

**RESPONSIBLE AI GUARDRAILS FOR ENTERPRISES: A MANAGERIAL
FRAMEWORK FOR BALANCING INNOVATION, GOVERNANCE, AND
TRUST A STRUCTURED LITERATURE REVIEW**

Yogesh Sabnis

Program Manager, Intuit Inc

Email: yogeshsabnis1@gmail.com

Abstract

The rapid adoption of artificial intelligence (AI) in enterprises has generated increasing concern about the ethical, regulatory, and managerial guardrails needed to ensure responsible deployment. While numerous frameworks emphasize principles such as fairness, transparency, and accountability, the literature reveals persistent gaps in translating these values into operational and organizational practice. This study addresses that gap by conducting a structured literature review of scholarship, standards, and policy documents published between 2018 and 2025. Using PRISMA guidelines, approximately thirty high-quality sources were identified and synthesized across three thematic clusters: ethical principles and governance, operational tools and practices, and managerial perspectives.

Building on this synthesis, the paper proposes a three-layer framework for responsible AI guardrails. The strategic layer emphasizes alignment with external regulations and internal governance commitments; the operational layer focuses on documentation, auditing, and risk management tools, and the managerial layer highlights accountability, leadership engagement, and integration into enterprise decision-making.

This study contributes to theory by consolidating fragmented streams of responsible AI scholarship into a unified, multi-layered governance framework. It contributes to practice by offering managers actionable guidance for embedding responsible AI across organizational levels, thereby enabling enterprises to balance innovation with accountability and trust.

Keywords: Responsible AI, AI governance, algorithmic accountability, ethical guardrails, enterprise risk management, artificial intelligence ethics.

1. Introduction

Artificial intelligence (AI) has moved from experimental deployments to mainstream adoption across industries, reshaping how enterprises operate, innovate, and compete. While the transformative potential of AI is well recognized, its rapid integration into business processes raises pressing concerns around fairness, accountability, transparency, and trustworthiness (Floridi & Cowls, 2019; Jobin et al., 2019). High-profile failures in algorithmic decision-making—ranging from biased recruitment tools to unreliable facial recognition systems—have underscored the risks of deploying AI without adequate oversight or governance (Raji &

Buolamwini, 2019; Mittelstadt, 2019). As a result, the global conversation has shifted from whether AI should be adopted to how it can be adopted responsibly.

Governments, standards bodies, and industry leaders have responded with a proliferation of principles, guidelines, and regulatory frameworks. Initiatives such as the European Union's *Ethics Guidelines for Trustworthy AI* (European Commission, 2019), the OECD Recommendation on AI (OECD, 2019), and the U.S. National Institute of Standards and Technology's AI Risk Management Framework (NIST, 2023) reflect a growing consensus on the need for ethical and responsible AI. These frameworks provide high-level principles that emphasize transparency, accountability, privacy, and human oversight. At the same time, companies such as Google and Microsoft have introduced internal governance structures and responsible AI toolkits to align product development with ethical standards (Mitchell et al., 2019; Gebru et al., 2021). Despite this progress, challenges remain in translating abstract principles into operational guardrails that managers can implement consistently across complex enterprise environments (Morley et al., 2021; ISO/IEC 38507:2022; Mökander & Floridi, 2021).

Academic and policy debates to date reveal two critical shortcomings. First, most contributions remain principle-driven rather than practice-driven, offering values without actionable pathways for integration into enterprise workflows (Stahl et al., 2021). Second, the majority of existing frameworks focus on technical or regulatory dimensions, with limited attention to the managerial perspective—how leaders and program managers can embed ethical guardrails into day-to-day decision-making, risk management, and organizational governance (ISO/IEC 38507:2022; Stahl et al., 2021). This misalignment between principle and practice contributes to what Raji et al. (2020) term the “AI accountability gap,” wherein enterprises adopt AI technologies faster than they establish systems to monitor and control them.

Against this backdrop, there is a need for scholarship that consolidates existing ethical frameworks, identifies gaps in managerial applicability, and proposes a conceptual structure tailored to enterprise settings. This paper responds to that need by conducting a structured review of recent literature and policy developments on responsible AI and guardrails. Drawing on insights from academic research, international standards, and corporate practices, the study proposes a three-layered framework—strategic, operational, and managerial guardrails—that can guide enterprises in balancing innovation with governance and trust. By addressing the practical dimension of responsible AI implementation, this paper contributes to both the academic literature and managerial practice, offering a pathway for enterprises to operationalize ethical principles in ways that are scalable, accountable, and aligned with regulatory expectations.

Building on these debates, this study makes three contributions. First, it integrates regulatory requirements (e.g., the EU AI Act), risk management frameworks (NIST AI RMF; ISO/IEC 23894), and governance system standards (ISO/IEC 42001; ISO/IEC 38507) into a single enterprise architecture that anchors responsible AI in managerial decision rights and oversight. Second, it reframes “guardrails” as an organizational capability by specifying how strategy,

product development, compliance, and risk functions interact to make ethics operational at scale. Third, it provides a practice-ready crosswalk that connects principles and tools to concrete managerial actions, thereby narrowing the accountability gap reported in prior work. Together these contributions extend existing AI governance models—which typically emphasize technical controls or compliance alone—by articulating a layered managerial pathway that enterprises can implement and evaluate.

1.1 Originality and value

This framework extends prior AI governance approaches in three ways. It widens the scope beyond technical safeguards by coupling legal and standards-based obligations with explicit managerial decision rights and accountability. It shifts the level of analysis from individual models or use cases to the enterprise, linking board-level oversight to day-to-day product and risk workflows. And it operationalizes adoption through a traceable mapping—from principles to tools to managerial actions—so that guardrails become part of routine governance rather than checklist compliance. In this sense, the framework complements and extends policy- and tool-centric models by specifying how managers make responsible AI durable across functions and time.

2. Methodology

2.1 Review design and rationale

This study employs a structured literature review (SLR) to develop a managerial framework for responsible AI guardrails. An SLR is appropriate for consolidating a rapidly expanding, multidisciplinary body of work and for translating dispersed insights into an actionable conceptual model for enterprises (Snyder, 2019; Tranfield et al., 2003). Consistent with best practice, we foreground transparency, reproducibility, and comprehensiveness, while keeping the process proportionate to a conceptual (non-empirical) article.

Scope and positioning. The review focuses on publications from January 2018 to September 2025, when research on responsible AI, governance, and guardrails accelerated alongside major regulatory and standards activity. We target three streams: (1) ethical principles and governance (e.g., OECD, EU, NIST, ISO/IEC), (2) operationalization tools (e.g., documentation, audits, risk management), and (3) managerial/organizational perspectives (leadership, accountability, assurance). The unit of analysis comprises peer-reviewed journal articles, tier-one conferences (AAAI/ACM/IEEE, FAccT/AIES), and authoritative standards or regulatory documents (e.g., EU AI Act, NIST AI RMF, ISO/IEC). We prioritize peer-reviewed and official sources; influential preprints or reports are included only when issued by recognized institutions and widely referenced.

The review is guided by four research questions: (RQ1) What guardrails for responsible AI are proposed across scholarship, standards, and policy? (RQ2) How are high-level principles translated into operational practices within enterprises? (RQ3) Where do managerial gaps persist, such as accountability, decision rights, and assurance? (RQ4) What integrated framework can align strategic, operational, and managerial guardrails?

Process backbone. We adopt an evidence-informed management review approach (Tranfield et al., 2003) and document the identification–screening–eligibility–inclusion flow using the PRISMA 2020 reporting logic for clarity and auditability (Page et al., 2021). Subsequent subsections detail the search strategy and sources (2.2), criteria and screening (2.3), data extraction and coding (2.4), synthesis approach (2.5), quality appraisal (2.6), and limitations (2.7).

The identification–screening–eligibility–inclusion flow for this review is shown in Figure 1.

2.2 Search Strategy and Sources

To ensure comprehensive and unbiased coverage of the literature, a structured search protocol was employed across multiple databases and repositories. Following established guidance on systematic and structured reviews in management and technology research (Snyder, 2019; Tranfield et al., 2003), the search strategy combined peer-reviewed scholarship with authoritative policy and standards documents.

Peer-reviewed literature was retrieved primarily from *Scopus* and *Web of Science*, which are recognized as the most authoritative indexing services for scholarly publications. Supplementary searches were conducted in *IEEE Xplore*, *ACM Digital Library*, and *SpringerLink* to capture conference proceedings relevant to AI ethics, auditing, and governance. Google Scholar was used as a secondary source to ensure completeness, although inclusion from this source was conditional on the outlet being indexed in Scopus or Web of Science.

Recognizing that AI guardrails are shaped as much by regulatory frameworks and industry standards as by academic debates, the search strategy also incorporated official sources. Documents were obtained directly from the European Commission (EU AI Act and *Ethics Guidelines for Trustworthy AI*), the OECD (Recommendation on AI), the National Institute of Standards and Technology (NIST) (AI RMF and playbooks), and the International Organization for Standardization (ISO/IEC) (risk management and governance standards). National regulatory guidance, such as the UK Information Commissioner’s Office reports on AI and data protection, was also included.

Search queries were developed iteratively using Boolean operators and phrase matching to balance precision with recall. Representative search strings included terms such as “Responsible AI” AND (framework OR guardrails OR governance), “AI ethics” AND (“enterprise” OR “management” OR “organizational”), “Algorithmic auditing” OR “AI accountability”, “AI governance” AND (“standards” OR “ISO/IEC” OR “NIST”), and “Trustworthy AI” AND (“regulation” OR “policy”). Searches were restricted to publications between January 2018 and September 2025 to capture contemporary debates and developments, while also ensuring consistency with the major regulatory and corporate initiatives launched during this period (Jobin et al., 2019; Morley et al., 2021).

2.3 Inclusion and Exclusion Criteria

To maintain rigor and transparency in the review, predefined inclusion and exclusion criteria were applied throughout the screening process. Only studies that directly addressed responsible AI, ethical guardrails, governance, or managerial implications were considered eligible. Peer-reviewed journal articles and tier-one conference proceedings were prioritized because of their scholarly credibility and indexing in Scopus or Web of Science. Official policy and standards documents were also included when published by recognized authorities such as the European Commission, OECD, NIST, ISO/IEC, and the UK Information Commissioner’s Office.

The review excluded sources that did not meet minimum standards of scholarly or institutional credibility. Articles published in non-indexed journals, opinion blogs, or outlets without peer review were omitted. To avoid the challenges experienced in earlier research projects with low-quality references, documents were admitted only when they could be verified as published in high-quality journals or issued by globally recognized organizations. The timeframe for inclusion was limited to publications between January 2018 and September 2025 to ensure contemporary relevance, reflecting the period when responsible AI became a mainstream research and policy concern (Jobin et al., 2019; Morley et al., 2021). Older works were incorporated selectively only if they provided seminal methodological foundations or frameworks that remain highly cited in the present discourse (Tranfield et al., 2003).

This approach ensured that the final dataset was both comprehensive and credible, capturing the breadth of ongoing debates in responsible AI while avoiding the dilution of quality that can result from indiscriminate inclusion. By combining peer-reviewed research with authoritative regulatory and standards sources, the review balances academic robustness with practical policy relevance, providing a strong foundation for developing a conceptual framework for enterprise-level guardrails. The criteria applied during screening are summarized in Table 1.

Table 1. *Inclusion and exclusion criteria used for structured literature review.*

Criterion	Inclusion	Exclusion
Publication type	Peer-reviewed journal articles, Scopus/Web of Science indexed conference proceedings, official policy and standards documents (EU, OECD, NIST, ISO/IEC).	Non-peer-reviewed articles, consultancy white papers, blogs, or reports lacking methodological transparency.
Timeframe	January 2018 – September 2025 (period of major AI ethics and governance developments).	Publications prior to 2018, unless foundational methodological works (e.g., Tranfield et al., 2003).

Content focus	Studies addressing responsible AI, ethical principles, governance, managerial implications, or operational guardrails (e.g., documentation, auditing, risk management).	Works only tangentially related to AI (e.g., general IT ethics, unrelated machine learning applications).
Credibility	Indexed in Scopus/Web of Science or published by globally recognized institutions.	Sources from non-indexed, low-quality, or predatory journals.
Relevance	Directly applicable to enterprise-level AI adoption and governance.	Literature without explicit connection to AI ethics, guardrails, or organizational governance.

2.4 Data Extraction and Coding

After applying the inclusion and exclusion criteria, eligible documents were subjected to a structured process of data extraction and coding. Each article, conference paper, or standards document was reviewed in full and summarized using a standardized template. The template captured bibliographic details, publication type, primary research focus, methodological orientation (where applicable), and the specific contributions related to AI ethics, guardrails, or governance. For policy and standards documents, extraction focused on stated principles, risk management provisions, and guidance for organizational adoption.

To synthesize insights across diverse sources, a coding scheme was developed that grouped findings into three overarching categories: ethical principles and governance, operational tools and practices, and managerial implications. This deductive structure was guided by the research questions and refined iteratively as additional themes emerged during the review. Thematic clusters were generated to highlight areas of convergence across scholarship and regulatory guidance, as well as to identify inconsistencies or gaps in how responsible AI is framed for enterprise adoption.

Coding was carried out manually, following recommendations from structured literature review methodology to ensure interpretive rigor and transparency (Snyder, 2019; Tranfield et al., 2003). Discrepancies in classification were revisited until consensus was achieved, thereby strengthening the reliability of the thematic synthesis. By integrating insights across academic, regulatory, and industry sources, the coding process created a coherent evidence base to inform the development of the conceptual framework presented later in this paper.

2.5 Quality Appraisal

To strengthen the credibility of the review, a quality appraisal process was applied to all included sources. For peer-reviewed journal articles and conference proceedings, quality was assessed on the basis of journal reputation, citation count, and methodological clarity. Articles

published in high-impact journals indexed in Scopus or Web of Science were prioritized, and studies from recognized conferences such as ACM FAccT, AAAI/ACM AIES, and IEEE venues were included only when they demonstrated a clear contribution to the discourse on AI ethics or governance. This ensured that the analysis was grounded in scholarship that met rigorous academic standards (Snyder, 2019).

Policy and standards documents underwent a different appraisal process, reflecting their normative rather than empirical orientation. Documents were included only if they were issued by globally recognized institutions such as the European Commission, OECD, NIST, ISO/IEC, or the UK Information Commissioner's Office. This criterion prevented the inclusion of industry white papers or consultancy reports that may lack transparency in methodology or institutional legitimacy. In this way, the appraisal process safeguarded against the quality issues that can arise when literature reviews incorporate unverified or non-peer-reviewed material (Tranfield et al., 2003).

In addition to source credibility, temporal relevance was considered. Consistent with established best practices in management and technology research, publications were limited primarily to those issued after 2018, the period during which responsible AI emerged as a mainstream area of study and regulation (Jobin et al., 2019). Older works were included sparingly, and only when they provided foundational methodological guidance that remains widely cited in current scholarship. By applying these criteria, the final dataset achieved a balance between breadth and depth, capturing both the most recent developments and the enduring frameworks that underpin contemporary debates.

2.6 Limitations

Although the structured review followed established methodological guidelines, several limitations must be acknowledged. First, the study did not include empirical validation through interviews, surveys, or case studies. While this aligns with the conceptual aim of the paper, it restricts the ability to test the proposed framework in real-world enterprise contexts. Future research should therefore complement this review with empirical studies to examine how responsible AI guardrails are implemented and adapted within organizations.

Second, despite the use of multiple databases and authoritative sources, it is possible that some relevant publications were not captured, particularly works published outside Scopus or Web of Science or those released very recently. To mitigate this risk, supplementary searches were conducted using Google Scholar, and policy and standards documents were obtained directly from institutional websites. Nevertheless, the review cannot claim to be exhaustive, and the rapidly evolving nature of AI governance means that new guidelines or frameworks may have emerged after the completion of the search process (Morley et al., 2021).

Third, the appraisal process, while rigorous, may have introduced subjectivity in judging source credibility, particularly when differentiating between widely cited preprints and formally peer-reviewed publications. This limitation is inherent in conceptual literature reviews and was

mitigated by prioritizing high-impact journals, tier-one conferences, and authoritative policy sources (Snyder, 2019).

Finally, the focus on publications from 2018 onward, while necessary for ensuring contemporary relevance, may have excluded earlier discussions that could provide additional historical perspective on the ethical governance of emerging technologies. However, the inclusion of selected seminal works helped preserve continuity with the foundational literature (Tranfield et al., 2003). Overall, these limitations do not undermine the validity of the findings but highlight the importance of viewing the proposed framework as a basis for further scholarly and practical exploration.

The search and selection process followed PRISMA 2020 guidelines (Page et al., 2021).

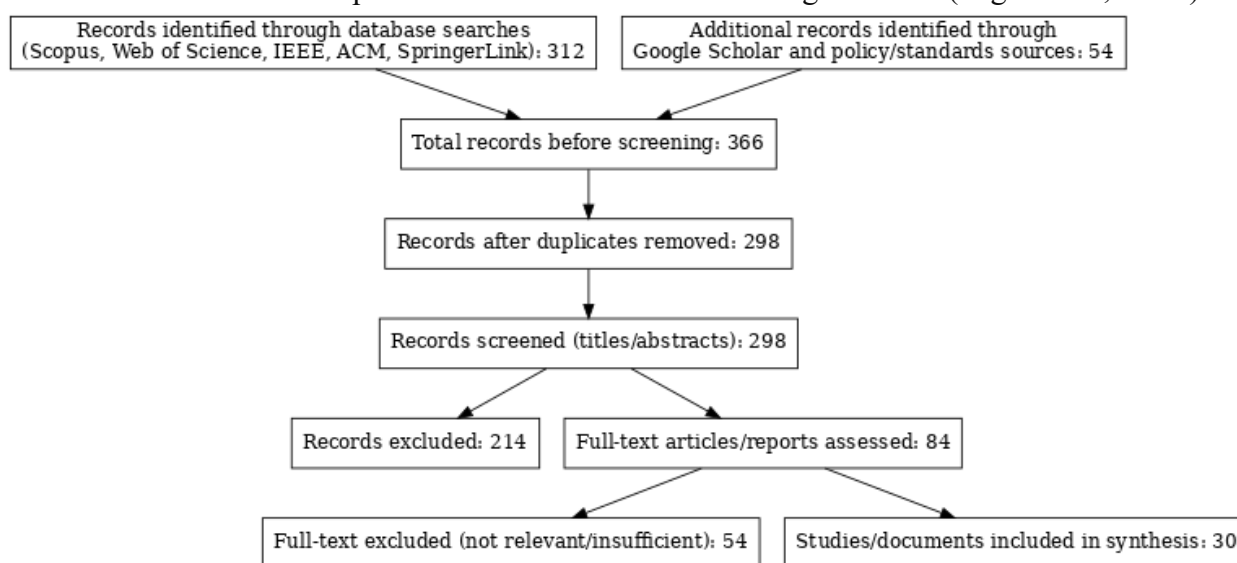


Figure 1. PRISMA flow diagram of literature search and selection process

3. Literature Review

3.1 Ethical Principles and Governance

The literature on responsible artificial intelligence is anchored in a set of widely endorsed ethical principles, often articulated as fairness, accountability, transparency, privacy, and human oversight. These principles emerged as consensus themes across a variety of ethical guidelines developed between 2018 and 2021, as governments, international organizations, and industry stakeholders sought to establish boundaries for the use of AI in society (Floridi & Cowls, 2019; Jobin et al., 2019). Although phrased differently across sources, these principles converge on the notion that AI must be deployed in ways that respect human rights, promote trust, and minimize harm.

International institutions have played a pivotal role in codifying these principles. The OECD's *Recommendation on Artificial Intelligence* (2019) is one of the earliest intergovernmental instruments to establish a globally recognized framework, emphasizing values such as inclusivity, accountability, and transparency. The European Union's *Ethics Guidelines for Trustworthy AI* (European Commission, 2019) further operationalized these principles into

three requirements: AI must be lawful, ethical, and robust. These guidelines provided a foundation for the later AI Act, which transformed ethical aspirations into legally binding obligations (European Commission, 2024). Together, these documents represent a shift from voluntary codes of conduct to enforceable governance structures that explicitly recognize the risks posed by high-risk AI systems.

Scholarly research complements these institutional efforts by critically assessing both the strengths and limitations of principle-driven approaches. Mittelstadt (2019) argues that high-level principles, while necessary, are insufficient to guarantee ethical AI because they do not specify mechanisms for accountability. Stahl, Markus, and Mittelstadt (2021) similarly highlight the risk of “ethics washing,” where organizations adopt the language of principles without demonstrating meaningful adherence. These critiques underscore the need to move from aspirational principles to operational frameworks that can be monitored and enforced in enterprise settings.

Despite the convergence around principles, gaps remain in how these values are interpreted and applied. Different institutions emphasize different priorities, and tensions often arise between innovation goals and compliance obligations. For example, while fairness and transparency are frequently cited as essential, their operationalization varies significantly across contexts and industries. This inconsistency reinforces the importance of developing guardrails that can translate principles into actionable guidance for organizations. Such translation requires not only technical interventions but also managerial oversight to ensure that principles are integrated into business processes rather than remaining abstract commitments (ISO/IEC 38507:2022; Morley et al., 2021).

3.2 Operational Tools and Practices

While ethical principles provide the normative foundation for responsible AI, the literature increasingly emphasizes the need for concrete tools and practices that translate values into measurable outcomes. Over the past five years, scholars and practitioners have introduced a range of mechanisms to improve transparency, accountability, and reliability in AI systems. Among the most influential contributions are documentation frameworks such as *Model Cards* and *Datasheets for Datasets*. Mitchell et al. (2019) propose model cards as standardized documentation to communicate model performance, intended use, and limitations, thereby enabling greater accountability in deployment. Similarly, Gebru et al. (2021) introduce datasheets as a method for documenting dataset characteristics, biases, and collection processes, ensuring that data provenance and risks are clearly understood before model training. Both tools illustrate the movement from abstract principles to operational guardrails that can be integrated into enterprise workflows.

Another major strand of operationalization literature focuses on auditing and assurance mechanisms. Raji and Buolamwini (2019) demonstrate how external audits of commercial facial recognition systems revealed systemic biases, prompting both regulatory attention and public accountability. Building on this work, Raji et al. (2020) propose an end-to-end framework for internal algorithmic auditing, emphasizing processes such as pre-deployment

review, bias testing, and ongoing monitoring. These approaches highlight the importance of embedding auditability into the AI lifecycle as a means of enforcing ethical commitments.

Policy-oriented institutions have similarly recognized the need for operational tools. The National Institute of Standards and Technology (NIST, 2023) developed the AI Risk Management Framework, which provides a structured approach organized around four functions—govern, map, measure, and manage. The associated playbook offers organizations detailed guidance for implementing risk controls, ranging from bias evaluation to incident response. Complementary standards such as ISO/IEC 23894:2023, which provides guidance on AI risk management, and ISO/IEC 42001:2023, the first management system standard for AI, reinforce the growing demand for formalized practices that enterprises can adopt.

Despite the progress made in developing such tools, the literature identifies persistent challenges. Many organizations lack the expertise or resources to implement these mechanisms consistently, and even when tools are adopted, they are often treated as checklists rather than embedded governance processes (Morley et al., 2021). Moreover, operational tools remain largely technical in orientation, leaving gaps in managerial oversight and accountability. These limitations highlight the need for frameworks that connect technical safeguards with organizational governance structures, ensuring that operational practices are supported by leadership, compliance teams, and cross-functional stakeholders.

3.3 Managerial Perspectives and Gaps

Although ethical principles and operational tools have received significant scholarly and regulatory attention, the managerial dimension of responsible AI remains comparatively underexplored. Most existing frameworks prioritize either technical interventions, such as bias detection and model auditing, or regulatory compliance mechanisms, such as conformity assessments under the European Union AI Act (European Commission, 2024). Yet, the successful adoption of responsible AI in enterprises depends equally on how managers, program leaders, and governance bodies embed guardrails into organizational processes and decision-making structures (ISO/IEC 38507:2022; Mökander & Floridi, 2021).

Research on managerial roles highlights several persistent challenges. First, there is a lack of clarity regarding accountability for AI-related risks within organizations. Ethical failures are often framed as technical shortcomings, while responsibility for oversight and remediation remains diffuse across business units and functions. Scholars emphasize that managers must assume active roles in defining boundaries for acceptable AI use, translating high-level ethical principles into concrete decisions about scope, risk tolerance, and operational trade-offs (ISO/IEC 38507:2022; Mökander & Floridi, 2021; Stahl et al., 2021). Second, the integration of ethical guardrails into program management practices is inconsistent. Studies show that while organizations may establish AI ethics boards or working groups, these bodies frequently lack the authority or resources to influence product development pipelines (Stahl et al., 2021).

A further gap lies in aligning responsible AI initiatives with broader enterprise governance systems. ISO/IEC 38507:2022 highlights the governance implications of AI, stressing the

importance of linking AI risk oversight to corporate governance and board-level accountability. However, the practical translation of these standards into organizational policies and management processes remains limited, with few empirical studies documenting successful enterprise-level adoption. This disconnection contributes to what Raji et al. (2020) identify as the “AI accountability gap,” where high-level commitments to ethics coexist with weak mechanisms for managerial enforcement.

Scholars also point to cultural and organizational barriers that hinder managerial engagement with responsible AI. These include tensions between innovation speed and compliance obligations, limited ethical literacy among managers, and inadequate incentives for prioritizing long-term trust over short-term performance gains (Morley et al., 2021). Such barriers suggest that the responsible AI discourse has advanced further in theory and policy than in managerial practice. Addressing these gaps requires frameworks that explicitly position managers as central actors in translating principles and tools into sustainable enterprise guardrails.

3.4 Summary of Literature Gaps

The review of ethical principles, operational tools, and managerial perspectives highlights important progress as well as persistent gaps in the responsible AI discourse. Ethical frameworks developed by intergovernmental organizations and regulators have created broad consensus on the values that should guide AI adoption, such as fairness, accountability, transparency, and human oversight (Floridi & Cowsls, 2019; OECD, 2019; European Commission, 2019). Operational tools including model documentation practices, auditing protocols, and risk management frameworks demonstrate tangible ways to implement these principles in practice (Mitchell et al., 2019; Gebru et al., 2021; Raji et al., 2020; NIST, 2023). However, while principles and tools provide necessary foundations, they remain insufficient for ensuring consistent and sustainable enterprise adoption.

The most significant gap lies in the limited attention given to managerial roles and organizational governance. Scholarship and policy often emphasize technical or regulatory solutions, leaving questions of accountability, leadership, and integration into enterprise decision-making underdeveloped (Markus, 2020; Stahl et al., 2021). Even where standards such as ISO/IEC 38507:2022 and ISO/IEC 42001:2023 outline governance implications, empirical studies documenting their translation into management practice are scarce. This disconnect reinforces the so-called “AI accountability gap” identified by Raji et al. (2020), in which commitments to responsible AI coexist with weak institutional mechanisms for oversight and enforcement.

In sum, the literature establishes a strong normative and technical foundation for responsible AI but falls short in bridging the gap between principles and practice at the managerial level. Addressing this shortcoming requires a framework that explicitly integrates strategic, operational, and managerial guardrails. The next section proposes such a framework, synthesizing insights from the literature to provide a practical model that enterprises can use to balance innovation, governance, and trust. Table 2 consolidates the themes emerging from the literature review, highlighting both progress and persistent gaps.

Table 2. *Thematic synthesis of literature on responsible AI principles, operational tools, and managerial gaps.*

Theme	Key References	Insights and Contributions
Ethical principles and governance	Floridi & Cows (2019); Jobin et al. (2019); OECD (2019); European Commission (2019, 2024); Mittelstadt (2019); Stahl et al. (2021)	Established consensus on principles such as fairness, accountability, transparency, and human oversight. EU and OECD guidelines provide normative frameworks, while scholars highlight risks of “ethics washing” and insufficient mechanisms for accountability.
Operational tools and practices	Mitchell et al. (2019); Gebru et al. (2021); Raji & Buolamwini (2019); Raji et al. (2020); NIST (2023); ISO/IEC 23894:2023; ISO/IEC 42001:2023; Morley et al. (2021)	Practical mechanisms for implementation, including documentation (model cards, datasheets), auditing (internal and external), and structured risk management frameworks. Challenges include inconsistent adoption, limited expertise, and checklist-style use rather than integrated governance.
Managerial perspectives and gaps	ISO/IEC 38507:2022; Mökander & Floridi (2021); Stahl et al. (2021); Morley et al. (2021); Raji et al. (2020)	Highlights accountability gaps and organizational barriers. Managers often lack clear decision rights or resources to enforce responsible AI. Standards suggest board-level responsibilities, but empirical adoption remains limited, reinforcing the need for managerial guardrails.

4. Proposed Framework

4.1 Overview and Rationale

The preceding review demonstrates that while responsible AI has been widely discussed in terms of principles and operational practices, there remains a pressing need for frameworks that directly address the managerial dimension of implementation. Principles articulated by the OECD (2019) and the European Commission (2019, 2024) provide high-level guidance, and operational tools such as model cards, datasheets, and algorithmic audits offer practical mechanisms for embedding transparency and accountability (Mitchell et al., 2019; Gebru et al., 2021; Raji et al., 2020). Yet these contributions are rarely integrated into organizational governance structures in ways that ensure sustained adherence or enterprise-wide alignment (Morley et al., 2021; ISO/IEC 38507:2022; Stahl et al., 2021).

To bridge this gap, the present study proposes a three-layer framework for enterprise guardrails in responsible AI, organized at the strategic, operational, and managerial levels. The purpose of this framework is to translate ethical aspirations into actionable governance structures that enterprises can adopt consistently across functions, products, and geographies. The strategic layer emphasizes alignment with laws, standards, and institutional ethics commitments; the operational layer focuses on technical and process-oriented practices for risk mitigation and auditing; and the managerial layer highlights leadership, accountability, and organizational integration. The three-layer structure is presented in Figure 2.

The rationale for a layered approach is twofold. First, responsible AI is inherently multidimensional, requiring integration across legal, technical, and organizational domains. No single layer of intervention is sufficient to address the complexity of AI risks. Second, enterprises operate within hierarchical structures in which strategic directives must be translated into operational practices and monitored through managerial oversight. A layered model therefore mirrors the governance realities of large organizations while addressing the accountability gap identified in prior scholarship (Raji et al., 2020; Stahl et al., 2021).

By synthesizing insights from ethical guidelines, operational tools, and management literature, this framework provides a structured yet adaptable model that enterprises can use to balance innovation with governance and trust. The following subsections elaborate on each of the three layers, describing their core components, institutional foundations, and interdependencies.

Figure 2 illustrates the three-layer framework, which integrates strategic, operational, and managerial guardrails to provide enterprises with a structured approach to responsible AI adoption.

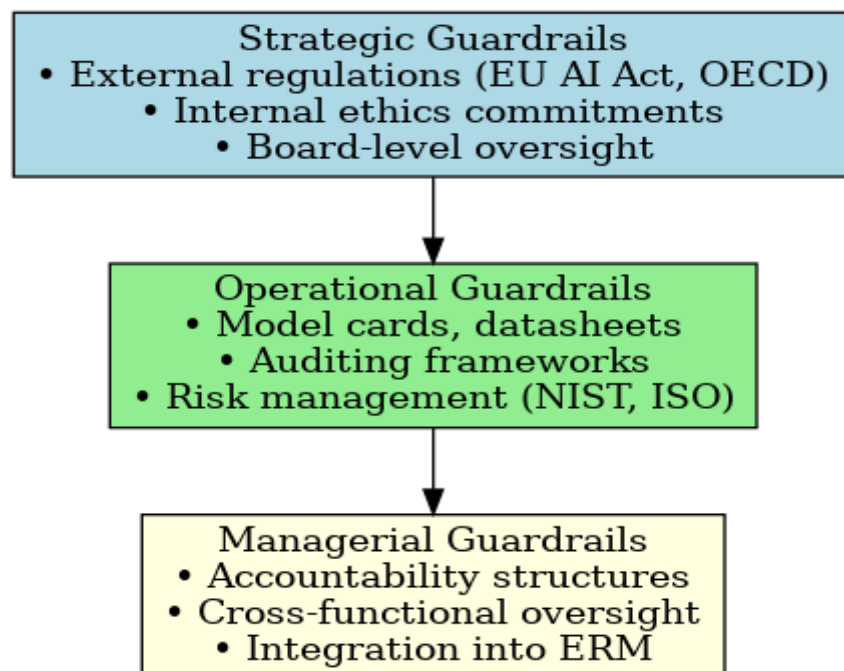


Figure 2. Conceptual framework of responsible AI guardrails at strategic, operational, and managerial levels

4.2 Strategic Guardrails

The strategic layer of the framework provides the foundation for responsible AI by ensuring that organizational practices align with legal, ethical, and societal expectations. This layer emphasizes compliance, institutional commitment, and value alignment, serving as the highest-level guardrail that shapes how enterprises adopt and govern AI. Strategic guardrails operate at the level of corporate governance, board oversight, and enterprise-wide policy, ensuring that ethical commitments are embedded into organizational strategy rather than treated as isolated technical measures.

Several key instruments illustrate the importance of this layer. The European Union's *AI Act* (European Commission, 2024) represents the most comprehensive regulatory framework to date, requiring enterprises to classify AI systems according to risk and implement mandatory safeguards for high-risk applications. Complementing this, the OECD Recommendation on AI (OECD, 2019) emphasizes values such as human-centered design, transparency, and accountability, and has been adopted by more than forty countries as a baseline for policy. Similarly, the NIST AI Risk Management Framework (NIST, 2023) provides a structured model for aligning organizational AI practices with risk-based governance functions. Collectively, these instruments establish external expectations that enterprises must address strategically to maintain legitimacy and trust.

In addition to external requirements, strategic guardrails involve internal governance commitments. Organizations increasingly publish responsible AI principles or ethics charters that signal their commitment to fairness, transparency, and inclusivity. While such statements are often criticized as insufficient on their own (Mittelstadt, 2019; Stahl et al., 2021), they can serve as entry points for embedding values into enterprise policies when combined with measurable objectives and accountability mechanisms. The integration of standards such as ISO/IEC 42001:2023, which establishes a management system for AI, and ISO/IEC 38507:2022, which outlines board-level responsibilities, provides organizations with structured pathways for institutionalizing these commitments.

The literature underscores that strategic guardrails must balance innovation with governance. Enterprises face ongoing tensions between leveraging AI for competitive advantage and ensuring responsible adoption. Without strong strategic direction, operational safeguards risk becoming fragmented or symbolic. By situating responsible AI within corporate governance and enterprise risk management, strategic guardrails create the conditions under which technical and managerial measures can function effectively. This top-level alignment is therefore indispensable to ensuring that responsible AI becomes a core element of organizational strategy rather than a peripheral concern.

4.3 Operational Guardrails

The operational layer translates strategic commitments into concrete processes, tools, and practices that can be embedded into the AI lifecycle. Whereas strategic guardrails emphasize

compliance and governance at the institutional level, operational guardrails function within the day-to-day workflows of AI development and deployment. Their purpose is to ensure that ethical principles such as fairness, accountability, and transparency are systematically integrated into data practices, model design, validation, and monitoring.

Documentation frameworks are central to operational guardrails. *Model Cards for Model Reporting* (Mitchell et al., 2019) and *Datasheets for Datasets* (Gebru et al., 2021) provide standardized mechanisms to disclose dataset provenance, model assumptions, and intended use cases. These practices improve transparency by offering stakeholders clear information about the conditions under which AI systems are designed to function. Beyond documentation, auditing has emerged as a critical practice. Raji and Buolamwini (2019) show how external audits exposed racial and gender biases in commercial facial recognition systems, while Raji et al. (2020) propose internal audit frameworks that embed accountability checks throughout the AI pipeline. Together, these tools demonstrate how operational guardrails move ethical commitments from aspirational statements to enforceable practices.

Risk management standards further reinforce the operational layer. The NIST AI Risk Management Framework (2023) provides organizations with structured guidance organized into four functions—govern, map, measure, and manage—that can be applied across the AI lifecycle. The ISO/IEC 23894:2023 guidance on AI risk management offers a complementary perspective by aligning risk controls with established enterprise risk management processes. Both frameworks emphasize the importance of iterative assessment, bias mitigation, and incident response as ongoing operational responsibilities.

Despite their promise, operational guardrails face notable challenges in practice. Organizations often treat documentation and auditing as compliance checklists rather than integrating them into iterative development processes (Morley et al., 2021). Furthermore, technical tools alone cannot resolve questions of accountability if they are not supported by managerial oversight and resource allocation. As a result, operational guardrails are most effective when situated within a broader framework that connects them to both strategic governance commitments and managerial accountability structures. This interdependence reinforces the layered nature of responsible AI, where operational measures provide substance but require alignment across all levels of the enterprise to achieve sustained impact.

4.4 Managerial Guardrails

The managerial layer addresses the critical role of leadership, accountability, and organizational integration in ensuring that responsible AI is more than a technical or regulatory exercise. While strategic guardrails establish the direction and operational guardrails provide the tools, managerial guardrails determine whether these measures are consistently applied, resourced, and enforced across the enterprise. Without this layer, responsible AI risks remaining aspirational, with commitments articulated at the strategic level but insufficiently embedded into organizational practice.

The literature on governance emphasizes that managerial guardrails must clarify accountability for AI-related risks. Recent studies emphasize that program leaders and managers act as ethical brokers who translate high-level principles into concrete decisions about scope, risk tolerance, and acceptable trade-offs in product development (*ISO/IEC 38507:2022*; *Mökander & Floridi, 2021*; *Stahl et al., 2021*). *ISO/IEC 38507:2022* reinforces this view by positioning boards and senior management as responsible for aligning AI use with organizational values and risk appetite. Yet studies show that accountability remains diffuse in practice, with oversight often fragmented across compliance, technical, and business units (*Stahl et al., 2021*).

Managerial guardrails also involve establishing organizational structures that enable cross-functional oversight. Enterprises increasingly form AI ethics boards, risk committees, or working groups to evaluate high-impact use cases. However, the effectiveness of these bodies depends on their authority and integration into decision-making processes. Without sufficient resources or influence, they risk becoming symbolic rather than substantive. Morley et al. (2021) note that cultural and organizational barriers—including limited ethical literacy among managers and the prioritization of short-term performance over long-term trust—can further undermine these structures.

Finally, managerial guardrails must ensure that operational tools are not treated as isolated checklists but are embedded into enterprise-wide processes. This includes integrating auditing practices into product release cycles, linking documentation frameworks to procurement and vendor management, and aligning risk assessments with broader enterprise risk management systems. By doing so, managers create the connective tissue that binds strategic commitments to operational practices. This integrated perspective addresses the accountability gap highlighted in earlier studies (Raji et al., 2020) and positions responsible AI as a core component of organizational governance. Table 3 summarizes the proposed framework, consolidating strategic, operational, and managerial guardrails, along with their corresponding instruments and managerial actions.

Table 3Summary of the Proposed Three-Layer Framework for Responsible AI Guardrails.

Layer	Description	Key Instruments / References	Managerial Actions
Strategic Guardrails	Align AI adoption with external regulations and internal ethical commitments at the enterprise level. Provides board-level direction and ensures long-term legitimacy.	EU AI Act (European Commission, 2024); OECD Recommendation (2019); NIST AI RMF (2023); ISO/IEC 42001:2023; ISO/IEC 38507:2022	Translate external regulations into corporate policies; establish ethics charters; integrate responsible AI into enterprise risk management and board oversight.

Operational Guardrails	Translate principles into concrete practices across the AI lifecycle, including documentation, auditing, and risk management.	Model (Mitchell et al., 2019); Datasheets (Gebru et al., 2021); Algorithmic Audits (Raji & Buolamwini, 2019; Raji et al., 2020); NIST (2023); ISO/IEC 23894:2023	Cards Embed documentation in workflows; conduct internal and external audits; implement risk controls; link tools to product release cycles and vendor oversight.
Managerial Guardrails	Ensure accountability, leadership engagement, and integration of responsible AI into decision-making structures. Bridge the accountability gap between principles and practice.	Stahl et al. (2021); Morley et al. (2021); Mökander & Floridi (2021); ISO/IEC 38507:2022; Raji et al. (2020)	Define accountability for AI risks; empower cross-functional ethics boards; allocate resources for governance; align AI oversight with enterprise-wide risk systems.

5. Discussion and Managerial Implications

The proposed three-layer framework highlights the necessity of integrating strategic, operational, and managerial guardrails to advance responsible AI adoption in enterprises. While ethical principles and technical tools are essential, the review demonstrates that they remain insufficient unless connected through governance mechanisms that reflect organizational realities. This section discusses the broader implications of the framework and outlines how managers can operationalize it to balance innovation with accountability and trust.

A first implication concerns the alignment of regulatory and strategic commitments. The growing regulatory landscape, exemplified by the European Union's AI Act (European Commission, 2024) and the OECD Recommendation on AI (OECD, 2019), establishes clear expectations that enterprises cannot overlook. Strategic guardrails ensure that organizations translate these external requirements into internal governance commitments that provide direction for product development, risk management, and compliance. For managers, this underscores the importance of maintaining visibility into evolving regulatory requirements and positioning responsible AI as a corporate priority rather than a technical afterthought.

A second implication lies in the institutionalization of operational tools. While practices such as model documentation (Mitchell et al., 2019; Gebru et al., 2021) and auditing frameworks (Raji et al., 2020) represent concrete advances, their effectiveness depends on how they are integrated into enterprise workflows. Managers play a central role in ensuring that these practices are embedded into development pipelines, procurement processes, and vendor oversight. By allocating resources, defining accountability, and setting performance metrics,

managers can move operational tools from isolated compliance exercises to meaningful governance mechanisms.

A third implication is the recognition of managerial accountability as the linchpin of responsible AI. Prior studies emphasize that governance failures often arise not from the absence of principles or tools, but from unclear lines of responsibility (ISO/IEC 38507:2022; Morley et al., 2021; Stahl et al., 2021). Managerial guardrails clarify decision rights and ensure that AI-related risks are addressed with the same rigor as financial, legal, or reputational risks. This requires cultivating ethical literacy among managers, empowering cross-functional governance bodies, and aligning AI oversight with enterprise risk management systems.

Finally, the framework points to the importance of balancing innovation with trust. Enterprises face competitive pressures to accelerate AI deployment, but short-term gains achieved at the expense of ethical practices risk long-term reputational damage and regulatory penalties. Embedding guardrails at all three layers makes responsible AI something leaders can manage, not just endorse. That shift matters in practice: teams ship under deadlines, and guardrails have to survive that reality.

This framework also directly addresses the four research questions posed in Section 2.1. Specifically, RQ1 is addressed by consolidating the diverse range of proposed guardrails; RQ2 is answered through the synthesis of operational tools and practices; RQ3 is resolved by identifying and analyzing managerial accountability gaps; and RQ4 is fulfilled by integrating these insights into the proposed three-layer framework. Together, these responses reinforce the framework's validity and highlight its contributions to both scholarly discourse and managerial practice.

In short, the framework translates convergent principles and emerging tools into a governance pathway managers can actually run, review, and improve over time.

6. Conclusion

This paper has examined the evolving landscape of responsible AI and highlighted the need for enterprise-level frameworks that move beyond principles and technical tools to encompass managerial governance. The structured literature review revealed that while ethical principles have been widely articulated (Floridi & Cowls, 2019; OECD, 2019; European Commission, 2019), and operational tools such as model documentation and auditing practices provide tangible pathways for implementation (Mitchell et al., 2019; Gebru et al., 2021; Raji et al., 2020), there remains a persistent accountability gap. Specifically, the literature offers limited guidance on how managers can institutionalize responsible AI within organizational structures, decision-making processes, and governance systems.

To address this gap, the paper proposed a three-layer framework of strategic, operational, and managerial guardrails. The strategic layer ensures alignment with external regulations and internal ethical commitments, the operational layer embeds technical practices and risk management processes, and the managerial layer clarifies accountability and integrates responsible AI into enterprise governance. By synthesizing insights from scholarship,

standards, and policy, the framework provides a structured yet flexible model for enterprises seeking to balance innovation with governance and trust.

The contribution of this study is both scholarly and practical. For researchers, it consolidates diverse streams of literature into a coherent framework that foregrounds the managerial dimension of responsible AI, an area that has been comparatively underexplored. For practitioners, it offers a roadmap for embedding responsible AI across organizational levels, thereby addressing regulatory pressures while strengthening trust and legitimacy.

At the same time, the study acknowledges limitations. As a conceptual paper, it does not provide empirical validation of the framework, and further research is needed to test its applicability in different organizational and industry contexts. Future work could involve case studies, comparative analyses, or surveys that examine how enterprises operationalize the proposed guardrails in practice. Such studies would enrich the literature by providing evidence of what works, under what conditions, and with what outcomes.

In conclusion, responsible AI requires more than principles and tools; it demands governance structures that connect ethics to practice through managerial accountability. By offering a layered framework that integrates strategic, operational, and managerial perspectives, this paper contributes to bridging the accountability gap and advancing both the theory and practice of responsible AI in enterprises.

6.1 Scholarly and Practical Significance

The proposed framework offers a testable basis for future research and a practical guide for executive education, PMO design, and enterprise governance. By clarifying the “unclear mandate” of middle managers, the study provides actionable levers for leaders seeking to sustain agility beyond initial adoption.

Funding

This research received no external funding.

Institutional	Review	Board	Statement
Not applicable. This study is a literature review and did not involve human subjects.			

Informed	Consent	Statement
Not applicable. This study did not involve human participants.		

Data	Availability	Statement
No new data were created or analyzed in this study. Data sharing is not applicable to this article.		

Conflicts	of	Interest
The author declares no conflict of interest.		

Use	of	AI	Tools
Digital tools were used to support language editing and formatting. All analysis, synthesis, and conclusions are the sole responsibility of the author.			

References

1. European Commission. (2019). *Ethics guidelines for trustworthy AI*. High-Level Expert Group on Artificial Intelligence. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
2. European Commission. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)*. *Official Journal of the European Union*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
3. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Philosophy & Technology*, 32(4), 687–703. <https://doi.org/10.1007/s13347-019-00347-0>
4. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
5. International Organization for Standardization. (2022). *ISO/IEC 38507:2022—Information technology—Governance of IT—Governance implications of the use of artificial intelligence by organizations*. <https://www.iso.org/standard/56641.html>
6. International Organization for Standardization. (2023a). *ISO/IEC 23894:2023—Information technology—Artificial intelligence—Guidance on risk management*. <https://www.iso.org/standard/77304.html>
7. International Organization for Standardization. (2023b). *ISO/IEC 42001:2023—Information technology—Artificial intelligence—Management system*. <https://www.iso.org/standard/81230.html>
8. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
9. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>
10. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
11. Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, 31(2), 323–327. <https://doi.org/10.1007/s11023-021-09557-8>
12. Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: An overview of AI ethics tools, methods and research to translate principles into practices. *AI & Society*, 36(1), 59–82. <https://doi.org/10.1007/s00146-020-00992-2>
13. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST Special Publication 1270). <https://doi.org/10.6028/NIST.SP.1270>

14. OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
15. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
16. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429–435). Association for Computing Machinery. <https://doi.org/10.1145/3306618.3314244>
17. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372873>
18. 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>
19. Stahl, B. C., Markus, M. L., & Mittelstadt, B. (2021). The ethics of ethics of AI: An evaluation of guidelines. *Minds and Machines*, 31(2), 239–265. <https://doi.org/10.1007/s11023-020-09517-8>
20. 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>