

Explainable AI within responsible AI: A definitional boundary framework for AI governance

Stephen Fabeyo, *Stevox Consulting Inc.*, stevefab2002@gmail.com

Abstract

Explainable AI (XAI) and responsible AI (RAI) have emerged as central organizing concepts in AI ethics and governance, yet the definitional boundary between them remains inconsistently theorized across the literature. Many governance frameworks treat explainability as one principle among many within responsible AI. In contrast, others argue that explainability functions as a foundational dependency without which responsible AI principles cannot be operationalized. This study presents a systematic review of 42 peer-reviewed publications and regulatory documents to map that boundary and to characterize the conceptual interdependencies between XAI and RAI. Four major themes were identified: (1) definitional divergence, with XAI as a technical construct and RAI as a normative governance framework; (2) XAI as a technical enabler of RAI, operationalizing transparency, accountability, and fairness; (3) necessity without sufficiency, as XAI is necessary but insufficient for responsible AI; and (4) regulatory convergence, as governance instruments increasingly mandate explainability for RAI compliance. The resulting Definitional Boundary Framework maps XAI and RAI across eight dimensions, establishing an asymmetric interdependence: XAI serves as the technical layer through which RAI principles are operationalized, while RAI provides the normative architecture that gives XAI its governance purpose.

Keywords: explainable AI, responsible AI, AI governance, algorithmic transparency, AI ethics, algorithmic governance, AI policy

Introduction

As artificial intelligence reshapes how societies govern healthcare, employment, finance, criminal justice, and public administration, scholars and policymakers press a fundamental question: what does it truly mean for an AI system to be both explainable and responsible? Researchers widely recognize explainable AI (XAI) and responsible AI (RAI) as cornerstones of AI ethics and governance, yet they continue to define these constructs inconsistently and struggle to articulate how the two genuinely relate to one another.

XAI is broadly defined as a collection of methods and techniques that make the predictions and decisions of machine learning models interpretable to human users. The field originated in computer science and engineering, where the opacity of high-performing models necessitated approaches to clarify their internal logic. Techniques such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), decision trees, and attention mechanisms now characterize the XAI landscape, offering both local and global explanations of algorithmic behaviour (Barredo Arrieta et al., 2020; Dwivedi et al., 2023; Hamida et al., 2024).

RAI, by contrast, occupies a broad normative terrain, drawing on ethics, governance, and policy. Scholars and policymakers organize RAI around core principles, including fairness, accountability, transparency,

safety, privacy, and human oversight, ensuring developers build and deploy AI systems in ways that reflect societal values and comply with legal standards (Barletta et al., 2023; Ibitoye et al., 2025; Lu et al., 2024; Sadek et al., 2025). Institutions worldwide actively embrace RAI frameworks, including the NIST AI Risk Management Framework (Tabassi, 2023), the EU Artificial Intelligence Act (Regulation (EU) 2024/1689, 2024), the OECD Principles on AI (OECD, 2024), and Canada's Directive on Automated Decision-Making (Treasury Board of Canada Secretariat, 2024).

Problem Statements

Despite the proliferation of XAI techniques and RAI frameworks, scholars across the literature inconsistently theorize the relationship between these concepts. Some scholars conceptualize XAI as a component of RAI transparency, applying technical methods to operationalize it (Batoool et al., 2025; Birkstedt et al., 2023). Others contend that explainability constitutes the foundation of responsible AI, serving as the technical basis for implementing other RAI principles (Barredo Arrieta et al., 2020). A third perspective emphasizes the influence of context, arguing that explainability carries ethical significance in high-stakes domains but does not universally demand it (Freyer et al., 2024).

This conceptual ambiguity creates practical challenges. When policymakers and developers treat XAI and RAI as equivalent, they may mistakenly view explanation tools as sufficient for responsible AI, a misalignment described as the “principles-to-practices gap” by Sadek et al. (2025). Conversely, if practitioners see XAI only as a technical supplement to RAI, they may overlook its essential role in achieving responsible AI.

The current literature lacks a systematic review of the relationship between XAI and RAI. While existing reviews address XAI techniques (Hamida et al., 2024; Mohamed et al., 2025) or RAI governance frameworks (Kuehnert et al., 2025; Papagiannidis et al., 2025) in isolation, none systematically examines their intersection, including the nature of their connection, shared definitions, and conceptual divergences.

Purpose of the Study

This study conducts a systematic review of 42 peer-reviewed publications and regulatory documents. It clarifies the definitional boundaries between XAI and RAI, identifies conceptual frameworks that describe their relationship, and synthesizes these insights into a formal Definitional Boundary Framework. The study aims to structure the ongoing debate and equip scholars, practitioners, and policymakers with a clear reference for distinguishing between explainability and responsibility in AI governance. The paper organizes its analysis across sections covering background and literature review, methodology, thematic findings, discussion, implications, limitations, and conclusion.

Research Questions

RQ1: What are the definitional boundaries between explainable AI (XAI) and responsible AI (RAI) across the peer-reviewed literature and regulatory frameworks?

RQ2: How are the conceptual interdependencies between XAI and RAI characterized within this literature and regulatory framework?

Background and Literature Review

The Definitional Landscape of XAI

XAI emerged as a formal research agenda to address the opacity of modern machine learning models. Barredo Arrieta et al. (2020) define explainability as a model's ability to provide human-understandable insights into prediction generation. Their taxonomy distinguishes between intrinsically interpretable models, which are transparent by design, such as linear regression and decision trees, and post-hoc

explanation methods, which researchers apply after training to approximate or visualize internal logic. This distinction is fundamental for understanding XAI's capabilities and limitations as a governance tool.

Dwivedi et al. (2023) expand this taxonomy by categorizing explanation methods by scope (local versus global), methodology (backpropagation-based versus perturbation-based), and application (intrinsic versus post-hoc). Local methods, such as LIME, explain individual predictions by approximating model behavior near a specific input. Global methods, such as SHAP, provide aggregated insights into feature importance across an entire dataset. These distinctions matter because they determine the types of accountability claims an explanation can support (Langer et al., 2021). Pradhan et al. (2022) further clarify this landscape in data management, distinguishing between model-specific and model-agnostic approaches and between local and global explanation granularity.

Despite these taxonomic advances, XAI terminology remains inconsistent. Vainio-Pekka et al. (2023) identify conceptual ambiguity around the terms "interpretability," "explainability," "transparency," and "comprehensibility," which researchers often use interchangeably despite their distinct meanings. Interpretability refers to the extent to which an observer can understand a model's mechanism. Explainability concerns communicating that understanding through human-readable outputs. Transparency is openness about model design and data, while comprehensibility relates to how well explanations are understood by specific audiences. This conflation complicates efforts to assess whether an XAI method advances responsible AI objectives. Mohamed et al. (2025), in a meta-review of 14 systematic reviews, confirm that this ambiguity persists across subfields and application domains, and that the field has yet to establish a unified definitional framework.

Ridley (2025) introduces human-centred XAI (HCXAI) to address the limitations of purely algorithmic definitions of explainability. HCXAI asserts that an explanation is meaningful only if it supports human stakeholders' understanding, trust, and decision-making. In this view, XAI is not only a technical property but also a sociotechnical practice evaluated by its effects on human behavior and comprehension. Weitz (2022) supports this perspective, drawing on cognitive psychology and human-computer interaction to emphasize that effective explanations must consider mental models, cognitive load, and user expectations. This reframing shifts the evaluation of explainability from technical fidelity to social utility, aligning XAI more closely with the normative concerns of Responsible AI (RAI). Herrera (2025) documents this evolution, tracing the field's progression from algorithmic transparency to human-AI collaboration as a defining trend shaping contemporary XAI research.

The Definitional Landscape of RAI

Scholars and institutions have defined and applied RAI in various contexts. Fundamentally, RAI encompasses the ethical, legal, and socially responsible development and use of artificial intelligence systems (Barletta et al., 2023; Lu et al., 2024; Sadek et al., 2025). Most definitions highlight key principles: fairness, accountability, transparency, safety, privacy, and human oversight. The OECD Principles on AI (OECD, 2024) and the NIST AI Risk Management Framework (Tabassi, 2023) both identify these as essential for trustworthy AI governance.

Although there is broad agreement on these principles, their abstract nature makes practical implementation challenging. Barletta et al. (2023) describe this as "principle proliferation," noting that organizations often interpret and apply RAI principles differently. Madanchian and Taherdoost (2025) highlight how ethical theories and governance models influence RAI strategies, underscoring the gap between ideals and practice. Singhal et al. (2024), in a scoping review of fairness, accountability, transparency, and ethics across social media and healthcare AI, confirm that the four principles are widely adopted but inconsistently operationalized across application domains. Sadek et al. (2025) identify five main barriers to RAI implementation: abstract guidelines, conflicting values, unclear metrics, fragmentation across the AI lifecycle, and insufficient organizational accountability.

Papagiannidis et al. (2025) outline a taxonomy of RAI governance practices, distinguishing structural practices (governance policies and oversight), procedural practices (ethics integration throughout the lifecycle), and relational practices (stakeholder engagement and communication). This approach highlights RAI as an ongoing sociotechnical governance process rather than a static checklist. Ibitoye et al. (2025) also argue that RAI should be seen as a dynamic, participatory practice rather than a fixed set of technical requirements. Johnson et al. (2023) show that effective RAI in healthcare depends on integrating technical, organizational, and governance mechanisms. Xia et al. (2023) find that organizations primarily implement RAI principles through governance and regulatory tools rather than relying solely on technical solutions.

XAI Within RAI: Three Positions in the Literature

Much of the reviewed literature addresses the relationship between XAI and RAI, either directly or implicitly. Three main positions emerge from this work.

The first position considers XAI as one principle within RAI. Frameworks such as the NIST AI Risk Management Framework (RMF) (Tabassi, 2023), the European Union AI Act (Regulation (EU) 2024/1689, 2024), and the OECD Principles (OECD, 2024) include explainability as part of responsible AI requirements. From this perspective, XAI is necessary but not the sole focus; responsible AI also requires fairness, safety, privacy, and accountability, with no single principle taking precedence. Kuehnert et al. (2025) note that explainability tools are widely used to implement RAI principles, but most governance tools focus on technical development and often overlook broader organizational and societal stakeholders.

The second position, described by Meske et al. (2022), argues that explainability is the foundation for all other RAI principles. In this view, XAI underpins responsible AI: without explainability, fairness cannot be assessed, accountability cannot be established, and trust cannot be built. Cali et al. (2021) and Barredo Arrieta et al. (2020) support this perspective, showing that accountability, bias detection, and fairness evaluation all require interpretability of algorithmic decisions.

The third position, presented by Freyer et al. (2024), holds that the need for explainability depends on context. In healthcare, mandatory explainability is ethically justified due to high stakes and its effect on patient autonomy and safety. In other fields, opaque but high-performing systems may be acceptable if practitioners rigorously validate and monitor them. Turchioe et al. (2024) add that context includes both the application domain and the identities of stakeholders, such as patients, whose explainability needs may differ from those of clinicians or developers. Domain-specific applications reinforce this point. Bagada and Patel (2025) demonstrate, in spinal metastasis imaging, that integrating XAI techniques with RAI principles is necessary to support clinical accountability, but the explainability layer alone does not satisfy the broader governance requirements of medical AI deployment. This diversity makes it difficult to set a universal XAI requirement within RAI frameworks.

Methodology

This study applies a systematic review methodology based on the PRISMA framework (Moher et al., 2009). The review synthesizes peer-reviewed literature and official regulatory documents to examine the conceptual relationship between XAI and RAI, with a focus on their definitions and interdependencies.

Search Strategy

The review involved a comprehensive search of six databases: ACM Digital Library, IEEE Xplore, SpringerLink, Elsevier ScienceDirect, SSRN, and Google Scholar. Search terms included Boolean queries, adapted to each database's syntax and restricted to 2020 to 2025, including "explainable AI," "responsible AI," "XAI," "RAI," "AI governance," "AI transparency," "algorithmic accountability," and "AI ethics." The

review sourced regulatory documents directly from the official websites of the OECD, NIST, the European Union, and the Treasury Board of Canada. The review focused on peer-reviewed publications from 2020 to 2025 alongside governance instruments such as the EU AI Act and the NIST AI RMF relevant to XAI and RAI.

Inclusion and Exclusion Criteria

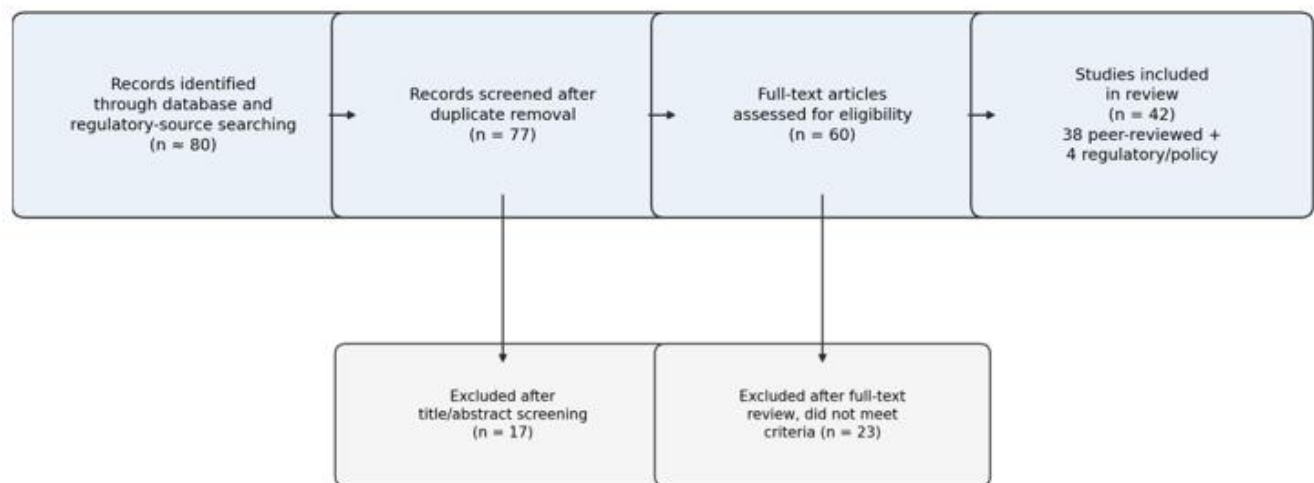
The review included articles that met the following criteria: (1) peer-reviewed publications or officially published regulatory and policy documents; (2) direct focus on XAI techniques, RAI frameworks, or their conceptual relationship; (3) analytical, empirical, or conceptual contributions to transparency, accountability, or explainability in AI systems; and (4) written in English. The review excluded articles that addressed XAI or RAI only tangentially, focused solely on technical benchmarks without governance or ethical analysis, or were duplicates. The initial search identified approximately 80 articles and policy documents. After screening titles and abstracts, the review examined 60 sources in full. Applying the inclusion and exclusion criteria, the review retained 42 sources: 38 peer-reviewed academic publications and 4 regulatory or policy documents. Table 1 summarizes the PRISMA selection process.

Table 1. PRISMA Study Selection Summary

Phase	Count
Records identified through database searching	~80
Records after duplicate removal	~77
Records screened (title and abstract)	77
Full-text articles assessed for eligibility	60
Studies excluded (did not meet criteria)	23
Peer-reviewed studies included	38
Regulatory and policy documents included	4
Total sources retained	42

Note. Adapted from PRISMA reporting guidelines (Moher et al., 2009).

Figure 1. PRISMA Flow Diagram of Study Selection



Analytical Approach

The review applied a thematic synthesis approach to analyze the selected literature and systematically coded each source for its treatment of explainable artificial intelligence (XAI) and responsible artificial intelligence (RAI) definitions, the nature of the XAI-RAI relationship (including dependency, subsumption, sufficiency, or independence), and any regulatory or governance implications. Codes were grouped into themes and refined through iterative cross-source comparison. This process produced four primary themes, which the analysis and results section presents.

Findings

Theme 1: Definitional Divergence: XAI as Technical Construct, RAI as Normative Framework

The literature consistently identifies a fundamental difference between XAI and RAI. XAI is primarily a technical and methodological concept, comprising tools, algorithms, and methods that generate human-interpretable outputs from machine learning models. In contrast, RAI is a normative governance framework, defined by ethical principles, institutional mechanisms, and regulatory requirements that drive the development and use of AI for social benefit.

Foundational texts make this distinction clear in how they define each domain. Barredo Arrieta et al. (2020) define XAI by its technical properties, such as the type and scope of explanation and applicable model architecture. Mohamed et al. (2025) also frame XAI as a technical pipeline, classifying models by transparency as white-box, grey-box, or black-box. Hamida et al. (2024) and Dwivedi et al. (2023) categorize XAI methods into backpropagation-based and perturbation-based approaches. Pradhan et al. (2022) further organize XAI by scope, methodology, and application context. Herrera (2025) traces the field's progression from algorithmic transparency to human-AI collaboration.

RAI definitions take a markedly different approach. Lu et al. (2024) and Sadek et al. (2025) define responsible AI as the ethical design and deployment of AI to promote social welfare, requiring value judgments about benefits, harms, affected parties, and the responsibilities of developers and deployers. Barletta et al. (2023) identify six to nine core principles in RAI frameworks, including transparency, fairness, accountability, safety, privacy, human oversight, normative commitments, and not technical specifications. Ibitoye et al. (2025) establish that RAI frameworks are rooted in ethical theory and function as sociotechnical governance processes within institutional, cultural, and political contexts. Xia et al. (2023) reinforce this by demonstrating that organizations and regulators operationalize responsible AI principles through governance and regulatory instruments, rather than solely through technical mechanisms.

This technical versus normative distinction carries significant implications for how XAI and RAI relate to one another. Because XAI operates at a technical level and RAI at a normative one, the two function at different levels of abstraction and cannot be directly equated. Vainio-Pekka et al. (2023) demonstrate that XAI lacks a unified framework connecting technical explanation methods to broader RAI governance objectives, and that most XAI research prioritizes technical fidelity over ethical adequacy. Terminological inconsistencies compound this misalignment; scholars use “explainability,” “interpretability,” and “transparency” interchangeably despite their distinct meanings, further undermining efforts to align XAI with RAI principles (Langer et al., 2021; Ridley, 2025). This definitional divergence exposes a real structural gap between what XAI delivers and what RAI demands.

Theme 2: XAI as Technical Enabler of RAI: The Dependency Relationship

Although XAI and RAI are distinct concepts, the literature consistently frames their relationship as one of dependency: XAI supplies the technical infrastructure needed to operationalize RAI principles. Across

sources, the core argument holds: without explainability, responsible AI principles such as transparency, accountability, and fairness cannot be effectively implemented.

Barredo Arrieta et al. (2020) articulate this argument most clearly, proposing a conceptual architecture in which XAI serves as the foundational layer of a responsible AI system. In their model, explainability enables transparency, which drives fairness auditing, accountability mechanisms, and trust. The dependency is both directional and structural: without explainability, the responsible AI framework loses its technical foundation. Other sources reinforce this view. Cali et al. (2021) position explainability as the core technical infrastructure for ethical oversight. Papagiannidis et al. (2025) characterize XAI as a technical dependency that enables organizations to implement RAI governance mechanisms. Cheong (2024) identifies explainability as the key technical mechanism for accountability and transparency in RAI.

The NIST AI Risk Management Framework (Tabassi, 2023) positions explainability and interpretability as foundational to trustworthy AI, making them essential to the governance functions of measuring, mapping, and managing AI risk. The framework's four governance functions require organizations to set explainability standards, assess risks, evaluate ethical compliance, and execute mitigation strategies, with explainability serving as both a measurable characteristic and an enabling condition for effective risk governance. Radanliev (2025) further demonstrates that, across international AI governance frameworks, transparency and explainability consistently intersect with fairness and accountability, reinforcing this dependency relationship across jurisdictions.

Regulatory instruments reinforce this dependency directly. Canada's Directive on Automated Decision-Making (Treasury Board of Canada Secretariat, 2024) requires departments to provide meaningful explanations to individuals affected by automated decisions, directly operationalizing RAI principles through XAI. The EU AI Act (Regulation (EU) 2024/1689, 2024) mandates transparency and documentation for high-risk AI systems, recognizing explainability as essential for responsible AI deployment. Kuehnert et al. (2025), in a meta-analysis of 220 responsible AI tools, find that explainability tools such as LIME and SHAP rank among the most widely used for implementing RAI principles during model development and evaluation, confirming XAI's central role in RAI implementation.

Theme 3: Necessity Without Sufficiency: The Limits of Explainability as a Governance Mechanism

While the dependency relationship is well-established, the literature consistently surfaces a key counterpoint: explainability is necessary but not sufficient for responsible AI. This theme cuts across technical AI research, healthcare ethics, organizational studies, and legal scholarship, making it a significant conceptual contribution to the XAI-RAI debate.

The necessity-sufficiency distinction matters because it challenges the misconception that deploying XAI tools equates to implementing responsible AI. Deck et al. (2024) confront this directly in their critical survey of fairness claims attributed to XAI. Their review of 175 academic publications reveals that many claims about XAI improving fairness lack clarity about the explanation methods used, the fairness definitions applied, or the stakeholders involved. This conflation overstates explainability's governance contribution by equating the ability to explain decisions with the ability to ensure fairness or ethics. Deck et al. draw a sharp distinction between XAI's epistemic role, which helps stakeholders understand model fairness, and its substantive role, which involves directly modifying model behavior to reduce unfair outcomes. Most XAI tools serve an epistemic function primarily and cannot alone deliver substantive fairness improvements; a distinction that carries significant implications for responsible AI governance.

Bach et al. (2025) reinforce this point through their systematic review of responsible AI practices in 45 empirical studies. They identify eleven key themes in RAI implementation, with explainability representing only one. Harm prevention, stakeholder participation, governance structures, continuous monitoring, and human-AI calibration all operate independently of and alongside explainability. Organizations cannot rely on XAI alone; they must build institutional and governance structures that address all RAI requirements.

Lu et al. (2024) similarly argue that responsible AI demands attention at the system level, not just the algorithmic level, organizing RAI practices into governance, process, and product patterns. In this framework, explainability occupies one product-level capability within a broader multi-layer architecture.

Organizational settings expose XAI's limitations most sharply. Cools (2025) identifies a "principles-to-practices gap" at the BBC, where adopting explainability principles failed to produce consistent implementation due to unclear responsibilities, varying interpretations, and absent standardized procedures. Sadek et al. (2025) report similar barriers, establishing that technical tools like XAI alone are insufficient because responsible AI is a sociotechnical challenge that demands organizational, institutional, and participatory mechanisms. Ibitoye et al. (2025) reinforce this view, arguing that explainability tools operate within complex sociotechnical systems where ethical outcomes depend more on institutional structures and cultural norms than on technical features alone. Tiwari and Farag (2025) extend this contextual argument to autonomous vehicles, where bias and fairness risks demand domain-specific RAI safeguards that go beyond generic explainability requirements. Wang et al. (2024) advance a cybernetic framing of this problem, arguing that managing dynamic black-box AI systems requires continuous feedback and control mechanisms at the organizational level, since technical explanations alone cannot regulate evolving model behaviour.

Langer et al. (2021) advance a human-centred perspective. Their stakeholder model of XAI demonstrates that explanation methods serve different purposes for different audiences: developers need explanations for debugging, regulators for compliance auditing, and end users for trust and understanding. No single explanation method satisfies all stakeholder needs at once. Consequently, even a technically rigorous XAI system can fail to meet responsible AI governance requirements if its explanations are not tailored to the intended audience. Fabeyo (2025), in a study of XAI in employment decision-making, provides empirical evidence of this gap. Explanation tools in algorithmic hiring systems routinely produce outputs that are technically precise but semantically inaccessible to lay users, such as HR professionals or job applicants. This evidence reinforces the human-centred critique and confirms that technical fidelity does not guarantee governance adequacy.

Theme 4: Regulatory Convergence: Institutionalizing the XAI-RAI Relationship

The fourth theme identifies a decisive trend in both international and national governance: regulatory frameworks are embedding explainability requirements into responsible AI compliance regimes, establishing the XAI-RAI dependency at the policy level. However, this convergence varies across jurisdictions, with notable differences in how regulatory contexts define the relationship between explainability and responsibility.

The EU Artificial Intelligence Act (Regulation (EU) 2024/1689, 2024) delivers the most comprehensive example. Applying a risk-based approach, the Act mandates transparency and documentation for high-risk AI systems and prohibits those that threaten fundamental rights. Organizations deploying high-risk systems in employment, healthcare, law enforcement, and critical infrastructure must maintain technical documentation, enable human review of AI decisions, and execute risk management throughout the AI lifecycle. The Act explicitly positions explainability as a compliance tool within responsible AI governance.

The NIST AI Risk Management Framework (Tabassi, 2023) takes a similar yet more process-oriented approach. Its four governance functions, Govern, Map, Measure, and Manage, require organizations to set explainability standards, assess context-specific risks, evaluate AI outputs for ethical compliance, and execute risk mitigation strategies. In contrast to the EU model, this framework is voluntary, reflecting a distinct regulatory philosophy. Criado et al. (2025) demonstrate how AI governance frameworks shape responsible AI practices across public administration, from national policy down to agency-level implementation.

Canada's Directive on Automated Decision-Making (Treasury Board of Canada Secretariat, 2024) presents a clear example of regulatory XAI-RAI integration. The Algorithmic Impact Assessment requires departments to evaluate AI systems by risk level and deliver meaningful explanations to individuals affected by automated decisions. This approach directly operationalizes the RAI principle of procedural fairness through XAI, confirming that the two concepts can be institutionally linked without conceptual convergence. The OECD Principles on AI (OECD Legal Instruments, n.d.) establish a reference standard at the intersection of regulatory convergence and voluntary guidance, shaping national regulatory approaches in member and partner countries. Birkstedt et al. (2023) find that although ethical AI principles command wide adoption, a gap persists between stated principles and their practical implementation, particularly in enforcement and accountability. Batool et al. (2025) further establish that governance frameworks remain fragmented and that enforcement consistently lags behind regulatory principles. Constantinides et al. (2024) respond to this gap by proposing a method for generating RAI guidelines that are grounded directly in regulatory text and usable across both technical and non-technical roles, illustrating an emerging design pattern for translating regulatory convergence into operational practice.

Collectively, the four themes construct a structured understanding of the boundary between XAI and RAI. Table 2 maps this boundary across eight analytical dimensions from the thematic analysis. This framework represents the review's central conceptual contribution, delivering a formal mapping the field currently lacks and equipping researchers, practitioners, and regulators to define, implement, and evaluate the relationship between explainability and responsible AI governance.

Table 2. XAI-RAI Definitional Boundary Framework

Dimension	Explainable AI (XAI)	Responsible AI (RAI)
Construct nature	Technical and methodological	Normative and ethical
Primary purpose	Interpretability of model predictions and decision logic	Ethical governance of AI systems and societal outcomes
Scope	Model-level and algorithmic	System-level and sociotechnical
Key outputs	Feature attributions, saliency maps, surrogate models, local and global explanations	Principles, policies, governance structures, and oversight mechanisms
Evaluation criteria	Fidelity, accuracy, comprehensibility, stakeholder usability	Fairness, accountability, safety, human oversight, trust
Dependency direction	Technical precondition enabling RAI operationalization	Normative framework that defines and frames XAI requirements
Sufficiency for ethical AI	Necessary but not sufficient for responsible AI outcomes	Requires XAI as technical infrastructure; extends beyond it institutionally
Regulatory expression	Transparency obligations, documentation requirements, and audit trails	Risk management frameworks, bias audits, human oversight, and redress mechanisms

To illustrate how the dependency direction in Table 2 operates in practice, consider the EU AI Act's (Regulation (EU) 2024/1689, 2024) treatment of high-risk AI systems. The Act, an RAI instrument, first defines which systems qualify as high risk and what governance outcomes they must achieve, then requires technical documentation and explanations sufficient to support human review of those systems. The XAI methods a developer selects, such as feature attribution

or surrogate models, are therefore chosen to satisfy an explanation standard defined by the regulation, not the reverse. This case shows concretely how RAI “defines and frames” XAI requirements: the normative instrument sets the explanation target, and the technical method is selected to meet it

Discussion of Findings

The four themes this review identifies construct a nuanced understanding of the XAI-RAI relationship, which resists simple definition. XAI and RAI are neither synonymous nor fully independent; they operate asymmetrically and at different levels of abstraction, connected by a structural dependency that is necessary but not sufficient for responsible AI governance.

Clarifying the Conceptual Architecture

The boundary framework in Table 2 formalizes a distinction that previous literature has described but not clearly defined: XAI and RAI differ in both content and ontological character. Building on the four themes above, this review defines XAI as the set of technical methods, tools, and design practices that render an AI system’s predictions and decision logic interpretable to relevant stakeholders, and defines RAI as the institutional and ethical governance framework, comprising principles, oversight mechanisms, and regulatory instruments, that directs the responsible development, deployment, and monitoring of AI systems and draws on XAI as one of its operational mechanisms. Conflating these concepts produces category errors with real-world consequences. Equating XAI with RAI ignores essential governance activities such as stakeholder engagement, organizational accountability, and ongoing monitoring. Conversely, treating RAI as a self-executing checklist, without the necessary technical infrastructure, reduces responsible AI to mere aspiration. The framework establishes that XAI serves as the technical layer within a multi-level RAI architecture, and practitioners must evaluate its governance adequacy against the normative standards RAI demands (Lu et al., 2024; Meske et al., 2022; Papagiannidis et al., 2025).

The Sufficiency Problem and Its Governance Implications

Theme 3’s necessity-sufficiency finding carries the clearest governance implication: organizations and regulators that treat XAI deployment as proof of RAI compliance commit the same “principles-to-practices gap” identified above (Cools, 2025; Deck et al., 2024; Sadek et al., 2025). Responsible AI governance must therefore distinguish between explanatory transparency, which describes what a model does, and normative adequacy, which determines whether those actions are ethically acceptable. XAI drives explanatory transparency, while normative adequacy belongs to the broader RAI governance system, of which XAI remains only one component (Vainio-Pekka et al., 2023).

Human-Centered Explainability as a Bridge

Ridley (2025) and Langer et al. (2021) argue that XAI must be evaluated not only on technical fidelity but also on its impact on stakeholder understanding, trust, and decision-making. This approach enables XAI research and RAI governance to pursue a shared goal: delivering explanations that are both accurate and meaningful for end users. Weitz (2022) calls for incorporating psychological concepts such as cognitive load and mental models into XAI design. Herrera (2025) frames this as a shift from algorithmic transparency to human-AI collaboration, directly aligning technical explainability with responsible AI objectives. However, Fabeyo (2025) confirms that this alignment continues to fall short in practice, as explanations in hiring systems routinely achieve precision but remain inaccessible to lay users. Human-centred XAI design is best positioned to close this gap.

Regulatory Implications

The regulatory convergence Theme 4 identifies confirms that the field is institutionalizing the XAI-RAI relationship at the policy level, though through divergent mechanisms, mandatory obligations, voluntary guidance, and tiered risk-based assessment, that signal lasting regulatory pluralism rather than a single prevailing model. Birkstedt et al. (2023) and Batool et al. (2025) expose the ongoing gap between regulatory principles and enforcement mechanisms as a defining challenge for AI governance, establishing that convergence in principles does not yet guarantee convergence in implementation and accountability.

Implications of Findings

The findings carry three key implications. First, the definitional divergence Theme 1 exposes the need for a shared vocabulary among researchers, standards bodies, and regulators to clearly distinguish between explainability, interpretability, and transparency. Without this consensus, evaluating compliance with governance instruments that require explainability, such as the EU AI Act (Regulation (EU) 2024/1689, 2024) and the NIST AI RMF (Tabassi, 2023), remains impossible because their standards stay unclear. Second, the boundary framework in Table 2 equips practitioners designing AI governance systems with a direct reference: XAI requirements must be addressed at the model and product layer, while accountability, oversight, and redress mechanisms must be embedded at the institutional and governance layer. These functions are complementary and not interchangeable. Third, the necessity-without-sufficiency finding compels regulators to move beyond treating explainability as a binary compliance item. Instead, they must define the governance conditions that make explanations meaningful, including who receives them, in what form, with what rights of challenge, and under what monitoring. Until regulatory frameworks specify these conditions, the principles-to-practices gap (Cools, 2025; Sadek et al., 2025) will persist, regardless of the technical quality of the explanations.

Conclusion

Limitations

This review carries several limitations worth noting. First, most of the literature reviewed draws from English-language publications by North American and European institutions. As a result, the perspectives on XAI and RAI may not capture all cultural, legal, or institutional contexts. Research from Asian, African, and Latin American regions may offer alternative views on the XAI-RAI relationship that this review underrepresents.

Second, because XAI and RAI are rapidly evolving, findings grounded in literature published through 2025 risk becoming outdated as new techniques, frameworks, and regulations emerge. For example, the EU AI Act entered into force in 2024, and its guidance and case law continue to develop. Third, as a qualitative systematic review, this study applies thematic synthesis rather than statistical meta-analysis. This constrains the ability to make quantitative claims about the frequency or strength of positions in the literature. The findings represent interpretive conclusions grounded in textual analysis, not empirically measurable outcomes. Fourth, this review interrogates the conceptual relationship between XAI and RAI, not empirical evidence of XAI's effectiveness in specific governance contexts. Future research must build on these conceptual findings by testing whether specific XAI techniques improve RAI outcomes in real organizational settings.

Conclusion

This study set out to clarify the definitional boundaries between explainable AI (XAI) and responsible AI (RAI), and to interrogate their conceptual interdependencies across peer-reviewed literature and regulatory frameworks. Four key themes emerged: definitional divergence, positioning XAI as a technical construct and RAI as a normative governance framework; XAI as a technical enabler of RAI, supplying the infrastructure for transparency, accountability, and fairness; necessity without sufficiency, confirming that XAI is essential but not sufficient for responsible AI; and regulatory convergence, as governance instruments actively demand explainability within responsible AI compliance regimes. From these themes, the study constructs the Definitional Boundary Framework in Table 2, an eight-dimensional mapping of the XAI-RAI relationship. This framework equips researchers, practitioners, and regulators to define constructs, design governance systems, and specify explainability requirements.

These findings deliver a clear practical implication: the field must stop equating XAI with RAI. Implementing explanation tools alone does not constitute responsible AI governance, any more than publishing an ethics policy guarantees ethical practice. Explainability is necessary, but effective governance also demands accountability structures, stakeholder engagement, regulatory oversight, and ongoing monitoring. At the same time, practitioners must not treat XAI as a minor addition to RAI principles. The evidence confirms that explainability is central to responsible AI, not because explanation equals responsibility, but because it supplies the technical foundation for other governance components.

Recommendation

Future research must pursue three areas: developing shared definitional standards for XAI and RAI to drive interdisciplinary communication and comparison; empirically testing the governance impact of specific XAI techniques in real organizational and regulatory settings; and advancing human-centred XAI design to ensure explanations reach and resonate with all stakeholders. AI systems in employment, healthcare, and public administration are making decisions that affect millions. Their responsibility rests on their explainability and on whether that explainability is anchored in robust governance structures.

Acknowledgements

The author acknowledges using AI writing assistance (Grammarly) for language editing and grammar refinement to improve clarity and readability. The author takes full responsibility for the originality, integrity and accuracy of the work.

References

- Bach, T. A., Kaarstad, M., Solberg, E., & Babic, A. (2025). Insights into suggested Responsible AI (RAI) practices in real-world settings: A systematic literature review. *AI and Ethics*, 5(3), 3185–3232. <https://doi.org/10.1007/s43681-024-00648-7>
- Bagada, A. A., & Patel, P. V. (2025). Spinal metastasis: Imaging using XAI and RAI techniques. In R. Malviya & S. Sundram (Eds.), *Explainable and responsible artificial intelligence in healthcare* (pp. 115–143). Wiley. <https://doi.org/10.1002/9781394302444.ch5>
- Barletta, V. S., Caivano, D., Gigante, D., & Ragone, A. (2023). A rapid review of responsible AI frameworks: How to guide the development of ethical AI. *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, EASE '23*, 358–367. <https://doi.org/10.1145/3593434.3593478>

- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics*, 5(3), 3265–3279. <https://doi.org/10.1007/s43681-024-00653-w>
- Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7), 133–167. <https://doi.org/10.1108/INTR-01-2022-0042>
- Cali, U., Kuzlu, M., Pipattanasomporn, M., Kempf, J., & Bai, L. (2021). Foundations of big data, machine learning, and artificial intelligence and explainable artificial intelligence. In U. Cali, M. Kuzlu, M. Pipattanasomporn, J. Kempf, & L. Bai (Eds.), *Digitalization of power markets and systems using energy informatics* (pp. 115–137). Springer International Publishing. https://doi.org/10.1007/978-3-030-83301-5_6
- Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6. Article 1421273. <https://doi.org/10.3389/fhumd.2024.1421273>
- Constantinides, M., Bogucka, E., Quercia, D., Kallio, S., & Tahaei, M. (2024). RAI guidelines: Method for generating responsible AI guidelines grounded in regulations and usable by (non-)technical roles. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), Article 388. <https://doi.org/10.1145/3686927>
- Cools, H. (2025). Navigating the responsible AI landscape: Unraveling the principles-to-practices gap of transparency and explainability at the BBC. *Information, Communication & Society*, 0(0), 1–21. <https://doi.org/10.1080/1369118X.2025.2553382>
- Criado, J. I., Sandoval-Almazán, R., & Gil-Garcia, J. R. (2025). Artificial intelligence and public administration: Understanding actors, governance, and policy from micro, meso, and macro perspectives. *Public Policy and Administration*, 40(2), 173–184. <https://doi.org/10.1177/09520767241272921>
- Deck, L., Schoeffer, J., De-Arteaga, M., & Kühn, N. (2024). A critical survey on fairness benefits of explainable AI. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1579–1595. <https://doi.org/10.1145/3630106.3658990>
- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, 55(9), 194:1-194:33. <https://doi.org/10.1145/3561048>
- Fabeyo, S. (2025). Explainable AI in employment decision-making: A systematic review of transparency methods in hiring algorithms. *Issues in Information Systems*, 26(3), 127–135. https://doi.org/10.48009/3_iis_2025_2025_110
- Freyer, N., Groß, D., & Lipprandt, M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: A systematic review of reasons. *BMC Medical Ethics*, 25(1), 104. <https://doi.org/10.1186/s12910-024-01103-2>

- Hamida, S. U., Chowdhury, M. J. M., Chakraborty, N. R., Biswas, K., & Sami, S. K. (2024). Exploring the landscape of explainable artificial intelligence (XAI): A systematic review of techniques and applications. *Big Data and Cognitive Computing*, 8(11), 149. <https://doi.org/10.3390/bdcc8110149>
- Herrera, F. (2025). Reflections and attentiveness on eXplainable artificial intelligence (XAI): The journey ahead from criticisms to human–AI collaboration. *Information Fusion*, 121, 103133. <https://doi.org/10.1016/j.inffus.2025.103133>
- Ibitoye, A. O., Nkwo, M. S., & Orji, R. (2025). Rethinking responsible AI from ethical pillars to sociotechnical practice. *AI and Ethics*, 5(6), 6207–6223. <https://doi.org/10.1007/s43681-025-00809-2>
- Johnson, M., Albizri, A., & Harfouche, A. (2023). Responsible artificial intelligence in healthcare: Predicting and preventing insurance claim denials for economic and social wellbeing. *Information Systems Frontiers*, 25(6), 2179–2195. <https://doi.org/10.1007/s10796-021-10137-5>
- Kuehnert, B., Kim, R. M., Forlizzi, J., & Heidari, H. (2025). The "who," "what," and "how" of responsible AI governance: A systematic review and meta-analysis of (actor, stage)-specific tools. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2991–3005. <https://doi.org/10.1145/3715275.3732191>
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys*, 56(7), Article 173. <https://doi.org/10.1145/3626234>
- Madanchian, M., & Taherdoost, H. (2025). Ethical theories, governance models, and strategic frameworks for responsible AI adoption and organizational success. *Frontiers in Artificial Intelligence*, 8. Article 1619029. <https://doi.org/10.3389/frai.2025.1619029>
- Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1), 53–63. <https://doi.org/10.1080/10580530.2020.1849465>
- Mohamed, A., Abdelqader, K., & Shaalan, K. (2025). Explainable artificial intelligence: A systematic review of progress and challenges. *Intelligent Systems with Applications*, 28, 200595. <https://doi.org/10.1016/j.iswa.2025.200595>
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology*, 62(10), 1006–1012. <https://doi.org/10.1016/j.jclinepi.2009.06.005>
- OECD. (2024). *Recommendation of the Council on artificial intelligence* (OECD/LEGAL/0449). <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), 101885. <https://doi.org/10.1016/j.jsis.2024.101885>

- Pradhan, R., Lahiri, A., Galhotra, S., & Salimi, B. (2022). Explainable AI: Foundations, applications, opportunities for data management research. *Proceedings of the 2022 International Conference on Management of Data, SIGMOD '22*, 2452–2457. <https://doi.org/10.1145/3514221.3522564>
- Radanliev, P. (2025). AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence*, 39(1), 2463722. <https://doi.org/10.1080/08839514.2025.2463722>
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Ridley, M. (2025). Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 76(1), 98–120. <https://doi.org/10.1002/asi.24889>
- Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, R. A., & Cave, S. (2025). Challenges of responsible AI in practice: Scoping review and recommended actions. *AI & Society*, 40(1), 199–215. <https://doi.org/10.1007/s00146-024-01880-9>
- Treasury Board of Canada Secretariat. (2024, July 2). *Directive on automated decision-making*. Government of Canada. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>
- Singhal, A., Neveditsin, N., Tanveer, H., & Mago, V. (2024). Toward fairness, accountability, transparency, and ethics in AI for social media and health care: Scoping review. *JMIR Medical Informatics*, 12, e50048. <https://doi.org/10.2196/50048>
- Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- Tiwari, A., & Farag, H. E. Z. (2025). Responsible AI framework for autonomous vehicles: Addressing bias and fairness risks. *IEEE Access*, 13, 58800–58822. <https://doi.org/10.1109/ACCESS.2025.3556781>
- Turchioe, M. R., Hermann, A., & Benda, N. C. (2024). Recentering responsible and explainable artificial intelligence research on patients: Implications in perinatal psychiatry. *Frontiers in Psychiatry*, 14, 1321265. <https://doi.org/10.3389/fpsyt.2023.1321265>
- Vainio-Pekka, H., Agbese, M. O.-O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023). The role of explainable AI in the research field of AI ethics. *ACM Transactions on Interactive Intelligent Systems*, 13(4), Article 26. <https://doi.org/10.1145/3599974>
- Wang, B. Y., Boell, S., Li, C. X., & Chen, E. (2024). Responsible management for dynamic black box AI: A cybernetic approach. In *Proceedings of the International Conference on Information Systems (ICIS 2024)* (Paper 11). Association for Information Systems. https://aisel.aisnet.org/icis2024/it_implement/it_implement/11
- Weitz, K. (2022). Towards human-centered AI: Psychological concepts as foundation for empirical XAI research. *IT—Information Technology*, 64(1–2), 71–75. <https://doi.org/10.1515/itit-2021-0047>
- Xia, B., Lu, Q., Perera, H., Zhu, L., Xing, Z., Liu, Y., & Whittle, J. (2023). *Towards Concrete and Connected AI Risk Assessment (C 2 AIRA): A Systematic Mapping Study* (p. 116). <https://doi.org/10.1109/CAIN58948.2023.00027>