions, and definitional variation. Relevant work does not always label itself "AI governance", and indexing may disadvantage studies from smaller organisations, emerging economies, and practice-oriented outlets. Moreover, governance is enacted within mixed sectors and organisational capacities, limiting the defensibility of one-size-fits-all prescriptions (Morley et al., 2020; Whittlestone et al., 2019). Accordingly, the findings are best

treated as transferable design tendencies rather than evidence of a single optimal framework (Schiff et al., 2020; Stahl, 2021).

Fourth, the dissertation's single-reviewer design constrains reliability. Screening, extraction, and coding were conducted by one researcher, increasing vulnerability to interpretive bias even where structured decision rules and audit trails are applied (Denyer & Tranfield, 2009; Tranfield et al., 2003). Future reviews should incorporate second-reviewer checks (at least on a subset), explicit consensus procedures for contested decisions, and reported reliability indicators.

Finally, and most importantly, the field remains methodologically better at documenting governance adoption than evaluating governance effects. Across the included studies, structures and artefacts (for example, committees, policies, documentation regimes) are more consistently reported than independent outcome evidence such as harm reduction, improved contestability, or durable trust effects (Mittelstadt, 2019; Morley et al., 2020; Raji et al., 2020). As a result, claims of effectiveness often rest on proxy indicators that may signal intent or maturity without demonstrating operational impact (Munn, 2023; Raji et al., 2020). The Aotearoa New Zealand public sector's Algorithm Charter makes this challenge explicit by framing algorithm use through transparency, accountability, bias mitigation, and Te Tiriti o Waitangi obligations, positioning governance as inseparable from social licence and public trust (Statistics New Zealand, 2020). Recent local experience likewise shows us that formal expectations and oversight do not, on their own, prevent material harm: the ManageMyHealth cyber incident involved unauthorised access to sensitive patient information despite established privacy governance expectations and regulatory settings, prompting response by the National Cyber Security Centre and an inquiry by the Office of the Privacy Commissioner (National Cyber Security Centre, 2026; Office of the Privacy Commissioner, 2026a) The implication is not that governance structures are dispensable, but that their presence cannot be treated as evidence of resilience or ethical performance.

These limitations point to a future research agenda that is more evaluative, comparative, and mechanism focused. The immediate priority is longitudinal and cross-sector study designs that test whether governance interventions change outcomes over time, rather than merely recording artefact presence (Morley et al., 2020; Schiff et al., 2020). Mixed-method designs are especially valuable here, triangulating organisational

evidence (for example, interviews, surveys, and document analysis) with technical assessments (for example, pre/post measures of bias, robustness, error, or security posture) to strengthen causal inference about what governance is doing and under what conditions (Mittelstadt, 2019; Morley et al., 2020). In Aotearoa New Zealand, such evaluation should also operationalise public trust and Treaty-aligned partnership obligations as measurable outcomes, rather than leaving them as aspirational claims (Statistics New Zealand, 2020).

A second priority is shared measurement infrastructure to support cumulative knowledge. More standardised, validated metrics for fairness, transparency, accountability, robustness, and downstream impacts, paired with shared repositories of incidents and near-misses would reduce the gap between governance rhetoric and evaluative capability (Morley et al., 2020; Raji et al., 2020). Effectiveness should also be assessed through human-centred outcomes, not only compliance indicators. Workforce constructs such as STARA awareness, job insecurity, and turnover intentions offer validated measures that can be adapted to evaluate governance as organisational change and to surface distributional impacts across groups (Brougham & Haar, 2017, 2018, 2020).

A third priority is scalable and inclusive governance. Given uneven organisational capacity, future work should test proportional governance designs that remain credible under constraint and examine how governance shapes participation, reskilling, retention, and unequal exposure to harm across sectors and national contexts (Brougham & Haar, 2018, 2020; Morley et al., 2020). Overall, the strongest contribution the field can make next is to move from describing governance to demonstrating, with methodologically defensible evidence, when and why particular governance arrangements succeed or fail across contexts (Stahl, 2021).

6.5 Final Reflection

Writing this dissertation has sharpened, and in some respects unsettled, how I think about technology governance at a time when organisational AI adoption is accelerating faster than the maturation of oversight. When I began, I approached "governance" largely through the lens that has shaped my professional life: as a senior technology practitioner at Auckland University of Technology (AUT) with 16 years' experience delivering complex technology change, where success is often defined by clear scope, disciplined risk management, both stakeholder and staff engagement and dependable delivery. That background has trained me to value control points, decision rights, assurance gates, and traceable artefacts. However, this research has made me confront a deeper, more uncomfortable truth: responsible AI governance cannot be reduced to a well-designed set of controls. It is an ongoing socio-technical capability that depends on organisational incentives, culture, and the moral seriousness with which people treat the consequences of technological decisions (Mittelstadt, 2019; Stahl, 2021).

In my day-to-day role, I have learnt that the difference between "governance on paper" and "governance in practice" is rarely technical. I have seen policies approved and committees convened, yet meaningful accountability can still be diluted when authority is unclear, when delivery pressure overrides caution, or when people do not feel safe to raise concerns that might slow down progress. This dissertation has given me a language for what I have often sensed but not formally named: governance fails when it becomes performative, when compliance theatre substitutes for genuine stewardship (Mittelstadt, 2019). It has also expanded my understanding of what is at stake. With AI systems, the harms that governance must mediate are not confined to cost overruns or service disruption; they can include inequitable outcomes, diminished autonomy, privacy erosion, and loss of institutional trust, which is particularly consequential in public and educational settings (European Commission, 2019; O'Neill, 2018). That shift has changed how I weigh trade-offs. I now see "delivery" not as the endpoint of responsible work, but as the point at which responsibility intensifies, because deployment is where impacts become real, contested, and socially distributed.

This research has therefore changed how I understand my own professional obligations. In the past, I might have treated ethical considerations as adjacent to governance, important, but ultimately outside the core machinery of project control. I no longer think that is defensible. Responsible AI governance, as this

dissertation argues, is fundamentally about alignment and integration: ensuring that organisational decisions, incentives, and routines are systematically aligned with stakeholder values and translated into durable practices that can be audited, challenged, and improved (O'Neill, 2018; Stahl, 2021). For me, that has practical implications in how I lead. I am more deliberate about asking, early, what "good" looks like beyond technical performance; who bears risk if the system is wrong; what evidence we will accept as assurance; and whether we have created genuine pathways for escalation and remediation rather than assuming that "no news is good news" (Mittelstadt, 2019; Raji et al., 2020). It also means I place greater weight on capability-building, helping teams build the literacy to identify ethical risk, and designing processes that normalise questioning rather than treating it as resistance (Stahl, 2021).

These reflections have become even more applied for me in light of AUT's recent institutional move to coordinate its AI direction at the highest level. The Vice-Chancellor's AI Taskforce report, Our AI Future, positions AI as strategically significant for AUT and calls for an integrated, university-wide approach across teaching, research, and operations (Auckland University of Technology, 2025c). Reading that as a practitioner, I interpret it as more than an aspirational statement: it signals that AI governance is high on the Vice-Chancellor's agenda and is being treated as an AUT priority with real implications for how we design, approve, and assure technology change. In parallel, Auckland University of Technology (2025a) establishment of the AI hub as an entry point for information, tools, resources, and guidance has shifted my thinking about what "operationalising governance" can look like in practice. It takes something that can feel abstract, responsible AI, and turns it into usable decision support for staff working under time pressure. In particular, Auckland University of Technology (2025b) guidance that sorts AI tools by trust rating reinforces a lesson that also surfaced strongly across the studies: risk is not uniform, and responsible governance depends on proportionate controls that are calibrated to context, impact, and uncertainty rather than applied as a one-size-fits-all checklist (Mittelstadt, 2019; Morley et al., 2020).

What has surprised me is how closely this aligns with existing, committee-led governance patterns that I already work with at AUT, and how those patterns become even more important when AI enters the picture. Like many organisations described in the literature, AUT relies on structured forums such as change advisory boards and technical design groups to surface dependencies, challenge assumptions, and create shared accountability for decisions that might otherwise be made in silos, consistent with findings from

Morley et al. (2020) and Raji et al. (2020). In my experience, these forums are where “governance” becomes real: they are spaces where colleagues can ask the uncomfortable questions, about data provenance, privacy, bias risk, user impact, security posture, monitoring responsibilities, and what we will do if something goes wrong. This dissertation has made me view those meetings differently. I used to treat them primarily as delivery safeguards, ways to protect service stability and ensure technical due diligence. I now see them as ethical and social safeguards as well: mechanisms for contestability, collective judgement, and legitimate critique, especially when a new AI capability is proposed and the downstream impacts are harder to foresee (O'Neill, 2018; Stahl, 2021). Put plainly, committee structures only matter if they have the authority and information to influence outcomes, and if the culture makes it safe for people to slow things down when the risks are not well understood (Mittelstadt, 2019). That cultural dimension, psychological safety, permission to question, and willingness to document and revisit decisions, has become, for me, one of the most practical and important insights of this research.

This dissertation has affirmed that responsible AI is a collaborative effort that transcends individual roles or teams. AUT, like other institutions, operates within broader regulatory and normative ecosystems, and credible governance depends on both internal controls and external accountability, through law, standards, and engagement with affected communities (European Commission, 2019; Stahl et al., 2021). This matters because trust is not granted by intention; it is earned through demonstrable assurance and the willingness to be held accountable (O'Neill, 2018). I finish this dissertation more convinced that mature AI governance is a journey rather than an endpoint: perfection may be unattainable, but continuous improvement is both necessary and feasible when organisations embed ethics into workflows, monitor outcomes, learn from failures, and treat evidence, not rhetoric, as the basis for legitimacy (Mittelstadt, 2019; O'Neill, 2018; Stahl, 2021). Personally, that is the most enduring change in my thinking: ethical AI does not occur by default; it occurs by design, and I now see that design as part of my responsibility as a technology leader at AUT, not an optional extra.

References

Albalawee, N., & Fahoum, A. A. (2024). A novel legal analysis of Jordanian corporate governance legislation in the age of artificial intelligence. *Cogent Business & Management*, *11*(1), 2297465.

Alston, P. (2019). *Report of the Special Rapporteur on extreme poverty and human rights (The Digital Welfare State)*. https://undocs.org/A/74/493

AlTawil, T. N., & Rahhal, A. (2025). Examining synergies between UAE corporate social responsibility laws and corporate governance frameworks. *Journal of Money Laundering Control*, *28*(2), 369-384.

Alzebda, S., & Matar, M. A. (2025). Factors affecting citizen intention toward AI acceptance and adoption: the moderating role of government regulations. *Competitiveness Review: An International Business Journal*, *35*(2), 434-455.

Attard-Frost, B., De los Ríos, A., & Walters, D. R. (2023). The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics*, *3*(2), 389-406. https://doi.org/10.2139/ssrn.4034804

Auckland University of Technology. (2025a). *AI Hub*. Retrieved 8th August from https://autuni.sharepoint.com/sites/Tuia/SitePages/AI-hub.aspx

Auckland University of Technology. (2025b). *AI tools guidance*. Retrieved 8th August from https://autuni.sharepoint.com/sites/Tuia/SitePages/AI-tools-guidance.aspx

Auckland University of Technology. (2025c). *Our AI future: Vice-Chancellor's AI Taskforce report*. https://www.aut.ac.nz/__data/assets/pdf_file/0003/1082091/AI-Taskforce-Report-Our-AI-Future-FINAL-1.pdf

Ballantyne, A., Style, R., Stubbe, M., Murton, S., & Dowell, T. (2025). Using AI scribes in New Zealand primary care consultations: an exploratory survey. *Journal of Primary Health Care*. https://doi.org/10.1071/HC25079

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: a systematic literature review. *AI and Ethics*, 1-15.

Bensa, G., & Fattore, G. (2024). Artificial Intelligence and the Administrative Side of Public Healthcare. *Economia & Management*(3).

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, *33*(7), 133-167.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Brougham, D., & Haar, J. (2017). Employee assessment of their technological redundancy. *Labour & Industry: a journal of the social and economic relations of work*, *27*(3), 213-231. https://doi.org/10.1080/10301763.2017.1369718

Brougham, D., & Haar, J. (2018). Smart Technology, Artificial Intelligence, Robotics, and Algorithms (STARA): Employees' perceptions of our future workplace. *Journal of Management & Organization*, *24*(2), 239-257. https://doi.org/10.1017/jmo.2016.55

Brougham, D., & Haar, J. (2020). Technological disruption and employment: The influence on job insecurity and turnover intentions: A multi-country study. *Technological Forecasting and Social Change*, *161*, 120276. https://doi.org/10.1016/j.techfore.2020.120276

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*),

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society*, *3*(1), 1-12.

Butcher, J., & Beridze, I. (2019). What is the State of Artificial Intelligence Governance Globally? *The RUSI Journal*, *164*(5-6), 88-96. https://doi.org/10.1080/03071847.2019.1694260

Capgemini Research Institute. (2020). *AI and the Ethical Conundrum. How organizations can build ethically robust AI systems and gain trust*. Capgemini Research Institute. Retrieved 9th November 2024 from https://www.capgemini.com/wp-content/uploads/2020/10/AI-and-the-Ethical-Conundrum-Report.pdf

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180080.

Critical Appraisal Skills Programme. (n.d., 2025/12/18). *CASP Checklists*. Retrieved 19 September from https://casp-uk.net/casp-tools-checklists/

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Future of Humanity Institute, University of Oxford. Retrieved 9th September from https://cdn.governance.ai/GovAI-Research-Agenda.pdf

References

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan & A. Bryman (Eds.), *The SAGE handbook of organizational research methods* (pp. 671-689). SAGE.

Diaz-Asper, C., Hauglid, M. K., Chandler, C., Cohen, A. S., Foltz, P. W., & Elvevåg, B. (2024). A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications, and solutions. *American Psychologist*, *79*(1), 79.

Emah, I., & Bennett, S. (2025). Algorithmic emergence? Epistemic in/justice in AI-directed transformations of healthcare. *Frontiers in Sociology*, *10*, 1520810.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

European Commission. (2019). *European Commission: High-Level Expert Group on Artificial Intelligence publishes Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. Retrieved 9th September from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *L*(2024/1689). http://data.europa.eu/eli/reg/2024/1689/oj

Fan, D., Breslin, D., Callahan, J. L., & Iszatt-White, M. (2022). Advancing literature review methodology through rigour, generativity, scope and transparency. *International Journal of Management Reviews*, *24*(2), 171-180. https://doi.org/10.1111/ijmr.12291

Farooq, M. S., Tahseen, R., & Omer, U. (2021). Ethical Guidelines for Artificial Intelligence: A Systematic Literature Review. *VFAST Transactions on Software Engineering*, *9*(3), 33-47.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. https://ssrn.com/abstract=3518482

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

References

Gale, N. K., Heath, G., Cameron, E., Rashid, S., & Redwood, S. (2013). Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology*, *13*, 117. https://doi.org/10.1186/1471-2288-13-117

Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing*, *21*(6), 58-62. https://doi.org/10.1109/MIC.2017.4180835

Hickok, M. (2023). *Written Comments Regarding Proposed Rules on NYC Local Law 144 of 2021 in relation to Automated Employment Decision Tools*. https://rules.cityofnewyork.us/wp-content/uploads/2022/12/DCWP_NYC-Public-Comment_Merve-Hickok_Jan2023.pdf

Hyiamang, O., & Liu, X. M. (2025). Artificial Intelligence (AI) strategies for organizational innovation, growth, and productivity: a multi-case study approach. *Issues in Information Systems,*, *26*(1), 20-36. https://doi.org/10.48009/1_iis_103

International Institute for Management Development. (2023). *How organizations navigate AI ethics - I by IMD*. IMD. Retrieved 9th November 2024 from https://www.imd.org/ibyimd/technology/how-organizations-navigate-ai-ethics/

International Organization for Standardization. (2022). ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations. In. Geneva: International Organization for Standardization.

Ismail, M. A. B., & Goh, M. L. (2024). Evaluating the Impact of Artificial Intelligence on Work Ethics within Malaysian Regulatory Bodies. *Pakistan Journal of Life and Social Sciences (PJLSS)*, *22*(2).

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature machine intelligence*, *1*, 389-399. https://doi.org/10.1038/s42256-019-0088-2

Khadka, K., & Ullah, A. B. (2025). Human factors in cybersecurity: an interdisciplinary review and framework proposal. *International Journal of Information Security*, *24*(3), 1-13.

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, M. A. (2022). Ethics of AI: A Systematic Literature Review of Principles and Challenges. Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering (EASE),

References

Kop, M. (2021). EU Artificial Intelligence Act: The European Approach to AI. *Transatlantic Antitrust and IPR Developments*, *2*(1), 1-11. https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/

Lacmanovic, S., & Skare, M. (2025). Artificial intelligence bias auditing–current approaches, challenges and lessons from practice. *Review of Accounting and Finance*(ahead-of-print).

Loufek, B., Vidal, D., McClintock, D. S., Lifson, M., Williamson, E., Overgaard, S., McNaughton, K., Lipford, M. C., & Pardi, D. S. (2024). Embedding Internal Accountability Into Health Care Institutions for Safe, Effective, and Ethical Implementation of Artificial Intelligence Into Medical Practice: A Mayo Clinic Case Study. *Mayo Clinic Proceedings: Digital Health*, *2*(4), 574-583.

Luo, X., Li, Y., Xu, J., Zheng, Z., Ying, F., & Huang, G. (2025). AI in Medical Questionnaires: Innovations, Diagnosis, and Implications. *Journal of Medical Internet Research*, *27*, e72398.

Mac, T. A. (2024). Bias and discrimination in ML-based systems of administrative decision-making and support. *Computer Law & Security Review*, *55*, 106070.

Manage My Health. (2026, 2026/01/06

2026/01/18). *MMH cyber breach update 6 January 2026*. Manage My Health. https://managemyhealth.co.nz/mmh-cyber-breach-update-6-january-2026/

Manage My Health Ltd v Unknown Defendants [2026] NZHC 2, (Courts of New Zealand 2026). https://www.courtsofnz.govt.nz/assets/cases/2026/2026-NZHC-2.pdf

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

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L.,…Oak, S. (2025). *Artificial Intelligence Index Report 2025*. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, *6*, 175-183. https://doi.org/10.1007/s10676-004-3422-1

MBIE. (2025). *Responsible AI guidance for businesses: Investing with confidence: Accelerating private sector AI adoption and innovation* (978-1-99-106955-9

978-1-99-106999-3). https://www.mbie.govt.nz/assets/responsible-ai-guidance-for-businesses.pdf

Medical Council of New Zealand. (2025). *Draft for consultation: Using artificial intelligence (AI) in patient care*. https://www.mcnz.org.nz/assets/Consultations/AI-Consultation/Draft-Statement-on-Using-AI-in-Patient-Care-consultation-draft.pdf

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA.

Mittelstadt, B. D. (2019). Principles Alone Cannot Guarantee Ethical AI. *Nature machine intelligence*, *1*, 501-507. https://doi.org/10.1038/s42256-019-0114-4

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics*, *27*, 44. https://doi.org/10.1007/s11948-021-00319-4

Monyela, M., & Tella, A. (2024). Leveraging artificial intelligence for sustainable knowledge organisation in academic libraries. *South African Journal of Libraries and Information Science*, *90*(2), 1-11.

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.

Mujtaba, B. G. (2025). Human-AI Intersection: Understanding the Ethical Challenges, Opportunities, and Governance Protocols for a Changing Data-Driven Digital World. *Business Ethics and Leadership*, *9*(1), 109-126.

Munn, L. (2023). The Uselessness of AI Ethics. *AI and Ethics*, *3*, 869-877. https://doi.org/10.1007/s43681-022-00209-w

Nalbandian, L. (2022). An eye for an 'I:'a critical assessment of artificial intelligence tools in migration and asylum management. *Comparative Migration Studies*, *10*(1), 32.

National Cyber Security Centre. (2026, 2026/01/05

2026/01/18). *ManageMyHealth (MMH) cyber security breach involving patient information*. National Cyber Security Centre. https://www.ncsc.govt.nz/news/managemyhealth-mmh-cyber-security-breach-involving-patient-information/

References

New York City Department of Consumer and Worker Protection. (2021). *Automated Employment Decision Tools (AEDT) - Local Law 144 of 2021*. Retrieved 9th September from https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page

O'Neill, O. (2018). Linking Trust to Trustworthiness. *International Journal of Philosophical Studies*, *26*(2), 293-300. https://doi.org/10.1080/09672559.2018.1454637

OECD. (2019a, 2025/12/18). *OECD AI Principles overview*. https://oecd.ai/en/ai-principles

OECD. (2019b). *Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449)*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Office of the Privacy Commissioner. (2020). *Health Information Privacy Code 2020*. Retrieved from https://www.privacy.org.nz/assets/New-order/Privacy-Act-2020/Codes-of-practice/Health-information-privacy-code-2020/Health-Information-Privacy-Code-2020-website-version.pdf

Office of the Privacy Commissioner. (2024). *Working with third-party providers: understanding your privacy responsibilities*. https://www.privacy.org.nz/assets/New-order/Resources-/Publications/Guidance-resources/2024-11-21-s11-third-party-providers.pdf

Office of the Privacy Commissioner. (2025a). *Biometric processing privacy code 2025* [Code of practice]. https://www.privacy.org.nz/assets/Codes-of-Practice-2020/Biometrics/060825-Biometric-Processing-Privacy-Code-2025-A1102662.pdf

Office of the Privacy Commissioner. (2025b). *Inquiry into Foodstuffs North Island trial use of facial recognition technology*. https://www.privacy.org.nz/assets/DOCUMENTS/20250603-FRT-Inquiry-Report-A1082856.pdf

Office of the Privacy Commissioner. (2026a). *Advisory notice from Privacy Commissioner to all primary care providers affected by Manage My Health data breach: Notifiable privacy breach reporting requirements*. https://www.privacy.org.nz/assets/DOCUMENTS/20260901-Notice-for-Primary-Care-Providers-re-MMH-data-breach-reporting-requirements-A1151130.pdf

Office of the Privacy Commissioner. (2026b, 2026/01/12

2026/01/18). *Information for people impacted by the Manage My Health data breach*. Office of the Privacy Commissioner. https://www.privacy.org.nz/tuhono-connect/statements-media-releases/information-for-people-impacted-by-the-manage-my-health-data-breach/

Office of the Privacy Commissioner. (2026c, 2026/01/09

References

2026/01/18). *MMH breach: Information for affected primary care providers*. Office of the Privacy Commissioner. https://www.privacy.org.nz/tuhono-connect/statements-media-releases/information-for-all-primary-care-providers-affected-by-manage-my-health-data-breach/

Office of the Privacy Commissioner. (2026d, 2026/01/09

2026/01/18). *Updated statement on Manage My Health cyber incident*. Office of the Privacy Commissioner. https://www.privacy.org.nz/tuhono-connect/statements-media-releases/statement-on-manage-my-health-cyber-incident/

Okun, S., Hanger, M., Browne-James, L., Montgomery, T., Rafaloff, G., & van Delden, J. J. (2023). Commitments for ethically responsible sourcing, use, and reuse of patient data in the digital age: cocreation process. *Journal of Medical Internet Research*, *25*, e41095.

Olawade, D. B., Weerasinghe, K., Mathugamage, M. D. D. E., Odetayo, A., Aderinto, N., Teke, J., & Boussios, S. (2025). Enhancing ophthalmic diagnosis and treatment with artificial intelligence. *Medicina*, *61*(3), 433.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S.,…Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Palladino, N. (2023). A ‘biased’emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices. *Telecommunications Policy*, *47*(5), 102479.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Paul, L., & Rena, R. (2025). Generative AI in South African Healthcare: Navigating Challenges and Harnessing Opportunities for Industry and Research. *Management Dynamics*, *25*(1), 6.

Qureshi, N. I., Choudhuri, S. S., Nagamani, Y., Varma, R. A., & Shah, R. (2024). Ethical Considerations of AI in Financial Services: Privacy, Bias, and Algorithmic Transparency. 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chikkaballapur, India.

References

Radio New Zealand. (2024, 2024/02/08). Privacy commissioner keeps close eye on supermarkets' facial recognition trial. *RNZ*. https://www.rnz.co.nz/news/national/508613/privacy-commissioner-keeps-close-eye-on-supermarkets-facial-recognition-trial

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. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*),

Rana, N. P., Pillai, R., Sivathanu, B., & Malik, N. (2024). Assessing the nexus of Generative AI adoption, ethical considerations and organizational performance. *Technovation*, *135*, 103064.

Renda, A., Arroyo, J., Fanni, R., Laurer, M., Maragkakis, M., Nadalin, G., & Pisharody, N. (2021). *Study to support the Parliament's assessment of the proposal for a Regulation on Artificial Intelligence*. https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2021)695482

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review*, *94*(1), 15-55.

Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the financial sector: The line between innovation, regulation and ethical responsibility. *Information*, *15*(8), 432.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation. *Ai & Society*, *36*(1), 59-77. https://doi.org/10.1007/s00146-020-00992-2

Rugiubei, R., & Stoica, V. (2025). Challenges in Adopting Artificial Intelligence Technologies in Supply Chain Management in Romanian Companies. *Revista De Management Comparat International*, *26*(1), 207-217.

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61-86.

Sandeep, M., Lavanya, V., & Balakrishnan, J. (2025). Leveraging AI in recruitment: enhancing intellectual capital through resource-based view and dynamic capability framework. *Journal of Intellectual Capital*, *26*(2), 404-425.

References

Sargiotis, D. (2024). Ethical AI in Information Technology: Navigating Bias, Privacy, Transparency, and Accountability. *Adv Mach Lear Art Inte*, *5*(3), 01-14.

Saw, S. N., & Ng, K. H. (2022). Current challenges of implementing artificial intelligence in medical imaging. *Physica Medica*, *100*, 12-17.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). *Principles to Practices for Responsible AI: Closing the Gap*. https://doi.org/10.48550/arXiv.2006.04707

Schuett, J. (2023). Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, *14*(4), 703-721. https://doi.org/10.1017/err.2023.1

Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. (2025). *The state of AI: How organizations are rewiring to capture value*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value

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

Stahl, B. C. (2021). *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Springer.

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Shaelou, S. L., Patel, A., Ryan, M., & Wright, D. (2021). Artificial intelligence for human flourishing–Beyond principles for machine learning. *Journal of Business Research*, *124*, 374-388. https://doi.org/10.1016/j.jbusres.2020.11.030

Statistics New Zealand. (2020). *Algorithm charter for Aotearoa New Zealand*. https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, 45. https://doi.org/10.1186/1471-2288-8-45

Tian, K., Zhu, Z., Mbachu, J., Moorhead, M., & Ghanbaripour, A. (2025). Artificial intelligence in construction risk management: a decade of developments, challenges, and integration pathways. *Journal of Risk Research*, 1-33.

Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory*, *30*(3), 167-186. https://doi.org/10.1177/0735275112457914

References

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207-222. https://doi.org/10.1111/1467-8551.00375

U.S. Department of Commerce, N. I. o. S. a. T. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.* NIST Trustworthy and Responsible AI, National Institute of Standards and Technology,. Retrieved 9th September from https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Ubaldi, B.-C., & Zapata, R. (2024). *Governing with artificial intelligence: Are governments ready?* [No. 20](OECD artificial intelligence papers, Issue. https://doi.org/10.1787/26324bc2-en

Veale, M., & Zuiderveen Borgesius, F. J. (2021). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, *22*(4), 97-112. https://ssrn.com/abstract=3896852

Wamba, S. F., Queiroz, M. M., Randhawa, K., & Gupta, G. (2025). Generative artificial intelligence and the challenges to adding value ethically. In (Vol. 144, pp. 103235): Elsevier.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. Honolulu, HI, USA.

Wright, L., Muenster, R. M., Vecchione, B., Qu, T., Cai, S., Metcalf, J., & Matias, J. N. (2024). Null compliance: NYC Local Law 144 and the challenges of algorithm accountability. *arXiv*. https://doi.org/10.48550/arXiv.2406.01399

Yang, J., Chu, S.-C., & Cao, Y. (2024). Adopting AI Advertising Creative Technology in China: A Mixed Method Study Through the Technology-Organization-Environment (TOE) Framework, Perceived Value and Ethical Concerns. *Journal of Current Issues & Research in Advertising*, 1-24.

Yue, C. A., Men, L. R., Mitson, R., Davis, D. Z., & Zhou, A. (2024). Artificial intelligence for internal communication: Strategies, challenges, and implications. *Public Relations Review*, *50*(5), 102515.

Zada, M., Khan, S., Mehmood, S., & Contreras-Barraza, N. (2024). Generative artificial intelligence in FinTech: Applications, environmental, social, and governance considerations, and organizational performance: The moderating role of ethical dilemmas. *Oeconomia Copernicana*, *15*(4).

Zhao, I. Y., Ma, Y. X., Yu, M. W. C., Liu, J., Dong, W. N., Pang, Q., Lu, X. Q., Molassiotis, A., Holroyd, E., & Wong, C. W. W. (2021). Ethics, integrity, and retributions of digital detection surveillance

systems for infectious diseases: systematic literature review. *Journal of Medical Internet Research*, *23*(10), e32328.

Zhou, D., Keogh, J. W., Ma, Y., Tong, R. K., Khan, A. R., & Jennings, N. R. (2025). Artificial intelligence in sport: A narrative review of applications, challenges and future trends. *Journal of Sports Sciences*, 1-16.

Zhou, T., Huang, Y., Avanasi, R., Brain, R. A., Prosperi, M., & Bian, J. (2025). Content hubs, information flows, and reactions for pesticide-related discussions on Twitter/X. *Integrated Environmental Assessment and Management*, *21*(3), 628-638.

Appendices

Appendix A: Databases searched by AUT-licensed EBSCO collections

Table 7. Summary of Databases Searched via EBSCOhost by Discipline

Category	Databases Included
Science, Technology & Engineering	• GreenFILE
	• Library, Information Science & Technology Abstracts
	• MathSciNet via EBSCOhost
	• Ergonomics Abstracts
	• Wiley e-textbook Subscription (Physical Sciences & Engineering)
Business, News & Hospitality	• Business Source Complete
	• Regional Business News
	• Newswires
	• Hospitality & Tourism Complete
	• Wiley e-textbook Subscription (Business)
Health, Medicine & Sports	• CINAHL Complete
	• CINAHL Ultimate
	• MEDLINE
	• Dentistry & Oral Sciences Source
	• SPORTDiscus with Full Text
	• Wiley e-textbook Subscription (Health Sciences)
Education & Social Sciences	• ERIC
	• Teacher Reference Center
	• SocINDEX with Full Text
	• Wiley e-textbook Subscription (Education)
	• Wiley e-textbook Subscription (Psychology)
	• Wiley e-textbook Subscription (Social Sciences & Humanities)
Arts & Humanities	• Art & Architecture Complete
	• Art Full Text (H.W. Wilson)
	• Art Index Retrospective (H.W. Wilson)
	• Humanities International Index
	• Communication & Mass Media Complete
Multidisciplinary & General Reference	• eBook Collection (EBSCOhost)
	• OpenDissertations
	• Australia/New Zealand Reference Centre Plus
	• Australia / New Zealand Reference Centre Plus eBook Subscription

Appendix B: Included Studies Index

This appendix lists the included studies using the study identifier employed in the coding matrix and the corresponding short-form citation used throughout the chapter.

Table 8. Included studies (N = 38)

Study ID	Short citation	Title
S01	Attard-Frost et al. (2023)	The ethics of AI business practices: a review of 47 AI ethics guidelines
S02	Munn (2023)	The uselessness of AI ethics
S03	Khan et al. (2022)	Ethics of AI: A Systematic Literature Review of Principles and Challenges
S04	Mäntymäki et al. (2022)	Defining organizational AI governance
S05	Batool et al. (2025)	AI governance: a systematic literature review
S06	T. Zhou et al. (2025)	Content hubs, information flows, and reactions for pesticide-related discussions on Twitter/X
S07	Nalbandian (2022)	An eye for an 'I:' a critical assessment of artificial intelligence tools in migration and asylum management
S08	Khadka and Ullah (2025)	Human factors in cybersecurity: an interdisciplinary review and framework proposal
S09	Monyela and Tella (2024)	Leveraging artificial intelligence for sustainable knowledge organisation in academic libraries
S10	Olawade et al. (2025)	Enhancing Ophthalmic Diagnosis and Treatment with Artificial Intelligence
S11	Luo et al. (2025)	AI in Medical Questionnaires: Innovations, Diagnosis, and Implications
S12	Okun et al. (2023)	Commitments for Ethically Responsible Sourcing, Use, and Reuse of Patient Data in the Digital Age: Cocreation Process
S13	Zhao et al. (2021)	Ethics, Integrity, and Retributions of Digital Detection Surveillance Systems for Infectious Diseases: Systematic Literature Review
S14	Ridzuan et al. (2024)	AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility
S15	Paul and Rena (2025)	Generative AI in South African Healthcare: Navigating Challenges and Harnessing Opportunities for Industry and Research

S16	Emah and Bennett (2025)	Algorithmic emergence? Epistemic in/justice in AI-directed transformations of healthcare
S17	Bensa and Fattore (2024)	Artificial Intelligence and the Administrative Side of Public Healthcare
S18	Mujtaba (2025)	Human-AI Intersection: Understanding the Ethical Challenges, Opportunities, and Governance Protocols for a Changing Data-Driven Digital World
S19	D. Zhou et al. (2025)	Artificial intelligence in sport: A narrative review of applications, challenges and fsuture trends
S20	Tian et al. (2025)	Artificial intelligence in construction risk management: a decade of developments, challenges, and integration pathways
S21	Yang et al. (2024)	Adopting AI Advertising Creative Technology in China: A Mixed Method Study Through the Technology-Organization-Environment (TOE) Framework, Perceived Value and Ethical Concerns
S22	Albalawee and Fahoum (2024)	A novel legal analysis of Jordanian corporate governance legislation in the age of artificial intelligence
S23	Diaz-Asper et al. (2024)	A Framework for Language Technologies in Behavioral Research and Clinical Applications: Ethical Challenges, Implications, and Solutions
S24	Ismail and Goh (2024)	Evaluating the Impact of Artificial Intelligence on Work Ethics within Malaysian Regulatory Bodies
S25	Rugiubei and Stoica (2025)	Challenges in Adopting Artificial Intelligence Technologies in Supply Chain Management in Romanian Companies
S26	Lacmanovic and Skare (2025)	Artificial intelligence bias auditing , current approaches, challenges and lessons from practice
S27	AlTawil and Rahhal (2025)	Examining synergies between UAE corporate social responsibility laws and corporate governance frameworks
S28	Sandeep et al. (2025)	Leveraging AI in recruitment: enhancing intellectual capital through resource-based view and dynamic capability framework
S29	Birkstedt et al. (2023)	Al governance: themes, knowledge gaps and future agendas
S30	Alzebda and Matar (2025)	Factors affecting citizen intention towards AI acceptance and adoption: the moderating role of government regulations

S31	Zada et al. (2024)	Generative artificial intelligence in FinTech: Applications, environmental, social, and governance considerations, and organizational performance: The moderating role of ethical dilemmas
S32	Saw and Ng (2022)	Current challenges of implementing artificial intelligence in medical imaging
S33	Yue et al. (2024)	Artificial intelligence for internal communication: Strategies, challenges, and implications
S34	Palladino (2023)	A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices
S35	Loufek et al. (2024)	Embedding Internal Accountability Into Health Care Institutions for Safe, Effective, and Ethical Implementation of Artificial Intelligence Into Medical Practice: A Mayo Clinic Case Study
S36	Mac (2024)	Bias and discrimination in ML-based systems of administrative decision-making and support
S37	Wamba et al. (2025)	Generative artificial intelligence and the challenges to adding value ethically
S38	Rana et al. (2024)	Assessing the nexus of Generative AI adoption, ethical considerations and organizational performance

Appendix C: Extracted Data

Table 9. Comparative Analysis of AI Governance Frameworks and Implementation Strategies

Article	Implementation Details (RQ1)	Governance Focus (RQ2)	Implementation Outcome/Effectiveness	Challenges Reported	Framework Components
Attard-Frost et al. (2023)	Conceptualises "AI business practices" as iterative political/economic behaviours spanning resourcing, design, development, deployment, and use, then uses this lens to review and code ethics guidelines. (S01, p. 2) Operationally foregrounds *internal* oversight (e.g., social impact assessments/audits, internal review boards/ethics bodies, worker involvement) and *external* oversight (policy/legal/regulatory frameworks) as governance mechanisms.	Argues current AI ethics guidance over-weights "algorithmic decision-making" and under-addresses the business/political economy in which AI systems sit, and reframes governance using the FAST principles (fairness, accountability, sustainability, transparency) tailored to business practice.	Finds the political and economic implications of AI business practices are "greatly underrepresented" in the guideline included studies reviewed, implying limited real-world governance coverage if organisations rely on typical guidelines alone.	Underrepresentation of business practice and political economy considerations in prevailing guidelines (a structural gap). Practical governance burden implied by the need for both internal and external oversight arrangements (audits, review bodies, regulatory scaffolding).	Concrete topic-level components coded across FAST include (examples): discrimination/data risk assessments, independent tests/audits/investigations, external oversight roles/institutions, post-harm accountability, right to challenge/appeal and redress, continuous monitoring/improvement, bias disclosure/mitigation, representative data, open data, and human-rights framing.

Munn (2023)	Explains why "principles" fail in practice (difficult to operationalise, poorly embedded in industry) and illustrates weak implementation via an external advisory council model with no veto power. Advances alternative implementation directions: broaden to AI justice (sociopolitical systems of oppression) and narrow to concrete controls (accuracy, auditing, governance).	Critiques principle-centric "AI ethics" as a governance substitute; argues effective responsible AI requires enforceable governance that is connected to real organisational power and constraints (e.g., auditability, governance mechanisms, and resistance to "business as usual").	Concludes the dominant turn to ethical principles is "largely useless" because it fails to mitigate harms and produces a persistent principle to practice gap.	"Meaningless, isolated, and toothless" principles: contested/incoherent, situated in ecosystems that ignore ethics, and lacking penalties (enabling corporate agendas). "Ethics washing" as a strategy to appear ethical while avoiding regulation.	Diagnostic framework of failure modes (meaningless / isolated / toothless) plus two alternative governance pathways: (1) justice-oriented governance; (2) operational controls such as accuracy, auditing, and governance.
Khan et al. (2022)	Synthesises how ethics is (and is not) implemented by mapping principles that designers are urged to consider, and by identifying barriers that hinder implementation; positions findings as inputs to a maturity model to guide capability improvement.	Principles-based governance focus highlights four most common principles, transparency, privacy, accountability, and fairness as dominant across the reviewed literature.	Offers an evidence-mapped inventory (22 principles; 15 challenging factors) rather than evaluating effectiveness of any single governance framework; proposes using these as inputs for a future maturity model to assess and improve ethical capabilities.	Most frequently cited barriers include lack of ethical knowledge and vague principles. Specific implementation blockers include lack of audit/monitoring, absence of legal frameworks, and business interest pressures (among others listed).	Catalogue of 22 principles + 15 challenging factors; challenge themes grouped into (i) knowledge and expertise, (ii) organisational management, and (iii) tools and technologies; preliminary maturity-model structure.

Mäntymäki et al. (2022)	Defines organisational AI governance as a *system* of rules, practices, processes, and technological tools used to ensure AI use aligns with organisational strategy/values, fulfils legal requirements, and meets ethical AI principles; explicitly intended to be action-oriented for implementation. (S04, p. 1,2) Extends implementation scope across the full AI system lifecycle (use case definition/design through maintenance and disposal).	Positions "AI governance" as the organisational mechanism through which AI ethics is translated into operational practice, connecting ethical, organisational, and technological aspects and anchoring them in legal and strategic alignment.	Conceptual contribution: provides definitional clarity to help practitioners identify constituent parts of the "complex problem" of translating ethics into practice; does not report empirical effectiveness outcomes.	Core implementation challenge: ethical principles (e.g., fairness) must be translated into practicable governance processes, yet the concept is complex and previously under-defined.	Core components are explicit in the definition: rules, practices, processes, technological tools; plus alignment targets (strategy/objectives/values, legal requirements, ethical principles) and lifecycle coverage.

Batool et al. (2025)	Uses an SLR organised around four implementation questions: who is accountable, what is governed, when governance occurs in the AI lifecycle, and how governance is implemented via frameworks/tools/policies/models. Classifies "how" solutions by governance level (team, organisational, industry, national, international) and reports most solutions sit at organisational level in their sample.	Practical governance mapping: frames responsible AI governance as selecting and aligning appropriate governance solutions across stakeholders, governed elements (notably data), lifecycle stages, and implementation instruments.	Intended effectiveness is decision-support addresses the difficulty stakeholders face in selecting suitable governance frameworks/tools/models by synthesising available solutions and their facets; does not claim universal effectiveness, but provides a structured basis for selection.	Stakeholders lack a clear picture of available governance solutions and how to choose among them. (S05, p. 1) Data governance challenges: security/privacy risks in data collection, and unstructured/non-standardised medical data affecting AI model quality. Notes there are "challenges and limitations" in existing governance solutions (treated as an explicit review target).	Four analytic elements (who/what/when/how); implementation instruments (frameworks, tools, policies, models); five governance levels (team → international); emphasis on governing data as a key object of governance.

T. Zhou et al. (2025)	Its "implementation" is methodological: it operationalises *content hubs*, maps information flows, and analyses reactions within pesticide-related discussions on X (Twitter). Any "governance" relevance is indirect (e.g., informing risk communication, moderation, or public health messaging strategies rather than governing AI systems).	Indirectly relevant to *information governance* (how influence, diffusion, and reactions form around risk topics), which could inform platform/community governance interventions.	Effectiveness is framed as explanatory/diagnostic: the study identifies how information clusters and reactions behave, rather than evaluating a governance framework's effectiveness.	Typical platform-data research constraints (platform dynamics, representativeness, and limits of inference) and the difficulty of translating descriptive findings into intervention design.	Content hub identification; network/flow analysis; reaction/engagement measures; discourse/sentiment-style interpretation of responses; comparative analysis across hub types.

Nalbandian (2022)	Describes real-world deployment of AI in migration/asylum management (e.g., NZ border/immigration efficiency tools; US biometric databases and matching; case management and data scraping). The paper frames implementation as involving explicit trade-offs between administrative efficiency and rights protections, implying that governance must be designed to prevent "rights-for-access" bargains.	A rights-based, critical governance stance: privacy and security as non-negotiables; transparency about tool use and impacts; attention to vulnerability and power asymmetries in migration contexts.	Not an evaluative governance trial; effectiveness is discussed as a tension: AI may improve administrative efficiency, but without robust safeguards it can degrade rights protections and legitimacy.	Risks of sacrificing human rights (privacy/security) in the name of efficiency; opacity and transparency deficits; heightened vulnerability for already-marginalised populations.	Implied governance components: human-rights safeguards, privacy/security protections, transparency requirements, and heightened scrutiny for high-stakes public-sector use cases (migration/asylum).
Khadka and Ullah (2025)	Proposes an *iterative* human-centric cybersecurity framework implemented via assessment → design → implementation → evaluation: behavioural audits; tailored/gamified training; socio-technical decision support; feedback loops; cross-sectional + longitudinal evaluation.	Governance embedded through *organisational culture and compliance*: policy, behaviour alignment, leadership development, insider-threat mitigation, and explicit AI ethics and privacy considerations for AI-enabled security tools.	Expected benefits (not yet empirically validated): improved literacy and compliance, reduced analyst fatigue, culturally inclusive practices, better human, system alignment, and more trustworthy/ethical AI security applications.	Human variability (stress, fatigue, culture), socio-technical misalignment, insider risk, ethical/privacy risks (bias, transparency), and limited real-world validation to date.	Four pillars: psychological dimensions, training & awareness, organisational culture & compliance, socio-technical integration (incl. AI ethics/privacy), plus inclusivity/cultural tailoring and evaluation/validation mechanisms.

Monyela and Tella (2024)	Recommends concrete library implementation steps: create/adhere to ethical standards; prioritise data protection (anonymisation, informed consent, lifecycle security); invest in staff education/training; collaborate with stakeholders; continuously monitor and evaluate AI applications.	"Ethical AI governance" in libraries emphasising user trust, fairness, equity, transparency, accountability, and mitigation of algorithmic bias and data protection risks.	Reports practice-level benefits (resource efficiency gains and improved user satisfaction) alongside the need for responsible governance to realise these benefits.	Ethical risks in adoption: data privacy, algorithmic transparency, and user empowerment; plus the need to manage bias, inclusivity, and ongoing alignment with user/community values.	Specific components include transparency/explainability, bias mitigation (audits/monitoring), inclusivity/accessibility, user education, partnerships, continuous assessment, community consultation, and an ethical AI governance structure with roles/duties and grievance channels.
Olawade et al. (2025)	Emphasises implementation controls for clinical AI: patient consent, anonymisation, and strong security (e.g., encryption) as part of data governance; bias mitigation through diverse/representative datasets; integration with existing health systems and regulatory compliance for deployment.	Responsible clinical governance anchored in privacy, fairness/bias control, safety, and compliance-oriented oversight (to ensure AI supports, rather than undermines, clinical decision-making).	Positions governance as enabling condition for safe clinical benefit (improved diagnosis/treatment potential), rather than reporting measured governance effectiveness outcomes.	Key barriers: privacy risks, bias in training data, and the operational burden of meeting regulatory and system-integration requirements (e.g., working with EHR environments).	Components include data governance protocols (consent/anonymisation/security), bias mitigation strategies, compliance mechanisms, and integration requirements with clinical information systems.

Luo et al. (2025)	Implementation emphasis is socio-technical: infrastructure upgrades and interoperability with EHRs (HL7/FHIR) to enable workflow integration; privacy/security compliance (e.g., GDPR/HIPAA); explicit consent procedures; clear roles and responsibilities plus liability allocation for AI-supported decisions.	Effective governance focuses on patient-data protection, accountability for clinical/administrative decision-making, and lawful/ethical operation of AI-enabled questionnaires across health information systems.	Discusses effectiveness in enabling responsible adoption (better integration → more usable tools; clearer governance → safer deployment), but does not present a formal evaluation of a governance framework.	Practical constraints: interoperability gaps, infrastructure costs, data privacy/security risks, consent management challenges, and medico-legal liability questions.	Components include interoperability standards (HL7/FHIR), governance role clarity, consent and privacy/security controls, and compliance frameworks (GDPR/HIPAA) with liability-aware oversight.
Okun et al. (2023)	A structured co-creation approach (9 months) using participatory methods (landscape analysis, listening sessions, survey), guided by biomedical ethics and "social licence", to produce a patient-first governance and accountability framework for data stewardship.	Governance foregrounds what is acceptable in data use/sharing (not merely technical analytics). It aims to strengthen legitimacy ("social licence") through privacy, transparency, openness, autonomy-respecting consent, and shared power between data stewards and communities.	Outcome is the "Commitments for the digital age" framework (six commitments) intended to guide ethically responsible governance between a data-collecting company and its community; effectiveness is framed as strengthening/maintaining social licence and setting enforceable expectations for conduct.	Core problem: organisations may lack social licence and "ethical maturity"; uncertainty about acceptable data uses; navigating privacy, autonomy, equity concerns in novel digital research environments.	Six framework components: continuous/shared learning; respect and empower individual choice; informed/understood consent; people-first governance; open communication and accountable conduct; inclusivity, diversity, and equity.

Zhao et al. (2021)	Reviews AI-augmented infectious disease surveillance and argues implementation must be underpinned by an ethical governance framework and cross-national/national governance structures (regulatory, medical, ethical, and legal), with mandated safeguards for cross-border data sharing.	Effective governance is framed as multi-level (individual, organisational, societal) and centred on trust, privacy/confidentiality, civil rights, and explicit governance structures (rather than leaving governance to lag behind technical capability).	Effectiveness is conditional: surveillance can be "highly effective and responsive", but acceptability and effectiveness depend on public trust in implementation and governance, plus transparent and independent oversight.	Public concerns include inadequate information, unclear governance frameworks, and lack of privacy protection, data integrity, and autonomy; risks of widening disparities/digital divides and exposing highly sensitive movement/contact data.	Components are operationalised as six domains/themes: awareness, digital integrity, trust, privacy/confidentiality, civil rights, governance; plus practical elements such as data ownership/control, data quality, bias, deidentification, and oversight standards.
Ridzuan et al. (2024)	Provides practical implementation guidance for financial services: risk management solutions (mapped in a table), governance recommendations, and an ethical/responsible approach drawing on ASEAN guidance; proposes a model governance framework including an AI governance committee for oversight and compliance.	Focuses on balancing innovation with regulation and ethical responsibility: responsible AI practices, compliance, and risk controls to address harms (e.g., bias, privacy/cyber risk) in AI-enabled financial decision-making.	Effectiveness is framed as risk mitigation and compliance enablement (a governance design contribution rather than empirical testing): adopting structured controls and oversight is positioned as the pathway to responsible AI use in financial services.	Governance complexity in highly regulated environments; managing AI risks and ensuring ongoing compliance; operationalising ethical guidelines into workable oversight and decision controls.	Components include risk management measures (Table 2), governance recommendations (Table 3), comprehensive ethical guidelines, and an AI governance committee embedded within a model governance framework.

Paul and Rena (2025)	Recommends practical governance actions for generative AI in SA healthcare: robust policy framework (data protection, "algorithmising" approaches, patient consent), public, private partnerships for infrastructure/R&D, capacity-building training, and patient/provider trust-building (information campaigns).	Emphasises regulatory specificity that balances innovation with ethical safeguards (privacy, consent, bias mitigation), backed by multi-stakeholder design (government, health sector, researchers, tech firms).	Conceptual/practice-oriented synthesis rather than an evaluated governance rollout; argues governance is a condition for realising benefits (diagnostic accuracy, cost reduction, accessibility) sustainably.	Data complexity and heterogeneity; privacy/security; infrastructure and skills shortages; policy/regulatory gaps; budget constraints; and risk of bias or misuse without safeguards.	Policy and regulatory architecture; PPP governance model; workforce AI literacy and training; trust/engagement mechanisms; localisation of models; cross-sector collaboration structures.
Emah and Bennett (2025)	Conceptual governance guidance: embed participatory engagement across system design and knowledge generation; adopt participatory accountability structures that recognise power differentials and prevent tokenistic "participation".	Frames "effective" governance as equitable health-AI governance grounded in data justice and algorithmic fairness, protecting autonomy, inclusivity, and meaningful patient/clinician agency.	No measured effectiveness (conceptual analysis), but provides governance implications intended to reduce epistemic/algorithmic injustice in healthcare AI assemblages.	Decontextualisation of qualitative clinical nuance; automation/acceleration of bias; weak transparency; risks of tokenistic engagement; structural power imbalances shaping evidence and system design.	Participatory governance mechanisms; accountability structures; data justice and fairness lens; continuous/critical review of evidence practices used for guidance and governance.

Bensa and Fattore (2024)	Advocates local experimentation (pilots) and public,private collaboration to implement administrative AI uses that can deliver efficiency while remaining lawful.	Focuses on navigating ethical, legal, organisational, and governance issues in administrative healthcare AI to ensure compliance and legitimacy.	Positions administrative AI as offering "relatively quick" wins in efficiency/effectiveness *if* governance issues are addressed; limited evaluation detail in the paper's abstract/metadata.	Ethical, legal, organisational, and governance constraints that must be resolved before scaling.	Local pilot design; public,private partnership arrangements; compliance-by-design emphasis; selection of administrative use cases.
Mujtaba (2025)	Proposes multi-level governance protocols (individual → departmental → organisational → societal): define roles/responsibilities; establish protocols and KPIs; invest in training; align AI use to values; integrate AI risk management with enterprise risk; require monitoring and periodic review; encourage multi-stakeholder collaboration and international coordination.	"Effective" governance foregrounds transparency, explainability, accountability, fairness, inclusivity, safety, and privacy, and stresses that accountability cannot be outsourced (even when procuring AI).	Conceptual framework (no implementation evaluation), aimed at reducing misuse/abuse, bias, and opacity through structured governance and culture change.	Opaque AI decisions; accountability gaps; surveillance/overreach; discrimination and bias; misuse/hallucinations; regulatory complexity and skills demand.	Governance principles (trustworthiness, privacy, fairness/bias detection, explainability, transparency, accountability); organisational AI strategy; protocols/guidelines; training; audit/monitoring; multi-stakeholder engagement.

D. Zhou et al. (2025)	Recommends governance for sport AI via robust data governance and security controls: clear access/retention rules; encryption/anonymisation; ethical review boards; athlete consent and education; privacy-preserving methods (e.g., federated learning, differential privacy); and explainable AI to improve transparency.	Effective governance is responsible and transparent AI for sensitive athlete data, using broader frameworks (e.g., OECD principles, EU trustworthy AI guidance) and legal standards (e.g., GDPR), plus sport-specific guidance (e.g., WADA).	Narrative review (not an evaluated governance deployment): argues these measures can protect rights, reduce misuse/discrimination , and build trust, while noting many methods are still early-stage in sport contexts.	Privacy and ethical risks (unauthorised access, misuse, discrimination); algorithmic bias; data governance complexity; real-time data quality/integration challenges; broader impacts (e.g., job displacement).	Governance frameworks (OECD, EU trustworthy AI, GDPR, WADA); governance rules for access/retention; ethical review boards; consent and education; technical privacy measures; XAI for transparency.
Tian et al. (2025)	Proposes an AI risk management lifecycle and a functional taxonomy mapping AI method to construction risk tasks/phases; recommends stakeholder co-governance/participatory forums, human-in-the-loop integration of domain expertise, and design choices that maintain interpretability and accountability in safety-critical contexts.	Frames governance as aligning AI-enabled risk management with broader risk epistemologies; emphasises responsible, resilient, explainable deployment in high-stakes environments and calls for ethical/legal framing (including liability and accountability implications).	Primarily a systematic synthesis: it consolidates evidence of AI's potential value in risk work, but the governance approach is mainly proposed rather than empirically evaluated as a standalone "framework effectiveness" test.	Fragmented integration across methods and phases; interpretability,performance tensions; legal/accountability and stakeholder trust concerns; data and organisational constraints limiting real-world uptake.	Functional taxonomy of AI methods ↔ risk phases; integrated lifecycle model; governance levers: explainability/interpretability, stakeholder co-governance, ethical/legal framing, and human-in-the-loop expert override.

Yang et al. (2024)	Focuses on *adoption and organisational adaptation* of AI advertising creative technology (China) using a mixed-method design. Governance is discussed mainly as enabling conditions: internal policies, oversight practices, and risk controls that shape adoption (e.g., managing ethical concerns, data handling, and responsible use expectations in creative work).	Not a full responsible-AI governance framework; rather, identifies what an "effective" enabling environment looks like for responsible adoption (controls that reduce ethical concerns and build perceived value/trust, alongside organisational support).	Effectiveness is expressed as adoption outcomes (e.g., factors associated with uptake and intention), rather than measured post-deployment impacts such as auditability, harm reduction, or accountability performance.	Ethical concerns (e.g., misuse, fairness, transparency), risk perceptions, governance/standards uncertainty, and organisational capability gaps that constrain adoption or shape how it is used.	Adoption determinants (e.g., perceived value and enabling conditions); risk/ethical concern constructs; organisational support and capability; implied controls such as policies/guardrails, role clarity, and oversight routines for responsible use.

Albalawee and Fahoum (2024)	Uses a descriptive,analytical legal approach and proposes a legal governance model for Jordanian enterprises, advocating clearer statutory treatment of AI within corporate governance; focuses implementation on board/senior management oversight, and strengthening transparency/disclosure practices supported by AI-enabled analytics.	Corporate governance legislation "in the age of AI": the emphasis is on aligning legal requirements, board duties, auditability, transparency, and disclosure with AI-enabled decision-making, and using AI to reduce unethical financial/managerial practices.	Effectiveness is argued largely in normative/legal terms (improved transparency, reduced errors, better disclosure), rather than tested through implementation evaluation in organisations.	Gaps/misalignment between existing legal frameworks and AI-driven practices; definitional ambiguity around AI in law; governance risks tied to opacity, accountability, and compliance capacity.	Proposed staged legal governance model; clearer legal definition of AI; board/senior leadership accountability; disclosure/transparency obligations; incentives/support for digital governance capability.
Diaz-Asper et al. (2024)	Provides a practical framework for deploying language technologies in behavioural/clinical settings: recommends early stakeholder involvement, explicit disclosure when users interact with AI, detailed documentation of model behaviour and data, and formal ethics oversight (e.g., IRB-style review) tailored to application context.	A socio-technical governance approach for language technologies that prioritises privacy/confidentiality, bias mitigation, transparency, and patient/participant protections, with procedural safeguards rather than relying on general ethics codes alone.	Effectiveness is positioned as risk reduction (e.g., reducing harm from misuse/misinterpretation) via process controls; it is not presented as a measured organisational outcome study.	Automation bias and misuse risk; limited enforceability of broad ethical guidance; structural inequities/discrimination risks; privacy and confidentiality pressures in clinical contexts.	Governance checklist across lifecycle (design → deployment); IRB/ethics review; stakeholder engagement; transparency/disclosure; user guidance; documentation of inferences/limits; privacy/confidentiality controls; human oversight.

Ismail and Goh (2024)	Discusses AI adoption in Malaysian regulatory bodies through an ethics lens and promotes strengthening codes of ethics, staff capability-building, and governance practices that preserve professional integrity (e.g., monitoring, accountability, transparency).	Focuses on how AI intersects with work ethics in public/regulatory contexts; governance emphasis is on maintaining ethical conduct, accountability, transparency, and appropriate use of AI in decision support.	Effectiveness is framed as supporting ethical behaviour and integrity while enabling AI-enabled performance gains; the paper largely provides evaluative discussion and recommended governance responses rather than a full implementation trial.	Ethical tensions in AI-assisted work; concerns about accountability, transparency, and professional judgement; organisational readiness and workforce impacts in regulated environments.	Ethics-forward governance measures: code of ethics for AI use, accountability mechanisms, transparency expectations, training/capability development, and oversight/monitoring practices.
Rugiubei and Stoica (2025)	Addresses adoption of AI in supply chain management; governance "implementation" appears as practical requirements for deploying AI in complex, inter-organisational settings (data governance, controls, process redesign, and oversight). Adaptation is strongly context-dependent (supply chain maturity, legacy systems, cross-partner data sharing).	Positions "effective governance" as the set of organisational controls that make AI adoption trustworthy and workable in supply chains: data governance, compliance, accountability, and transparency, embedded into operational decision-making.	Effectiveness is discussed inferentially: governance measures are presented as necessary to mitigate risks and enable sustainable adoption, rather than evaluated through outcome metrics of a specific framework.	High implementation cost, poor/fragmented data, cybersecurity and privacy risks, regulatory compliance burden, skills shortages, change resistance, integration challenges with legacy systems, and concerns about transparency/accountability of AI-supported decisions.	Data governance (quality, ownership, access); security/privacy controls; regulatory compliance; accountability and role clarity; transparency/traceability expectations; monitoring and continuous improvement; change management and capability building.

Lacmanovic and Skare (2025)	Synthesises how bias audits are conducted in practice and recommends embedding auditing into the AI lifecycle: scoping → measurement/evaluation → verification/validation, plus documentation and governance integration; highlights the need for repeatable processes and organisational support for audit access and follow-through.	Bias auditing as a governance mechanism: stresses fairness assessment, assurance practices, and linking audits to broader governance regimes (standards, risk classification, compliance expectations).	Effectiveness is presented as conditional: auditing can surface bias and improve accountability when resourced and institutionally supported, but outcomes vary by context, access, and clarity of standards.	Lack of consensus standards; metric and definition disputes (what counts as "fairness"); limited access to models/data; socio-technical complexity; auditing gaps between technical findings and organisational remediation.	Audit lifecycle model (scope,evaluate,verify); documentation/reporting; governance integration (roles, responsibilities, remediation pathways); ongoing monitoring and re-auditing; socio-technical stakeholder considerations.
AlTawil and Rahhal (2025)	Not AI-specific. Examines how CSR laws can be integrated with corporate governance structures in the UAE. "Implementation" is through formal governance mechanisms (board oversight, compliance processes, disclosure, enforcement arrangements), which can be *analogically* useful when thinking about hard-law vs soft-law approaches to AI governance.	Effective governance is framed as coherent alignment between legal duties (CSR requirements) and internal corporate governance, strengthening accountability and oversight. While not about AI, the architecture (law → governance controls → oversight) is transferable as a governance design pattern.	Effectiveness is conceptual: argues that alignment/synergy can strengthen accountability and responsible conduct, but does not evaluate a deployed framework's impact empirically.	Implementation and enforcement complexity, variability in organisational compliance capacity, reporting burden, and the risk that formal requirements become performative without robust oversight.	Board and committee oversight; compliance and enforcement mechanisms; disclosure/reporting; stakeholder accountability; whistleblowing and assurance/audit-type functions; integration of CSR strategy into governance routines.

Sandeep et al. (2025)	Organisational implementation is framed through HR capability: recommends continuous human oversight, regular audits, collaboration with AI developers, and compliance with data protection (e.g., consent/privacy safeguards) to adapt AI recruitment tools ethically over time.	Governance is embedded in HRM: ensuring AI recruitment aligns with organisational values and legal/ethical requirements, especially around fairness, transparency, explainability, and privacy in hiring decisions.	Reports relationships suggesting that strong capabilities/support enable more effective AI adoption and organisational advantage; ethical governance is treated as a necessary condition for trust and legitimacy rather than a separately quantified governance "effectiveness" score.	Algorithmic bias, discrimination, and lack of explainability; privacy and lawful data processing constraints; trust deficits among stakeholders; capability gaps and cultural resistance.	Capability-based framework: HR competency, open innovation, IT infrastructure, financial support; governance controls: audits, human-in-the-loop oversight, transparency/explainability practices, privacy/compliance safeguards.
Birkstedt et al. (2023)	Finds implementation knowledge is limited in the literature and calls for stronger empirical work on how AIG is enacted; proposes practical actions such as establishing an AI oversight unit, plus training/awareness and stakeholder-facing governance routines to translate ethics into practice.	Defines AI governance as a system of rules/practices/processes aligning AI use with strategy, values, legal requirements, ethical principles, and stakeholder requirements; identifies four governance themes: technology, stakeholders/context, regulation, processes.	Highlights that the effectiveness of ethical principles and regulation remains uncertain and under-evidenced; positions future research agendas to strengthen operationalisation and evaluation of AIG processes.	Fragmentation and few explicit definitions; weak evidence on implementation; limited attention to context; uncertain real-world impact of principles/regulation; insufficient operationalisation of governance processes.	Working definition of AIG; four themes (technology; stakeholders/context; regulation; processes); recommended organisational mechanisms: oversight unit, collaborative governance, training/awareness, and process operationalisation.

Alzebda and Matar (2025)	Emphasises practical levers that sit "around" adoption: public education/awareness, user-centred design, ethical AI development techniques, multi-stakeholder partnerships, and ongoing monitoring systems, with government regulation shaping how adoption drivers translate into intention.	Positions formal government regulation as a key governance enabler, alongside privacy/security protections and risk mitigation (bias, misuse).	Empirical (acceptance/adoption model): reports that adoption determinants (and regulation as a moderator) significantly shape citizens' intention to accept/adopt AI, rather than testing a specific organisational governance framework.	Risk perceptions; privacy/security fears; concerns about algorithmic bias and "unethical" outcomes; trust deficits.	Adoption model constructs (e.g., usefulness/ease, risk, trust/social influence), plus government regulation as a moderating governance variable; monitoring/feedback loops as an operational mechanism.
Zada et al. (2024)	Frames implementation as organisational adoption choices that should be aligned with ESG priorities: embed compliance-oriented practices, proactively manage ethical dilemmas, and design processes that reduce bias and improve transparency in FinTech deployments.	"Effective governance" is presented through an ESG-oriented lens (especially the "G"): trust, transparency, compliance, risk oversight, and bias mitigation while pursuing innovation and performance.	Empirical (PLS-SEM): finds GenAI adoption is associated with exploratory/exploitative innovation and organisational performance; ethical dilemmas are treated as a contextual factor (not always a significant moderator).	Ethical dilemmas in FinTech contexts; algorithmic bias risks; regulatory/compliance complexity; tensions between speed-to-market and responsible safeguards.	TOE framing (technology,organisation,environment) plus ESG considerations; constructs representing ethical dilemmas/risks; performance and innovation outcomes.

Saw and Ng (2022)	Describes an implementation pipeline for clinical contexts (data acquisition → model development → validation with clinical evidence → regulatory approval → clinical implementation). Stresses data governance (data sharing best practice), privacy protection, and stakeholder consensus to create trustworthy AI policies and regulatory frameworks.	Governance is risk- and safety-oriented: trustworthy AI policies, data governance, privacy/cybersecurity safeguards, and robust algorithms that are fair, transparent, and clinically reliable.	Narrative review: argues that weak methodological validation and governance gaps limit real-world translation; governance maturity (policies, standards, stakeholder alignment) is positioned as necessary for scalable clinical deployment.	Ethical/legal concerns; privacy and cyber risk; insufficient data sharing norms; lack of clear accountability/policy frameworks; limited clinical trust/skills readiness.	Four-part challenge structure frequently used as "components": algorithm robustness (fairness/transparency/trust), data governance, privacy protection, and trustworthy policies/regulatory frameworks; multi-stakeholder guidance/standards.
Yue et al. (2024)	Recommends internal implementation practices: education and training, a "co-piloting" human,AI approach, transparent communication, ethical governance, and an experimental culture. Also highlights the need for practical guidelines that embed organisational values into AI-enabled communication workflows.	Governance emphasises data ethics, transparency, accountability, privacy, and trust in workplace settings,particularly where internal communications may involve monitoring, sensitive employee data, or reputational risk.	Evidence is primarily practice-oriented (qualitative/interpretive): demonstrates AI's potential to improve efficiency and communication capacity; governance "effectiveness" is discussed as enabling responsible uptake rather than measured outcomes.	Authenticity concerns; accuracy/error rates in AI outputs; data privacy concerns; cultural/HR readiness barriers; trust and acceptance issues among employees.	Training and capability building; human,AI co-production ("copilot"); transparency and open communication; ethical governance structures; decision-making guidelines aligned to organisational values.

Palladino (2023)	Analyses how high-level ethical principles are translated into operational tools and practices, often by private actors. Implementation is characterised as a shift from broad ethics claims to narrower, tool-driven governance that can legitimise rather than constrain.	Critiques the emerging governance regime as structurally biased: effective governance must include social and institutional mechanisms (not only technical tools), and guard against value-narrowing driven by dominant actors.	Conceptual/critical analysis: argues the emergent regime tends to reframe ethics in constrained ways; "effectiveness" is questioned (i.e., governance may be performative or partial).	Institutional and power asymmetries; private-sector dominance; narrowing of ethics to what is measurable/auditable; gaps in social accountability.	Ethical principles/guidelines; translation mechanisms (standards, toolkits, audits/metrics); governance ecosystem (multi-actor regime); social accountability mechanisms beyond technical compliance.
Loufek et al. (2024)	Directly addresses ethical AI implementation in healthcare via *internal accountability*. Mayo Clinic operationalises governance by (i) building internal expertise, (ii) establishing a centralised Software as a Medical Device (SaMD) Review Board, and (iii) aligning development/deployment with regulations, standards, and best practices across the AI/DHT lifecycle.	A risk-based, institutionally embedded governance framework: multidisciplinary oversight, regulatory interpretation, and lifecycle quality management to ensure AI-enabled DHTs are safe, effective, and ethical in clinical practice.	Reports early organisational value (e.g., review capacity supporting multiple product teams and standardising risk/regulatory guidance). Effectiveness is described in terms of strengthened governance capacity and safer translation to practice, rather than comparative outcome evaluation.	Evolving regulation; limited embedded expertise in many institutions; resourcing constraints; potential bottlenecks; and acknowledged limitations such as strengthening patient/societal representation and broadening stakeholder representation in governance.	Internal capability building (regulatory/QA expertise); central review board; risk classification and regulatory determination; mitigation recommendations; integration with existing governance (legal/privacy/IRB); lifecycle controls (requirements/traceability, change/configuration management, verification/validation, cybersecurity, data management); continuous improvement.

Mac (2024)	Focuses on governance-by-design for administrative decision-making: stresses enforceable accountability and transparency measures (documentation, explainability, review/appeal pathways, auditing), and clearer allocation of responsibility when ML supports or makes decisions.	Effective governance is framed in legal-administrative terms: procedural fairness, contestability, transparency obligations, and accountability for discriminatory impacts in ML-assisted public decision-making.	Legal/analytical contribution: offers governance recommendations rather than outcome evaluation; effectiveness is argued as improved fairness and reduced discriminatory harm if safeguards are implemented.	Opacity ("black box" decision support); discrimination and disparate impacts; accountability gaps; difficulties operationalising transparency and meaningful explanation.	Transparency requirements; accountability allocation (roles/responsibility); audits/impact assessments; human oversight and review rights; procedural safeguards (appeal/contestability).
Wamba et al. (2025)	As an editorial synthesis, it highlights implementation needs across GenAI use: responsible governance, strategic alignment, and human,AI collaboration; calls for organisational and policy responses that operationalise transparency, fairness, and accountability in value creation.	Governance is positioned as the condition for "ethical value" from GenAI: controls and oversight to manage risks while enabling innovation; attention to societal impact and policy development.	Editorial synthesis (special issue framing): identifies opportunities and tensions; does not test a framework, but argues that responsible governance is required for sustainable value.	Ethical tensions; transparency/fairness/accountability concerns; risk and volatility; regulatory uncertainty; misalignment between rapid deployment and safeguards.	Responsible governance and oversight; strategic alignment; human,AI collaboration; risk management controls; policy/regulatory engagement.

Rana et al. (2024)	Provides explicit organisational governance actions: establish governance frameworks with clear roles/responsibilities for AI decision-making and issue handling; maintain documentation to ensure transparency; embed ethical principles into adoption processes under institutional pressures.	Integrates institutional theory with ethical guidelines: coercive/normative/mimetic pressures plus ethical principles (fairness, accountability, transparency, accuracy, autonomy) as drivers of responsible GenAI use and governance maturity.	Empirical (survey; PLS-SEM): finds institutional pressures and ethical principles influence GenAI use; GenAI use influences organisational performance; organisational innovativeness moderates the GenAI,performance relationship.	Institutional constraints; ethical principle trade-offs; transparency and accountability demands; governance capacity and role clarity; regulatory risk.	Institutional pressures (coercive/normative/mimetic); ethical principles (FAT + accuracy/autonomy); governance frameworks (roles, responsibilities, escalation); transparency via documentation; organisational innovativeness; performance outcomes.

Appendix D: Reflexive Memo Log

Table 10. Selected excerpts aligned to the dissertation

Date	Stage	Trigger	Decision made	Evidence anchor (study IDs)	Subtheme link	Reflexive note (bias/uncertainty + check)
05 July 2025	Screening → inclusion calibration	Noticed my inclusion decisions were drifting towards "ethics principles" papers rather than "governance in organisations"	Re-checked inclusion logic against RQ1/RQ2: only keep papers with organisational governance mechanisms and/or implementation content	S01, S02, S34	T1c principles-to-practice failure	I am drawn to influential "ethics principles" critiques; I will only keep these where they clearly inform organisational enactment and failures and explicitly label them as implementation gap evidence rather than effectiveness proof.
06 July 2025	Extraction (RQ1 vs RQ2 separation)	Several studies blend "governance structures" with "ethical goals", making it easy to conflate mechanisms with outcomes	Split extraction notes into: (a) governance mechanisms (structures/processes/tools) and (b) outcomes/effectiveness evidence	S04, S26, S35	T4a + T4b + T5d	I have a tendency to treat "having controls" as equivalent to "being effective". I will require explicit outcome indicators (trust, performance, reduced harms, compliance impact) before coding anything as effectiveness.
08 July 2025	Coding (operationalisation instruments)	Documentation tools appeared across disparate contexts (audits, checklists, impact assessments), risking a messy "grab bag" code	Consolidated into one code family: "operationalisation instruments", with subcodes for impact assessments, documentation, audit routines	S02, S26	T1a operationalisation instruments	My prior professional experience biases me towards audits as a "gold standard". To counter this, I will code what the tool does (documentation, assurance, accountability) rather than assuming effectiveness.

10 July 2025	Coding (lifecycle embedding boundary)	"MLOps/lifecycle controls" were being coded both as governance and as technical practice	Defined a boundary rule: lifecycle embedding counts as governance only when it creates decision rights, gates, monitoring obligations, or review routines	S04, S11, S17, S29, S36	T1b lifecycle embedding	Risk: I might under-count governance embedded in engineering practice. Check: when re-reading extraction, confirm whether lifecycle practices are tied to accountability and oversight structures.
12 July 2025	Theme development (principles-to-practice failure)	Three studies strongly problematise performative governance and "ethics washing", but in different language	Created a single subtheme to retain the shared phenomenon without flattening nuance (symbolic compliance vs skewed implementation)	S01, S02, S34	T1c principles-to-practice failure	I'm aware this subtheme can sound accusatory. I will write it with calibrated language ("may", "can", "in some contexts") and anchor it to the evidence base in the included studies.
14 July 2025	Sector coding (avoid over-generalisation)	"Healthcare governance" studies were rich, but risked becoming the implied default model of governance	Introduced explicit sector subthemes so that governance patterns are interpreted as context-contingent	S10, S12, S17, S35	T2a healthcare	I may be overweighting healthcare because it offers more concrete implementation descriptions. Check: ensure sector contrasts are used to limit generalisation, not to privilege one sector as universal.
16 July 2025	Coding (accountability structures vs assurance mechanisms)	Oversight bodies (boards/committees) and assurance practices (audits/monitoring) were blending into one large "governance" theme	Split Theme 4 into: internal accountability structures vs assurance mechanisms, to preserve analytic clarity	S35, S26, S30	T4a vs T4b	I tend to centre formal committees. I will check whether assurance mechanisms and lifecycle gates are actually doing more "governance work" than committees in some studies.

18 July 2025	Coding (external regulation alignment)	Large proportion of studies reference standards/regulation; risk that "compliance" dominates interpretation	Treated external alignment as a governance condition (scaffolding) rather than a proxy for effectiveness	S05, S14, S22, S36	T4c external regulation alignment	Bias risk: equating "more law" with "better governance". Check: in the discussion, distinguish rule presence from organisational capability to comply meaningfully.
20 July 2025	Theme refinement (risk-based governance)	Risk-based approaches appeared in both finance and broader governance discussions	Kept "risk-based/proportional governance" as its own subtheme, and used it to explain why governance designs differ by system criticality	S14, S20, S34, S35	T5a risk-based/proportional governance	I might be importing EU-style risk logic into all contexts. Check: ensure risk-based framing is described as a tendency in the evidence, not a universal requirement.
22 July 2025	Theme refinement (fairness and discrimination scope)	Fairness was cited widely, but often without operational detail	Coded "fairness/discrimination" separately from "mechanisms", and noted when studies provide concrete operationalisation vs principle statements	S01, S03, S16, S36	T5b fairness/equity & discrimination	Personal values make me want to treat fairness as central. Check: write cautiously where the evidence shows fairness is discussed more than it is implemented.
24 July 2025	GenAI-specific synthesis	GenAI papers emphasised value creation and adoption pressures; risk of treating novelty as justification for weaker governance	Added a distinct GenAI challenges subtheme; interpreted GenAI as amplifying existing governance gaps rather than creating wholly new ones	S31, S37, S38	T5c GenAI governance challenges	Bias risk: "GenAI exceptionalism". Check: compare GenAI governance claims with baseline governance mechanisms in Themes 1 and 4 to show continuity and where it genuinely shifts practice.

26 July 2025	Effectiveness claims calibration (writing)	Many studies imply effectiveness without measuring it; risk of overstating RQ2	Wrote a rule for Chapter 5: treat effectiveness as plural (trust, performance, value, harm reduction) and context-specific; downgrade certainty where outcomes are asserted, not evaluated	S31, S35, S37	T5d effectiveness outcomes	I am motivated to answer "what works". Check: explicitly signal evidential limits, and use the traceability matrix to support each effectiveness claim with the contributing studies.

Appendix E: Reflexive Memo Log

Table 11. Screening and inclusion decisions for selected excerpts

Date	Stage	Trigger (candidate)	Decision made	Evidence anchor	Screening rule applied	Reflexive note (bias/uncertainty + check)
02 July 2025	Gatekeeper screening (title/abstract)	CHATGPT - learning accelerator or demolisher of foreign language teaching and learning? (ID 189)	Excluded at Stage 1 (fails governance-in-organisations scope)	Candidate_ID 189 - decision: exclude	Gatekeeper trinity: AI + governance mechanism + ethical orientation (must all be present)	I initially hesitated because it mentions ChatGPT; however it is about language teaching, not governance mechanisms in organisations. Re-anchored to RQ1/RQ2 and excluded.
02 July 2025	Gatekeeper screening (title/abstract)	Blockchain and AI-based solutions for healthcare management: liver disease detection… (ID 186)	Excluded at Stage 1 (technical healthcare AI; governance not central)	Candidate_ID 186 - decision: exclude	Exclude primarily technical contributions lacking governance/ethics engagement	Healthcare can look "ethics-adjacent", but the title signals solution design rather than organisational governance. Applied technical-exclusion to avoid scope creep.
02 July 2025	Gatekeeper screening (title/abstract)	A blockchain framework for securing ambient health care systems (ID 225)	Excluded at Stage 1 (security/blockchain focus; AI governance not foregrounded)	Candidate_ID 225 - decision: exclude	Governance must be substantive (oversight/policies/audits/standards), not only technical security design	I noted my tendency to treat cybersecurity frameworks as governance by default. Here governance mechanisms were not explicit; excluded to keep governance meaning "controls + accountability".

03 July 2025	Gatekeeper screening (title/abstract)	Online exploitation: drawing the line for surveillance advertising in Europe (ID 1)	Borderline: retained as "maybe" pending abstract check	Candidate_ID 1 - decision: maybe	Retain policy papers only if they connect to organisational enactment (compliance/controls)	Regulatory interest can inflate inclusion. Tagged as borderline: governance language is strong, but organisational implementation is unclear.
03 July 2025	Stage 2 scoring (6/6 threshold)	Leveraging blockchain and AI for ethical decision-making in healthcare firms… (ID 2)	Advanced to scoring (needs proof of implementation/evaluation)	Candidate_ID 2 - decision: maybe	6/6 rubric: org context (+1), implementation (+2), evaluative evidence (+2), empirical basis (+1)	"Framework" language triggered positive bias. Scoring forces evidence of enactment + evaluation, not rhetoric.
04 July 2025	Stage 2 scoring calibration	Human-AI intersection: ethical challenges… and governance protocols… (ID 3)	Kept as "maybe" until scoring; likely to fall below threshold if only conceptual	Candidate_ID 3 - decision: maybe	Calibration check to prevent drift towards principle-only work	I caught myself awarding "implementation" points for protocol-level discussion. Tightened interpretation: implementation requires organisational processes/roles/controls.
05 July 2025	Inclusion decision (candidate list: include)	Artificial intelligence governance: ethical considerations and implications for social responsibility (ID 8)	Included for synthesis	Candidate_ID 8 - decision: include	Must contribute directly to governance focus (RQ2) and/or implementation (RQ1)	I checked it was more than generic "responsibility" talk; retained because it centres governance implications rather than only principles.

05 July 2025	Inclusion decision (candidate list: include)	A management control oriented governance framework for artificial intelligence (ID 74)	Included (framework is operationalisable; supports component mapping)	Candidate_ID 74 - decision: include	Include frameworks that specify governance controls/structures that can be operationalised	I have a preference for control-oriented governance; inclusion was justified by explicit control mechanisms (monitoring/accountability), not comfort with the paradigm.
06 July 2025	Soft law vs hard law boundary	Mitigating adverse effects of AI with the EU AI Act: hype or hope? (ID 19)	Included, but flagged for "implementation inference risk"	Candidate_ID 19 - decision: include	Regulatory papers included only when linked to organisational governance obligations/compliance mechanisms	Regulation can masquerade as "effectiveness evidence". Any org-effectiveness claims from legal text will be framed as *inferred* unless empirically evaluated.
07 July 2025	Scope creep check (generative AI)	Organising projects for responsible use of generative AI in project management (ID 35)	Retained as "maybe" pending confirmation governance mechanisms are central	Candidate_ID 35 - decision: maybe	Only include GenAI/project papers that outline governance controls (committees, audits, policies, risk gates).	Selection may be biassed by GanAI's salience. Until it exhibits governance architecture (rather than merely project methodologies), it remains borderline.