STIESIA
ICOBUSS
e-ISSN 2746-5667
5th INTERNATIONAL CONFERENCE ON BUSINESS & SOCIAL SCIENCES
RESPONSIBLE AI AND HUMAN-CENTERED DESIGN: BRIDGING ETHICS AND INNOVATION

OPERATIONALIZING RESPONSIBLE AI PRINCIPLES THROUGH A GLOBAL ACCOUNTABILITY FRAMEWORK MODEL

Hairul Hatami[1*], Rizky Salsabila Fitri [2], Achmad Fauziannor [3] Meiske Claudia[4]

[1] Faculty of Economic and Business, Lambung Mangkurat University, Banjarmasin, Indonesia
[2] Faculty of Economic and Business, Lambung Mangkurat University, Banjarmasin, Indonesia
[3] Faculty of Economic and Business, Lambung Mangkurat University, Banjarmasin, Indonesia
[4] Faculty of Economic and Business, Lambung Mangkurat University, Banjarmasin, Indonesia

*Corresponding Author: zuliahatami@gmail.com

Abstract: The rapid expansion of artificial intelligence (AI) has intensified concerns about its ethical, social, and governance implications. While numerous frameworks and guidelines for Responsible AI have been issued globally, a gap persists between high-level principles and their operationalization. This study addresses this gap by developing an accountability framework through a systematic literature review (SLR) and validating it with a case study of IBM. The SLR analyzed 78 scholarly and policy sources, producing five thematic domains: values, governance instruments, metrics, barriers, and pathways, with accountability emerging as the integrative theme. To test this framework, the IBM case was examined across eight key documents, including the shortcomings of Watson Health and the company's subsequent development of governance tools such as FactSheets 360, fairness toolkits, and counterfactual testing. Findings show that while IBM's commitments align with global principles, accountability remains most vulnerable in metrics and real-world validation. Nevertheless, IBM's innovations and participation in global standardization illustrate how accountability can be reinforced in practice. The study contributes theoretically by reframing accountability as the linchpin of Responsible AI, and practically by offering pathways for organizations and policymakers to strengthen AI governance.

Keywords: Responsible AI, Accountability, Governance Framework, Systematic Literature Review, Case Study

1. Introduction

The rapid diffusion of artificial intelligence (AI) across sectors has generated both optimism and concern. AI is increasingly deployed in healthcare, finance, public administration, and education to enhance decision-making, efficiency, and innovation. Yet its growing ubiquity raises profound challenges related to fairness, transparency, privacy, trust, and accountability. International organizations and governments have responded by issuing guidelines and policy frameworks to ensure the development and deployment of AI remains aligned with human values. The OECD Recommendation on AI (2019) stands as one of the most influential, articulating five foundational principles inclusive growth, human-centered values, transparency, robustness and safety, and accountability which provide a global baseline for Responsible AI. Building on this, the NIST AI Risk Management Framework (AI RMF 1.0) (2023a) operationalizes trustworthy AI through a structured cycle of Govern, Map, Measure, and Manage, embedding essential characteristics such as validity, safety, explainability, fairness, and accountability within the AI lifecycle.

Despite the proliferation of such principles and frameworks, a persistent gap remains between ethical aspirations and practical implementation. Corrêa et al., (2023) show that while more than 200 AI ethics guidelines exist globally, their real-world application is fragmented and inconsistent across sectors. Veale et al., (2023) emphasize that global AI governance suffers from jurisdictional divides and regulatory fragmentation, undermining harmonization. Gianni

et al., (2022) describe the prevalence of "ethics-washing," in which organizations adopt ethical commitments symbolically without embedding them into operational processes. Roberts et al., (2024) and Birkstedt et al., (2023) further highlight knowledge gaps and barriers that limit accountability in AI governance. Together, these insights suggest that principles alone are insufficient unless translated into enforceable mechanisms, measurable metrics, and context-sensitive practices.

Within this landscape, accountability emerges as the linchpin. Novelli, Taddeo, and Floridi (2024) argue that accountability ensures not only the assignment of responsibility but also answerability when AI systems cause harm. Yet accountability is also the principle most prone to failure, often undermined by weak validation mechanisms, insufficient fairness metrics, and the absence of redressability. The case of IBM Watson Health illustrates these shortcomings vividly: launched with ambitious promises of personalized, evidence-based medical support, Watson struggled with reliance on curated datasets, lack of clinical validation, poor integration into medical workflows, and high costs, ultimately eroding trust among healthcare professionals and limiting adoption (Insights, 2023). This case underscores the difficulty of operationalizing accountability and demonstrates that even leading organizations face profound challenges in translating principles into practice.

Against this backdrop, this study aims to develop an accountability framework that bridges Responsible AI principles with operational mechanisms. Using a systematic literature review (SLR) of 78 scholarly and policy sources, the research synthesizes thematic insights into values, governance instruments, metrics, barriers, and pathways. This framework is subsequently validated through a case study of IBM, an organization that has both faced accountability failures (e.g., Watson Health) and pioneered innovative governance tools such as AI FactSheets 360, fairness toolkits, and counterfactual testing for bias mitigation. By combining literature synthesis with case analysis, this study seeks to advance both theoretical understanding and practical pathways for embedding accountability in AI governance.

The remainder of this paper is structured as follows. Section 2 reviews prior scholarship and frameworks on Responsible AI and accountability. Section 3 outlines the methodology, including the systematic review and case study design. Section 4 presents the results and discussion, integrating insights from the literature with empirical evidence from the IBM case. Section 5 concludes by summarizing the key contributions, implications, and limitations of the study.

2. Literature Review

The rise of artificial intelligence (AI) has accelerated scholarly and policy discussions about its ethical, social, and governance implications. As AI systems increasingly permeate critical domains such as healthcare, finance, education, and public services, the risks associated with bias, opacity, privacy violations, and safety lapses have become urgent concerns. To address these challenges, international and national bodies have developed principles and frameworks under the umbrella of Responsible AI. The OECD Principles on AI (2019) are widely regarded as a global benchmark, highlighting five commitments: inclusive growth and sustainable development; human-centered values and fairness; transparency and explainability; robustness, security, and safety; and accountability. These principles emphasize that AI should benefit people and the planet, operate with integrity, and include mechanisms of accountability to mitigate risks. In parallel, the NIST AI Risk Management Framework (AI RMF 1.0) (2023) provides an operational model structured around four functions Govern, Map, Measure, and Manage and identifies key trustworthiness characteristics such as validity, safety,

explainability, privacy, fairness, and accountability. In Europe, the EU AI Act (2024) introduces the first comprehensive regulatory framework, adopting a risk-based classification of AI systems and legally binding requirements, positioning accountability as both a legal and ethical imperative (Veale et al., 2023).

By contrast, initiatives in Asia and Southeast Asia are still evolving and often take the form of voluntary frameworks. The ASEAN Guide on AI Governance and Ethics (2020) promotes transparency, fairness, accountability, and human-centric values, providing a set of high-level principles but lacking enforceable obligations. Among ASEAN members, Singapore has emerged as a regional leader, introducing the Model AI Governance Framework in 2019, updated in 2020, and most recently extended in 2024 to cover generative AI. The 2024 framework introduces nine dimensions including accountability, content provenance, incident reporting, testing and assurance, security, safety research, and AI for public good and explicitly references the EU AI Act as a global benchmark for more stringent regulation (Infocomm Media Development Authority (IMDA), 2024). In Indonesia, the Strategi Nasional Kecerdasan Artifisial (Stranas KA 2020–2045) outlines a roadmap for AI development in alignment with ethical principles rooted in Pancasila and the G20 AI Principles. Stranas identifies five national priority sectors health, education and research, food security, smart mobility, and bureaucracy reform supported by enablers such as talent development, infrastructure, and governance (Kementerian Riset dan Teknologi / BRIN, 2020). These regional and national efforts reflect growing awareness of Responsible AI, but they remain less prescriptive than European models and face challenges of implementation across diverse political and regulatory contexts.

Within these frameworks, the literature converges on several core principles of Responsible AI that have gained near-universal recognition. Transparency is consistently emphasized as necessary for enabling users, regulators, and affected communities to understand how AI systems function and make decisions (Felzmann et al., 2019). Fairness or non-discrimination is central to preventing biased outcomes and promoting equity across demographic groups (Memarian & Doleck, 2023). Explainability further complements transparency by ensuring that AI systems provide clear, comprehensible justifications for their outputs, enabling accountability by design (Kim et al., 2020). Human oversight is widely endorsed as a safeguard against over-reliance on automated systems, ensuring that human judgment remains central in high-stakes decisions (Rozenblit et al., 2025). In addition, privacy and data security are foundational in protecting individuals' rights and fostering trust in AI applications (Saeidnia et al., 2024). Finally, monitoring and audit mechanisms are increasingly recognized as necessary for ensuring that AI systems remain compliant, reliable, and contestable throughout their lifecycle (Mökander & Floridi, 2023).

Although these principles are broadly endorsed, scholars point out that they frequently remain at the normative level rather than being translated into operational practices. Corrêa et al. (2023) demonstrate that while more than 200 AI ethics guidelines exist worldwide, their application is fragmented and inconsistent, varying significantly by jurisdiction and industry. Novelli, Taddeo, and Floridi (2024) highlight that accountability although universally acknowledged is often underdeveloped in practice, with organizations lacking mechanisms to ensure genuine answerability for AI outcomes. Gianni, Lehtinen, and Nieminen (2022) introduce the concept of "ethics-washing," in which institutions adopt ethical language to signal responsibility while failing to embed accountability into governance systems. Roberts et al., (2024) and Birkstedt et al., (2023) further argue that regulatory fragmentation and knowledge gaps persist, leaving accountability vulnerable to being eroded in real-world implementation. Xia et al., (2024) add that the absence of standardized accountability metrics

limits the capacity of organizations to measure, monitor, and demonstrate compliance with ethical norms.

Beyond principles, recent studies have examined governance instruments designed to operationalize Responsible AI. These include regulatory standards (such as the EU AI Act), organizational policies, and technical tools. IBM's development of AI FactSheets 360, AI Fairness 360, and counterfactual testing exemplify efforts to embed accountability into AI development processes (OECD, 2022). Similarly, Mökander & Floridi, (2023) emphasize ethics-based auditing as a promising instrument to translate abstract principles into actionable accountability mechanisms. However, as many scholars argue, such tools are not yet widely adopted, and their effectiveness remains limited by context-specific implementation challenges (Li & Chignell, 2022).

The barriers to operationalizing Responsible AI are also well-documented. Key obstacles include limited technical expertise and talent shortages (Corrêa et al., 2023), resistance within organizations to invest in governance infrastructure (Gianni et al., 2022), and the costs of implementing fairness or privacy-preserving techniques (Roberts et al., 2024). Regulatory gaps and fragmentation further complicate accountability, as varying jurisdictions impose inconsistent requirements, leading to uneven adoption (Veale et al., 2023b). Moreover, cultural and geopolitical differences shape how accountability is understood and prioritized, with Western frameworks often emphasizing individual rights and litigation, while Asian contexts may focus more on collective welfare and voluntary compliance (Ayana et al., 2024).

In Southeast Asia, these challenges are particularly acute given the region's diverse political and economic contexts. The ASEAN Guide (2020) reflects a commitment to harmonization, but its non-binding nature limits enforcement. Singapore's leadership provides a model for operationalization, yet its advanced infrastructure and regulatory capacity may not be easily replicable in neighboring states. Indonesia's Stranas KA (2020–2045) demonstrates ambition, but implementation remains in early stages, constrained by resource limitations and the absence of binding regulation. These contrasts underscore the need for frameworks that can bridge normative aspirations with practical realities, tailored to regional contexts yet informed by global best practices.

Taken together, the literature establishes a paradox: accountability is simultaneously the most cited and the most fragile principle of Responsible AI. Frameworks at global, regional, and national levels emphasize transparency, fairness, explainability, human oversight, privacy, and auditing, but their operationalization remains weak. Real-world cases such as IBM Watson Health, which failed due to inadequate validation, biased datasets, and weak integration into workflows, exemplify the consequences of insufficient accountability mechanisms (Insights, 2023b). This unresolved tension between normative principles and practical implementation defines the research gap addressed in this study. Our research responds to this gap by developing an accountability framework that integrates values, governance instruments, metrics, and barriers, and by validating this framework through the IBM case study, thereby advancing both theoretical and practical understanding of how accountability can be institutionalized in AI governance.

3. Method

This study adopts a qualitative research design that combines a Systematic Literature Review (SLR) with a case study. The integration of these methods allows both the identification of thematic patterns across academic and policy discourse, and the validation of these themes against empirical evidence from an organizational context (Snyder, 2019; Yin, 2017).

Systematic Literature Review (SLR)

The SLR followed the standard protocol of identification, screening, eligibility, and inclusion (Moher et al., 2009). Searches were conducted in databases such as Scopus, Web of Science, and IEEE Xplore using combinations of keywords: Responsible AI, AI governance, accountability, transparency, fairness, explainability, and ethics. The initial search yielded 162 records, which were then subjected to duplicate removal and relevance screening. Studies were included if they (a) explicitly addressed principles or frameworks of Responsible AI, (b) discussed governance mechanisms or accountability practices, and (c) were published in peer-reviewed journals, policy reports, or reputable white papers. Exclusion criteria applied to purely technical studies without governance focus, editorials, or non-English publications. After the full-text review, 78 articles were retained for analysis.

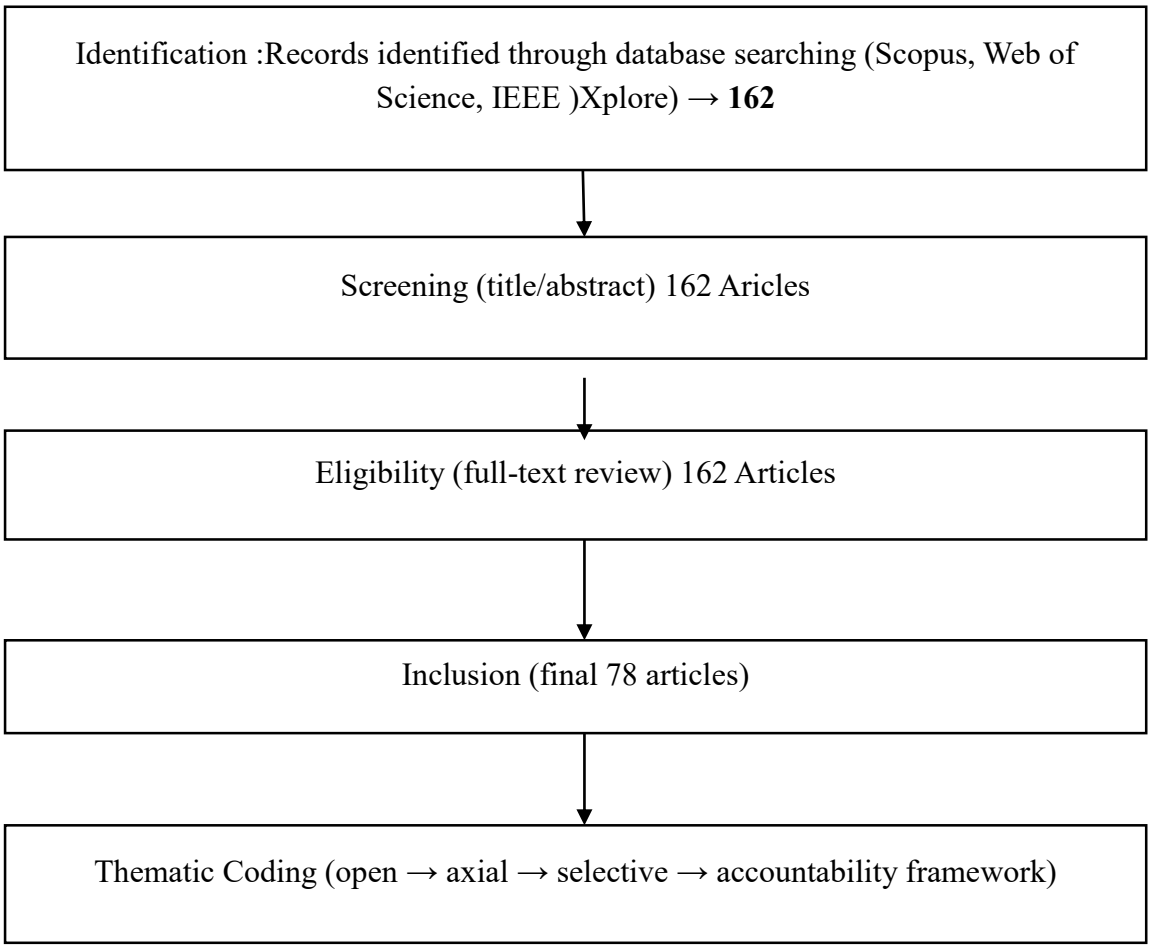

Figure 1: PRISMA Flow Diagram of the SLR Process

To ensure rigor, all selected articles were subjected to quality appraisal. The assessment classified articles based on journal quartiles (Q1–Q3), indexed but non-quartiled journals, and non-indexed sources, following established approaches to quality evaluation in SLRs (Tranfield et al., 2003). The majority of the literature originated from high-quality outlets, with Q1 journals comprising nearly half of the dataset, followed by Q2 and other indexed sources. This quality distribution underscores the robustness of the knowledge base while also including non-indexed but policy-relevant reports for practical insights.

After the final inclusion step, the entire review process was documented in a flow diagram to enhance transparency and replicability (see Figure 1). The diagram summarizes the progression from the initial identification of 162 records through screening and eligibility checks, leading to the final inclusion of 78 articles. This systematic visualization clarifies the logic of the review and highlights how the dataset was refined for analysis.

To further ensure rigor, the included articles were subjected to a quality appraisal. Articles were categorized based on journal quartiles (Q1–Q3), indexed but non-quartiled sources, and non-indexed yet policy-relevant publications. The distribution of article quality is presented in a bar chart (see Figure 2), which shows that Q1 journals account for the largest share, underscoring the robustness of the evidence base while also incorporating diverse perspectives from other sources.

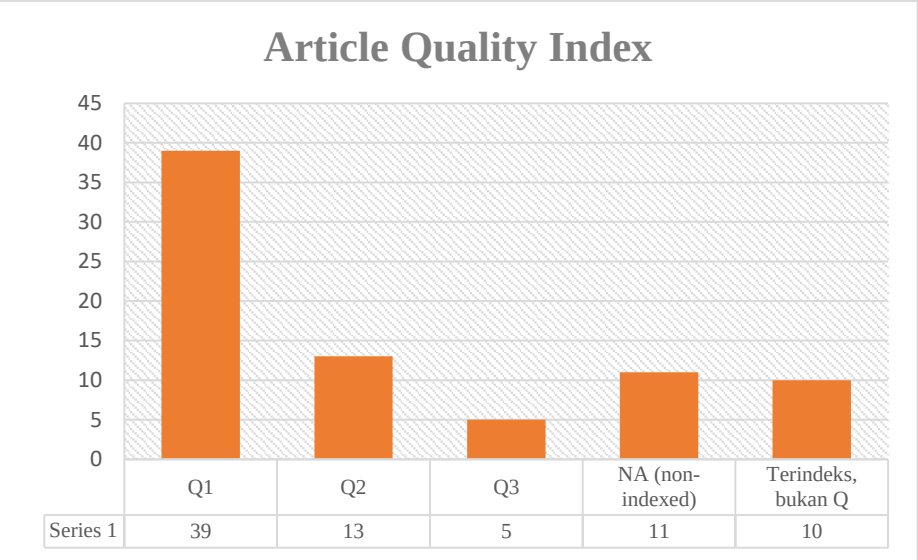

Figure 2: Article Quality Index of the Selected Studies

Data Analysis (SLR)

The review employed a thematic analysis approach consistent with Braun and Clarke (2006). First, open coding was conducted to extract key ideas from the corpus of 78 articles. Second, axial coding clustered related codes into broader categories (Corbin & Strauss, 2015). Third, a cross-article analysis was performed to identify convergence and divergence across studies. Finally, selective coding consolidated the categories into core themes, revealing accountability as the integrative principle linking values, governance instruments, metrics, barriers, and pathways.

Case Study

To refine and validate the accountability framework, a case study of IBM was conducted. IBM was chosen for two reasons. First, it has been a global frontrunner in advancing Responsible AI through tools such as AI FactSheets 360, AI Fairness 360, and counterfactual testing for bias mitigation (IBM Research, 2023; OECD, 2022). Second, it has also experienced widely recognized accountability challenges, most notably with IBM Watson Health, which revealed the risks of weak validation and operational failures (Insights, 2023).

Eight documents were analyzed, including IBM white papers, technical frameworks, policy contributions, and independent evaluations of Watson Health. These were subjected to the same thematic coding process as the SLR, enabling a systematic comparison between normative frameworks and organizational practice. Triangulation across sources enhanced the validity of

findings and provided insights into both successes and shortcomings in operationalizing accountability (Yin, 2017).

4. Result and Discussion

Introduction to the Synthesis

The SLR produced a rich body of insights from 78 scholarly and policy sources, allowing us to identify recurring themes and conceptual tensions in the discourse on Responsible AI. Across diverse contexts, five broad domains consistently emerged: values, governance instruments, metrics, barriers, and pathways. These domains, while analytically distinct, interact to shape how AI governance is imagined and implemented. At the center of this synthesis lies accountability, which repeatedly surfaced as both the foundation of Responsible AI and the principle most prone to failure.

A distinguishing feature of this review is its integration of perspectives across global, regional, and national frameworks. International principles such as those from the OECD (2019a) and NIST (2023) are complemented by regional initiatives like the EU AI Act (2024), ASEAN Guide on AI Governance (2020), and Singapore's updated Model AI Governance Framework (2024). National strategies, such as Indonesia's *Stranas KA 2020–2045*, further localize these discourses. By synthesizing these layers, the SLR provides a holistic understanding of how Responsible AI is articulated across contexts.

The subsequent sections unpack each domain in detail, showing not only points of consensus but also gaps and contradictions. Together, they set the stage for the construction of a conceptual framework of accountability, which will later be validated against empirical evidence from IBM.

Core Responsible AI Values

Responsible AI is anchored in normative values that guide governance and practice. The most prominent are transparency, fairness, explainability, privacy, inclusivity, and accountability. Transparency enables external scrutiny and trust (Felzmann et al., 2019), while fairness ensures equitable outcomes across social groups (Memarian & Doleck, 2023). Explainability complements transparency by providing human-understandable rationales for AI decisions (Kim et al., 2020). Privacy and data security protect individual rights and maintain legitimacy, especially in sensitive domains (Saeidnia et al., 2024). Inclusivity emphasizes the equitable distribution of AI's benefits across diverse populations. Accountability functions as the meta-value, ensuring that these principles are enforceable and actionable (Novelli et al., 2024).

However, the literature reveals variation in emphasis. Western frameworks, such as the EU AI Act, link transparency and accountability directly to liability and compliance (Veale et al., 2023a). By contrast, Asian frameworks prioritize inclusivity and collective welfare, framing accountability more as a voluntary commitment (ASEAN, 2020). This divergence reflects broader socio-political differences while affirming global consensus on the importance of these values.

Importantly, many studies caution that without accountability, these values risk remaining symbolic aspirations. Ethical guidelines are often published by governments and corporations but lack mechanisms for enforcement, leading to accusations of ethics-washing (Gianni et al., 2022). This gap underscores why accountability must be treated as more than a principle—it must be operationalized.

Governance Instruments

Governance instruments are the mechanisms through which values are translated into practice. The SLR highlights three categories: regulatory frameworks, auditing mechanisms, and technical tools. The EU AI Act (2024) exemplifies a legally binding approach, imposing strict requirements on high-risk AI systems. ASEAN's Guide (2020) and Singapore's framework (Infocomm Media Development Authority (IMDA), 2024) represent softer approaches, emphasizing best practices and industry guidance rather than enforceable law.

Auditing mechanisms are increasingly advocated to provide independent oversight. Ethics-based auditing, for example, systematically evaluates AI systems for compliance with ethical principles (Mökander & Floridi, 2023). However, their adoption is uneven, with more advanced economies leading in implementation while resource-limited contexts lag behind.

Technical tools bridge the gap between policy and practice. IBM's open-source AI Fairness 360 and AI FactSheets 360 exemplify instruments that embed accountability within development processes (IBM, 2023) These tools document datasets, monitor performance, and assess bias, providing practical pathways for accountability. Yet, as the literature notes, tools alone are insufficient without organizational structures and cultural commitment to governance (OECD, 2022).

Measurement and Validation

Measurement and validation emerged as the weakest link in Responsible AI. While there is broad consensus on the importance of metrics, no standardized accountability indicators exist. Bias audits and fairness tests are widely used, but they vary across domains and lack comparability (Xia et al., 2024). Explainability metrics provide partial insights, but technical heterogeneity makes standardization difficult (Kim et al., 2020).

Validation practices are also inconsistent. Few AI systems undergo independent third-party validation or long-term monitoring. This absence undermines accountability, as ethical commitments cannot be verified without evidence. Watson Health exemplifies this problem: despite IBM's strong ethical declarations, reliance on curated datasets and lack of clinical validation eroded its credibility (Insights, 2023d).

Some scholars advocate for universal accountability benchmarks to ensure comparability across industries (Roberts et al., 2024). Others caution that rigid metrics risk overlooking contextual differences and unintended consequences (Birkstedt et al., 2023). Regardless, the literature agrees that without measurable indicators, accountability remains rhetorical rather than operational.

Barriers and Constraints

Barriers to Responsible AI span organizational, systemic, and socio-political dimensions. At the organizational level, lack of technical expertise, resource constraints, and resistance to investing in governance infrastructure are common (Corrêa et al., 2023). Smaller firms in particular struggle to allocate resources to accountability practices.

Systemic barriers include regulatory fragmentation, with jurisdictions adopting divergent standards that hinder harmonization (Veale et al., 2023). Socio-political contexts also shape adoption. In Western contexts, accountability is tied to compliance and litigation, while in Asian contexts it is often framed as voluntary or collective responsibility (ASEAN, 2020).

Cultural inertia and ethics-washing further exacerbate these challenges. Gianni, Lehtinen, and Nieminen (2022) describe how organizations often adopt ethical principles rhetorically without substantive implementation. These constraints suggest that accountability cannot be achieved through technical measures alone it requires structural and cultural transformation.

Pathways Forward

Despite these challenges, the literature identifies several pathways forward. Global harmonization is frequently proposed, with OECD (2019) and G20 AI Principles serving as reference points. Ethics-based auditing offers a structured approach to embedding accountability within organizations (Mökander & Floridi, 2023).

Human–AI collaboration is highlighted as a practical balance between automation and oversight, ensuring that humans retain control in high-stakes contexts (Rozenblit et al., 2025). Stakeholder engagement and inclusivity are emphasized to ensure accountability mechanisms reflect societal values and build trust (Ayana et al., 2024).

Finally, the integration of sustainability indicators into accountability frameworks is an emerging trend. The Standardization Summit (2025) calls for metrics on energy consumption, carbon emissions, and lifecycle impacts, expanding accountability to encompass environmental stewardship. This broadens the scope of Responsible AI beyond ethics to sustainability.

Integrative Core: Accountability

Synthesizing these domains reveals accountability as the integrative core. It links values, instruments, and metrics while being undermined by barriers and reinforced by pathways. Accountability is simultaneously the most cited principle and the most fragile in practice. Without it, other principles risk being symbolic; with it, they can be operationalized and enforced (Novelli et al., 2024).

This paradox reflects the central research gap: accountability is universally recognized yet inconsistently implemented. The findings thus highlight the need for frameworks that not only articulate accountability but also provide mechanisms for its measurement and enforcement.

Conceptual Framework and Transition to Case Analysis

The findings of the SLR are consolidated in a conceptual framework that illustrates the dynamics of Responsible AI governance. The framework positions accountability as the central node, linking values, governance instruments, and metrics, while showing how barriers weaken and pathways strengthen accountability.

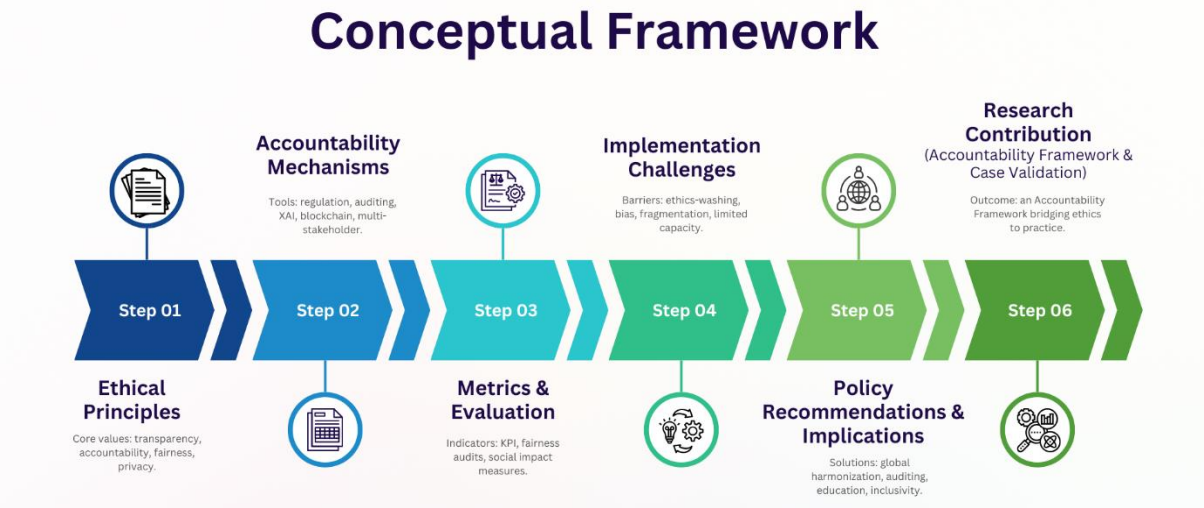

Figure 3: Conceptual Framework for Operationalizing Responsible AI through Accountability

This framework provides a structured lens for understanding AI governance. However, its robustness must be tested against empirical realities. Therefore, the next stage of this study involves a case analysis of IBM, chosen for its dual role as a pioneer in AI governance and as the developer of Watson Health, one of the most prominent accountability failures.

Case Study Analysis: IBM and the Operationalization of Accountability in AI Governance

The case study of IBM provides a concrete illustration of how the principles of Responsible AI are articulated, challenged, and refined in practice. IBM has consistently emphasized trust, transparency, fairness, and accountability as foundational values, framing AI as a tool to augment rather than replace human intelligence (IBM, 2022; Veale et al., 2023). These commitments are evident in its public principles on trust and transparency, as well as in its alignment with global standards such as the OECD AI Principles and the NIST AI Risk Management Framework (Roberts et al., 2024). IBM has also operationalized these values through the development of governance instruments including AI FactSheets 360, AI Fairness 360, Explainability 360, and tools like GYC for counterfactual bias testing, which together reflect a deliberate effort to embed accountability into the design and monitoring of AI systems (IBM, 2023).

Despite these commitments, IBM's trajectory demonstrates the persistent gaps in accountability when principles are translated into real-world applications. The experience of IBM Watson Health highlights the risks of overpromising and underdelivering: reliance on curated datasets, lack of clinical validation, high costs, and poor integration into medical workflows eroded trust among healthcare professionals and limited adoption (Insights, 2023). This failure underscores the limitations of values and governance tools when accountability metrics are insufficiently developed or poorly implemented. At the same time, it illustrates the broader challenges faced by AI governance, including data bias, cost constraints, and regulatory fragmentation, which hinder the consistent operationalization of Responsible AI (Gianni et al., 2022).

IBM's subsequent initiatives indicate an organizational learning process aimed at reinforcing accountability. Through innovations such as FactSheets 360 and the integration of

sustainability considerations into AI governance, IBM has moved towards a more robust and measurable approach to accountability that encompasses not only fairness and transparency but also environmental responsibility (OECD/BIAC, 2022). Its engagement in international standardization efforts further reflects an effort to harmonize practices and contribute to global governance (Birkstedt et al., 2023). Taken together, the IBM case demonstrates both the fragility and necessity of accountability: it is the principle most at risk of failure in practice, yet it is also the linchpin that guides corrective pathways, from technical tools to global frameworks (Mökander & Floridi, 2023).

Summary of Key Findings

The combined findings of the SLR and IBM case study confirm that accountability is the central and most fragile principle of Responsible AI. The SLR identified five domains values, governance instruments, metrics, barriers, and pathways that consistently recur across global, regional, and national frameworks (Corrêa et al., 2023; Roberts et al., 2024). While values such as transparency, fairness, privacy, inclusivity, and explainability are widely emphasized, the analysis shows that they often remain aspirational without accountability mechanisms to enforce them. The IBM case study reinforces this point. On the one hand, IBM's tools such as AI FactSheets 360 and AI Fairness 360 operationalize accountability through documentation and auditing (Research, 2023). On the other hand, the failure of Watson Health demonstrates the risks of neglecting robust validation and integration, leading to erosion of trust among professionals and the public (Insights, 2023b). Together, these insights underscore that accountability is not merely an ethical ideal but a practical necessity, and that its absence can undermine even the most well-intentioned AI initiatives.

Interpretation for the Framework

The conceptual framework developed in this study positions accountability as the binding force connecting values, instruments, metrics, barriers, and pathways. IBM's trajectory both validates and refines this model. The Watson Health case illustrates the vulnerability of accountability when metrics are insufficient bias audits and internal validations were inadequate to meet clinical standards, leaving the system open to criticism (Insights, 2023d). Conversely, the introduction of FactSheets 360 demonstrates how documentation can institutionalize accountability, allowing stakeholders to track datasets, objectives, and model performance over time. Furthermore, IBM's shift towards incorporating sustainability indicators such as energy consumption and carbon emissions highlights how accountability frameworks are expanding beyond ethical and social dimensions to include environmental responsibility (Birkstedt et al., 2023). This suggests that accountability must be understood as a dynamic process: constantly tested by failures, strengthened through organizational learning, and adapted to new governance demands.

Theoretical Implications

Theoretically, this research advances Responsible AI discourse by reframing accountability as the integrative principle rather than a secondary value. Existing studies often privilege transparency or fairness as central (Felzmann et al., 2019; Memarian & Doleck, 2023), yet our synthesis shows that without accountability these principles remain symbolic. The findings also highlight the dual character of accountability: as a social construct tied to answerability, liability, and stakeholder trust, and as a technical property embedded in audits, documentation, and validation (Mökander & Floridi, 2023). By incorporating sustainability into the framework, this study extends theoretical debates to include environmental stewardship, situating

accountability as a holistic principle that integrates ethical, technical, and ecological responsibilities. Moreover, the IBM case illustrates that accountability is iterative, evolving through cycles of failure, critique, and adaptation a view that enriches governance theory by treating accountability as processual and adaptive rather than static.

Practical Implications

For practitioners, the study demonstrates that accountability must be embedded across the AI lifecycle rather than treated as a compliance afterthought. Organizations are encouraged to adopt structured documentation tools like FactSheets to ensure transparency, traceability, and oversight. Independent auditing and third-party validation should complement internal monitoring, preventing reputational damage caused by overpromising, as in Watson Health. Policymakers must move beyond aspirational guidelines to develop enforceable accountability metrics and reporting requirements (Tabassi, 2023). In ASEAN, where frameworks remain advisory, insights from the EU can inspire stronger regulatory designs adapted to regional contexts (ASEAN, 2020; Infocomm Media Development Authority (IMDA), 2024). Civil society and stakeholders also have practical roles in contesting AI decisions, demanding transparency, and ensuring inclusivity in governance processes. Ultimately, accountability should be reframed not merely as a compliance burden, but as a strategic asset for building long-term trust, sustaining innovation, and ensuring legitimacy.

Limitations

Despite its contributions, this study faces limitations. The SLR was confined to 78 articles, and while these were carefully selected and quality-appraised, perspectives from non-English or emerging literature may be underrepresented (Snyder, 2019). The case study was limited to IBM; while instructive, a single case cannot capture the diversity of AI governance practices across industries or regions (Yin, 2017). Reliance on publicly available documents further restricts depth, as internal decision-making processes remain inaccessible. Methodologically, thematic coding introduces interpretive subjectivity, although triangulation between SLR findings and case evidence strengthens reliability (Braun & Clarke, 2006). These limitations point to avenues for future research, including comparative case studies, development of standardized accountability metrics, and longitudinal studies of organizational learning in AI governance

5. Conclusions

This study set out to explore how principles of Responsible AI can be operationalized through an accountability framework, drawing on both a systematic literature review and a case study of IBM. The literature review revealed accountability as the integrative core linking values such as transparency, fairness, and privacy with governance instruments, emerging metrics, and pathways forward, while also highlighting barriers including ethics-washing, regulatory fragmentation, and limited validation mechanisms. The IBM case study validated these insights in practice, demonstrating how accountability commitments can be both undermined and strengthened. While IBM Watson Health exemplified accountability failures marked by data bias, lack of clinical validation, and diminished trust subsequent initiatives such as AI FactSheets 360, fairness toolkits, and sustainability-oriented governance reflect corrective efforts to embed accountability into organizational processes and global standards.

By synthesizing insights from theory and practice, this study advances accountability as not only a normative principle but also a practical and adaptive mechanism for Responsible AI. The proposed framework captures accountability as the linchpin that ensures ethical values are

translated into operational mechanisms, measurable outcomes, and sustainable practices. The case of IBM reinforces the notion that accountability is most vulnerable in metrics and real-world implementation, yet it also demonstrates how organizations can learn, adapt, and contribute to the development of global governance standards.

Ultimately, the findings underscore that accountability is both the greatest challenge and the greatest necessity in Responsible AI. For scholars, the study contributes a multidimensional framework that enriches theoretical debates on AI governance. For practitioners and policymakers, it offers lessons on the risks of neglecting validation and the value of documentation, fairness tools, and global alignment. By situating accountability at the heart of Responsible AI, this research calls for continuous innovation, collaboration, and regulatory support to ensure that AI systems are not only technically advanced but also ethically grounded, socially trustworthy, and environmentally sustainable.

Theoretical and Practical Contributions. Theoretically, this research refines the discourse on Responsible AI by establishing accountability as the integrative principle that unites ethical values, governance instruments, and sustainability considerations. It extends prior models by demonstrating that accountability is not static but evolves through iterative cycles of failure, learning, and adaptation. Practically, the study provides organizations with concrete pathways—such as implementing FactSheets, fairness toolkits, and counterfactual testing to operationalize accountability, while offering policymakers insights into the need for global harmonization and enforceable accountability metrics. Together, these contributions strengthen the bridge between ethical aspirations and real-world applications, advancing both academic understanding and practical governance of AI.

References

ASEAN. (2020). *ASEAN Guide on AI Governance and Ethics*. https://asean.org

Ayana, G., Dese, K., Daba Nemomssa, H., Habtamu, B., Mellado, B., Badu, K., Yamba, E., Faye, S. L., Ondua, M., & Nsagha, D. (2024). Decolonizing global AI governance: assessment of the state of decolonized AI governance in Sub-Saharan Africa. *Royal Society Open Science*, *11*(8), 231994.

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

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

Corbin, J., & Strauss, A. (2015). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4th ed.). SAGE Publications.

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., & Terem, E. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10).

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 2053951719860542.

Gianni, R., Lehtinen, S., & Nieminen, M. (2022). Governance of Responsible AI: From Ethical Guidelines to Cooperative Policies. *Frontiers in Computer Science*, *4*, 874653.

IBM. (2022). *IBM Principles for Trust and Transparency*.

IBM. (2023). *AI FactSheets 360: A Transparency Tool for AI Systems*.

IBM Research. (2023). *What is Trustworthy AI?* https://research.ibm.com

Infocomm Media Development Authority (IMDA). (2024). *Model AI Governance Framework for Generative AI*. https://imda.gov.sg

Insights, H. (2023). IBM Watson: From Healthcare Canary to a Failed Prodigy. *Healthark Insights White Paper*.

Kementerian Riset dan Teknologi / BRIN. (2020). *Strategi Nasional Kecerdasan Artifisial 2020–2045*. https://stranas.ai

Kim, T., Park, J., & Suh, A. (2020). Explainability in artificial intelligence: A survey and conceptual framework. *Information Systems Frontiers*, *22*(2), 411–427.

Li, J., & Chignell, M. (2022). FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI and Ethics*, *2*(4), 837–850.

Memarian, B., & Doleck, T. (2023). Algorithmic fairness in education: A systematic review. *British Journal of Educational Technology*, *54*(5), 1523–1542.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Bmj*, *339*.

Mökander, J., & Floridi, L. (2023). Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics*, *3*(2), 451–468.

National Institute of Standards, & Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (Issue NIST AI 100-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in Artificial Intelligence: What It Is and How It Works. *AI & Society*.

OECD. (2019). *OECD Principles on Artificial Intelligence*. https://oecd.ai/en/ai-principles

OECD. (2022). *AI FactSheets 360: Documenting the AI Lifecycle for Transparency*. https://oecd.ai

OECD/BIAC. (2022). *Implementing the OECD AI Principles: Challenges and Best Practices*.

Research, I. B. M. (2023). *IBM Researchers Check AI Bias with Counterfactual Text*.

Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: barriers and pathways forward. *International Affairs*, *100*(3), 1275–1286.

Rozenblit, L., Price, A., Solomonides, A., Joseph, A. L., Koski, E., Srivastava, G., Labkoff, S., Bray, D., Lopez-Gonzalez, M., & Singh, R. (2025). Toward responsible AI governance: Balancing multi-stakeholder perspectives on AI in healthcare. *International Journal of Medical Informatics*, 106015.

Saeidnia, H. R., Hashemi Fotami, S. G., Lund, B., & Ghiasi, N. (2024). Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact. *Social Sciences*, *13*(7), 381.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339.

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

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.

Veale, M., Matus, K., & Gorwa, R. (2023). AI and Global Governance: Modalities, Rationales, Tensions. *Annual Review of Law and Social Science*, *19*, 231–252.

Xia, B., Lu, Q., Zhu, L., Lee, S. U., Liu, Y., & Xing, Z. (2024). Towards a responsible ai metrics catalogue: A collection of metrics for ai accountability. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 100–111.

Yin, R. K. (2017). Case study research and applications: Design and methods. *(No Title)*.