AI Governance Is Not Enough to Prove Responsibility

A Conceptual and Testable Architecture for Demonstrable Responsibility in AI Systems

Raphael A. La Touche
La Touche Academy Ltd, London, United Kingdom

Canonical public research edition | Version 1.0 | Published 8 August 2026

Edition: Canonical public research edition Status: Published Version: 1.0 Publication date: 8 August 2026
Peer-review status: Not peer reviewed Publisher: La Touche Academy Ltd Publication index: https://responsibilityinfrastructure.com/publications Related protocol: RI-PROTOCOL-CORE-01

Recommended citation: La Touche, R. A. (2026). AI Governance Is Not Enough to Prove Responsibility: A Conceptual and Testable Architecture for Demonstrable Responsibility in AI Systems. Canonical public research edition, version 1.0. La Touche Academy Ltd.

Public edition note: This is version 1.0 of the canonical public research edition. It is the author's official public reference version and has not undergone formal peer review. It identifies the author and publisher, includes the full declarations, and links the conceptual argument to the published Responsibility Infrastructure Protocol. Independent institutional, regulatory, and empirical validation remain to be established.

Research status. This is a problem-definition and research-agenda paper. It does not establish that the Operational Responsibility Gap is widespread, that Responsibility Infrastructure is necessary, or that the proposed architecture is effective. Current evidence consists of conceptual analysis, public protocol materials, and an internal prototype; independent institutional authority, regulatory recognition, empirical effectiveness, broad adoption, and external reliance remain to be established through external testing and institutional development.

Abstract

Artificial intelligence governance frameworks increasingly define organisational duties relating to risk management, documentation, transparency, and human oversight. Those frameworks are necessary, but they do not necessarily produce a reconstructable account of responsibility in a specific operational event: who held authority, who accepted responsibility, what action followed, what evidence supports the outcome, and how the resulting claim was assessed. This paper defines that limitation as the Operational Responsibility Gap and proposes Responsibility Infrastructure as a conceptual architecture for representing and reconstructing responsibility claims across organisational boundaries. Its central claim is limited: where responsibility must support external reliance, allocation, agency, and accountability may need to be jointly evidenced rather than inferred from fragmented governance records. The proposal is conceptual and testable; it has not yet been empirically validated across organisations, sectors, or implementations. This paper does not claim to have solved the problem or established the effectiveness of Responsibility Infrastructure; it defines a bounded and testable architectural hypothesis and identifies diagnostic and pilot conditions through which its necessity, proportionality, and practical value may be independently assessed.

Keywords: *Artificial Intelligence; AI Governance; Responsibility; Accountability; Verification; Responsibility Infrastructure*

1. Introduction

Artificial intelligence increasingly supports consequential organisational and public decisions. When one of those decisions is challenged, the practical question is not only whether the AI system was governed, but whether responsibility for the specific event can be reconstructed with sufficient clarity, authority, and

evidence. Existing governance frameworks may establish duties, controls, and oversight without necessarily showing who held authority, who accepted responsibility, what action followed, and how the resulting claim was assessed.

This paper distinguishes the existence of responsibility from its demonstrability. Responsibility may exist in principle even where the relevant records are incomplete, contested, or dispersed across systems. The issue examined here is therefore narrower than responsibility in the abstract: it concerns the conditions under which responsibility can be demonstrated to another party without depending materially on inference, memory, or institutional self-assertion.

Organisations already produce extensive documentation around AI systems, including policies, risk assessments, model documentation, technical logs, approval records, audit trails, and incident reports. These materials remain valuable, but they are often created by different actors, for different purposes, and at different stages of the system lifecycle. They may therefore fail to form a coherent account of authority, acceptance, required action, supporting evidence, assessment, and current standing. The problem is not necessarily too few records; it is the absence of a common structure through which responsibility can be reconstructed across them.

This paper calls that condition the Operational Responsibility Gap: the situation in which governance expectations exist, but the evidence required to reconstruct responsibility for a particular event is incomplete, fragmented, non-authoritative, or difficult for an independent party to assess. The term identifies an evidential and operational problem, not an absence of responsibility itself.

The gap should also be distinguished from poor behaviour or weak culture. An organisation may have committed staff, extensive policies, and established oversight processes while remaining unable to show how responsibility moved through a particular event. The problem may therefore remain invisible until the organisation tests whether its records demonstrate responsibility rather than merely record fragments of activity.

The article makes three contributions. First, it identifies the Operational Responsibility Gap between governance obligations and event-level responsibility evidence. Second, it distinguishes responsibility governance from the infrastructure needed to make responsibility reconstructable. Third, it proposes a bounded conceptual architecture through which responsibility may be represented, accepted, evidenced, assessed, and relied upon across organisational boundaries.

The proposed architecture is relevant to organisations that deploy, procure, regulate, insure, audit, or otherwise rely upon AI-assisted decisions. Its practical purpose is to help those parties determine who held authority, who accepted responsibility, what action followed, what evidence supports the record, how the claim was assessed, and what standing may currently be relied upon.

Responsibility Infrastructure is presented as complementary to governance, law, auditing, provenance, and existing operational systems. The paper asks when those arrangements may require an additional evidential structure because responsibility must be demonstrated to a regulator, court, affected party, or independent reviewer. It first defines its method and scope, then reviews relevant governance frameworks and literature, develops the Operational Responsibility Gap, and sets out the proposed architecture and its limitations.

The minimum public interoperability requirements associated with this architecture are published separately as RI-PROTOCOL-CORE-01, Responsibility Infrastructure Protocol. The protocol defines minimum interoperability requirements rather than internal implementation design.

The practical purpose of this paper is therefore not only to define the proposed problem, but to make it testable. An organisation may examine a bounded operational event, determine whether responsibility can already be reconstructed from existing records, introduce only the minimum additional responsibility-record requirements needed by the hypothesis, and compare the resulting evidential completeness, reconstruction burden, and suitability for independent reliance. In this sense, the paper opens a programme of empirical testing rather than presenting Responsibility Infrastructure as an already validated institutional solution.

2. Method and Scope

This paper uses a conceptual and interdisciplinary method. It compares selected primary AI governance instruments, management-system standards, ethical frameworks, and developer governance documents; reviews literature on responsibility attribution, distributed agency, accountability, auditing, reviewability, and

provenance; and develops an architectural response to the evidential problem identified through that comparison.

The source selection is purposive rather than systematic. Materials were selected to represent structurally different forms of governance and were examined against one common question: do they require or produce decision-level evidence showing authority, acceptance, action, evidential support, assessment, and a claim capable of independent evaluation or reliance? The method is therefore designed to clarify a conceptual boundary, not to measure the prevalence of the gap across sectors.

The paper concerns consequential AI-assisted decisions in which an AI system provides an input, recommendation, or output that is incorporated into a decision by or on behalf of an organisation. It does not evaluate a specific implementation or prescribe databases, interfaces, APIs, workflows, or commercial models.

The analysis does not provide legal advice and does not claim that Responsibility Infrastructure establishes compliance with the EU AI Act or any other legal instrument. References to law and standards illustrate the distinction between governance obligations and event-level responsibility evidence.

The primary unit of analysis in this paper is a responsibility claim concerning a defined operational event or decision, together with the authority, allocation, acceptance, agency, action, evidence, assessment, and standing required to reconstruct that claim.

3. The Rise of AI Governance

The past several years have witnessed a proliferation of AI governance frameworks at the international, national, and organisational levels. This section reviews the most significant frameworks and identifies their shared objective.

3.1 Regulatory Frameworks

The EU AI Act (Regulation (EU) 2024/1689) is the first comprehensive legal framework for AI. It establishes a risk-based classification system, with ‘unacceptable risk’ applications prohibited, ‘high-risk’ systems subject to requirements including risk management, data governance, technical documentation, and human oversight, and limited-risk systems subject to transparency obligations (European Union 2024). Its obligations apply on a staggered timeline, as amended by Regulation (EU) 2026/1744: prohibited-practice provisions took effect on 2 February 2025; governance and general-purpose AI obligations became applicable on 2 August 2025; Article 50 transparency obligations apply from 2 August 2026; requirements for high-risk systems classified under Article 6(2) and Annex III apply from 2 December 2027; and requirements for systems classified under Article 6(1) and Annex I apply from 2 August 2028 (European Union 2024, 2026). These dates establish the paper’s timeliness rather than provide a compliance timetable, and readers should verify the current legal position before relying on them.

Several provisions bear on decision-level responsibility without fully addressing it. Article 12 requires automatic logging capabilities supporting traceability and monitoring; such logs may provide important evidence about system activity, but they do not by themselves establish a complete responsibility claim for an individual decision. Article 13 requires that high-risk systems be designed to ensure transparency sufficient for deployers to interpret system outputs. Article 14 requires that high-risk systems be designed for effective human oversight; system design for oversight is distinct from evidence that a particular person possessed authority, understood the relevant output, intervened where necessary, and exercised judgement in a specific case. Article 17 creates significant organisational quality-management duties; a functioning quality-management system may demonstrate that appropriate organisational processes exist without, by itself, producing a complete responsibility record for every consequential decision (European Union 2024).

Two further provisions are directly relevant. Article 26 requires deployers of high-risk AI systems to assign human oversight to natural persons who have the necessary competence, training, authority, and support (European Union 2024). This raises a question the Act does not itself answer: how can an organisation later demonstrate that the assigned individual possessed those attributes and actually exercised the relevant authority in the decision under scrutiny? Article 27 requires certain deployers — bodies governed by public law, private entities providing public services, and deployers of specified credit-scoring and insurance-pricing systems — to conduct a fundamental rights impact assessment before first use, describing the system’s intended purpose, the period and frequency of use, and the persons likely to be affected (European Union

2024). Impact assessments of this kind describe intended processes, risks, and oversight arrangements before deployment; decision-level responsibility evidence, by contrast, concerns whether those arrangements were subsequently exercised in practice for a specific decision. Taken together, these provisions are necessary governance inputs, but the paper identifies no generally accepted, interoperable mechanism for translating them into verifiable responsibility claims for individual AI-assisted decisions.

Responsibility Infrastructure, as proposed in this paper, is not presented as a legally recognised means of demonstrating conformity with the EU AI Act, nor as a substitute for conformity assessment, harmonised standards, regulatory supervision, or legal interpretation. Its proposed role is narrower: to support the creation of responsibility evidence that may assist organisations and third parties in examining whether governance and oversight arrangements were exercised in particular cases.

The NIST AI Risk Management Framework (AI RMF 1.0), published in January 2023, provides a voluntary framework organised around four functions: Govern, Map, Measure, and Manage (NIST 2023). The Govern function "cultivates and implements a culture of risk management within organizations designing, developing, deploying, evaluating, or acquiring AI systems" and "outlines processes, documents, and organizational schemes that anticipate, identify, and manage the risks a system can pose" (NIST 2023, p. 21). The framework addresses the full AI lifecycle and emphasises trustworthiness characteristics including validity, safety, security, accountability, transparency, explainability, privacy enhancement, and fairness. The NIST AI RMF is explicitly designed to be a cross-cutting governance framework that informs and is infused throughout the other functions. It does not, however, specify how responsibility for individual AI-assisted decisions should be captured, evidenced, or independently verified.

3.2 Standards

ISO/IEC 42001:2023 is the world's first international standard for AI management systems. It "specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organizations" (ISO 2023). Built on the Plan-Do-Check-Act methodology, the standard helps organisations define responsibilities for AI use, identify and assess AI-related risks, ensure transparency and accountability, manage data quality, and monitor AI systems throughout their lifecycle (ISO 2024). As a management system standard, ISO/IEC 42001 provides a structured way to manage AI-related risks at the organisational level. It does not prescribe mechanisms for producing verifiable responsibility claims at the decision level. The related standard ISO/IEC 42006:2025 specifies requirements for bodies that audit and certify AI management systems (ISO 2025), establishing institutional infrastructure for organisational certification — but not for decision-level responsibility verification.

3.3 Ethical Principles

The OECD AI Principles, adopted in 2019 and updated in 2024, provide values-based principles for responsible AI, including inclusive growth, human-centred values and fairness, transparency and explainability, robustness and security, and accountability (OECD 2024). The principles have been adopted by over forty countries and serve as a reference point for national AI strategies.

The UNESCO Recommendation on the Ethics of Artificial Intelligence, adopted by 193 Member States in November 2021, is the first global standard on AI ethics. It is grounded in the protection of human rights and dignity, with core principles including transparency, fairness, environmental sustainability, and human oversight of AI systems (UNESCO 2021). The Recommendation's extensive Policy Action Areas allow policymakers to translate principles into action across domains including data governance, environment, gender, education, health, and social wellbeing (UNESCO 2024).

3.4 Frontier AI Developer Frameworks

Beyond governmental and intergovernmental frameworks, frontier AI developers have published their own governance documents. Anthropic's Responsible Scaling Policy establishes capability thresholds and safety requirements that scale with model capabilities, committing to internal and external evaluations before deploying increasingly capable systems (Anthropic 2023). OpenAI's Preparedness Framework (Version 2, April 2025) outlines the company's approach to tracking and preparing for frontier capabilities that create risks of severe harm, focusing on biological and chemical capabilities, cybersecurity, and AI self-improvement, with explicit thresholds and safeguard requirements (OpenAI 2025). Google's Secure AI Framework (SAF) provides a structured approach to AI security governance across six elements: expand the security cycle to include AI-

specific threats; extend security evaluations to the AI ecosystem; adapt security protections to include AI-powered systems; align AI governance with existing security and privacy governance; promote a holistic approach to security across the AI ecosystem; and adapt red-teaming practices to address AI-specific threats (Google 2023).

3.5 National Safety Institutes

The UK AI Security Institute (AISI), originally established as the AI Safety Institute in 2023 and renamed in February 2025, is a government organisation dedicated to AI safety and security research. Its mission is to equip governments with a scientific understanding of the risks posed by advanced AI. AISI has conducted evaluations of frontier AI systems since November 2023 across domains including cyber, chemistry and biology capabilities, safeguards, and societal impacts (AISI 2025). AISI's research includes work on the "Loss of Oversight" — mapping how current AI oversight rests on foundations likely to erode as systems become more complex (AISI 2026).

3.6 Common Objective

Despite their differences in scope, authority, and approach, these frameworks share a common objective: they help organisations govern AI. They define principles, establish processes, specify documentation requirements, and create oversight structures. They address how AI systems should be designed, deployed, monitored, and governed. This is important and necessary work. However, as the following sections argue, governance and demonstrable responsibility are not equivalent, and the latter requires additional infrastructure that current frameworks do not provide. Table 1 summarises the comparison.

The 'question outside current scope' framing below should be read as a division of labour rather than a deficiency. These frameworks were not designed to answer every decision-level responsibility question, and their not doing so is not itself a shortcoming of the frameworks.

Framework	Primary level	Main contribution	Question outside current scope
EU AI Act	Legal / system / organisational	Binding duties (logging, oversight, quality management)	How was responsibility exercised in a specific decision?
NIST AI RMF	Organisational / system	Risk-management structure	How is individual decision responsibility demonstrated?
ISO/IEC 42001	Management system	Organisational AI controls	How does control translate into a decision-level claim?
OECD / UNESCO	Principles / policy	Normative direction	How are principles evidenced in operation?
Developer frameworks	Organisational / frontier safety	Internal thresholds and commitments	Who held responsibility for a particular deployment decision?

Table 1: Governance frameworks compared against the responsibility-evidence question each leaves open.

4. The Responsibility Gap and Algorithmic Accountability: Literature Review

The concept of a "responsibility gap" in the context of AI and autonomous systems has been the subject of significant philosophical and ethical debate, alongside adjacent bodies of work on algorithmic accountability, auditing, and socio-technical documentation. This section reviews four streams of that literature — moral and philosophical responsibility gaps; role responsibility, control, and distributed agency; algorithmic accountability, auditing, and reviewability; and provenance, documentation, and socio-technical records — and positions the present paper's contribution against each.

4.1 Moral and Philosophical Responsibility Gaps

The term "responsibility gap" was introduced by Andreas Matthias in a landmark 2004 paper. Matthias argued that autonomous, learning machines create a new situation in which "the manufacturer/operator of the machine is in principle not capable of predicting the future machine behaviour any more, and thus cannot be held morally responsible or liable for it" (Matthias 2004, p. 175). Matthias framed this as a dilemma: either

society rejects the use of learning machines whose behaviour cannot be predicted, or it faces "a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription" (Matthias 2004, p. 175).

This formulation has generated extensive philosophical debate. Sparrow (2007) extended the argument to autonomous weapons systems, arguing that responsibility gaps arise when no human agent can be held morally responsible for outcomes caused by autonomous machines. Vallor and Vierkant (2024) reframed the responsibility gap as a "vulnerability gap," arguing that the traditional focus on the epistemic and control conditions of moral responsibility misses a deeper relational asymmetry between human agents and AI systems, and describe it as "a core challenge for the effective governance of, and trust in, AI and autonomous systems" (Vallor & Vierkant 2024).

Not all scholars accept that a responsibility gap exists. Tigard (2020) argued that "there is no techno-responsibility gap," contending that moral responsibility is a dynamic and flexible process that can encompass emerging technological entities. Veluwenkamp and Hindriks (2024) went further, arguing that the concept of a responsibility gap is "incoherent, misguided, and diverts attention from the core issue" of control gaps, which they define as discrepancies between the causal control an agent exercises and the moral control it should possess (Veluwenkamp & Hindriks 2024). Danaher (2022) offered a different perspective, arguing that some techno-responsibility gaps are "virtuous" because they enable a reduced-cost form of delegation and allow societies to avoid the burden of tragic moral choices. Coeckelbergh (2019) approaches the same territory from a relational angle, arguing that responsibility attribution for AI systems is bound up with what explainability can plausibly deliver — a point this paper returns to in Section 8, where evidence assembly is distinguished from the underlying decision's justifiability.

4.2 Role Responsibility, Control, and Distributed Agency

A related stream examines how responsibility is distributed across roles rather than located in a single agent. Nyholm (2017) examined the attribution of agency to automated systems, noting that "much of the focus has been on locating potential responsibility-gaps, viz. cases where it is unclear or indeterminate who is morally or legally responsible for some outcome or event" (Nyholm 2017). Porter et al. (2024) develop this further, arguing that responsibility for AI-assisted outcomes is typically "unravelled" across a chain of actors — developers, deployers, operators, and oversight bodies — rather than held by any single party, and that governance frameworks which assume a single responsible agent may misdescribe how responsibility actually functions in practice. Stahl (2023) situates this within the broader project of embedding responsibility in intelligent systems, arguing that responsible-AI ecosystems require institutional as well as individual mechanisms. These accounts inform the paper's own move, in Section 5, away from a purely individual conception of responsibility toward a specific agent, role, or body.

4.3 Algorithmic Accountability, Auditing, and Reviewability

A parallel body of literature focuses on algorithmic accountability and auditing as practical mechanisms for addressing AI system risks. Busuioc (2020) examined accountability challenges from a public administration perspective, mapping the implications of AI systems for public accountability. Raji et al. (2020) introduced a framework for end-to-end internal algorithmic auditing to "close the accountability gap" in AI system development; as Raji and colleagues acknowledge, auditing frameworks nonetheless face significant challenges, since audits "remain poorly defined" and lack "widely used standards or regulatory frameworks" (Costanza-Chock et al. 2022). Birhane et al. (2024) assessed the state of AI auditing more critically, finding that "only a subset of AI audit studies translate to desired accountability outcomes." Cobbe, Lee, and Singh (2021) introduced the concept of "reviewability" as a framework for improving the accountability of automated decision-making, drawing on an understanding of such systems as socio-technical processes — a framing closely related to, but distinct from, the evidential and recognition concerns of this paper. Mökander and Floridi (2023) examined ethics-based auditing as a governance mechanism, acknowledging that "important aspects of EBA… have yet to be substantiated by empirical research." Terzis, Veale, and Gaumann (2024) analysed the emerging political economy of algorithmic audits under regulatory mandates such as the EU's Digital Services Act.

Many conventional audits are periodic or system-level assessments. Even where monitoring is continuous, the resulting records do not necessarily establish the allocation, acceptance, authority, and exercise of responsibility for each consequential decision — the gap this paper terms operational rather than moral (Section 7).

4.4 Provenance, Documentation, and Socio-Technical Records

A further stream concerns how socio-technical systems document their own operation. Moss et al. (2022) argue for a relational approach to algorithmic accountability and assessment documentation, treating accountability as an ongoing relationship between an organisation and those affected rather than a static artefact produced once and filed away. This relational framing is compatible with the present paper's argument but operates at a different level: Moss et al. are concerned with what documentation should represent about a socio-technical relationship, while this paper is concerned with the narrower question of how a specific decision-level claim can be captured, evidenced, and independently assessed. Provenance mechanisms such as model cards (Mitchell et al. 2019), datasheets for datasets (Gebru et al. 2021), and the W3C PROV standard (W3C 2013) establish the origin, transformation, or lineage of data, models, and content; they do not by themselves establish which agent accepted operational authority for the decision in which those artefacts were used, a distinction developed further in Section 8.5.

4.5 Positioning the Present Contribution

The literature reviewed here already provides substantial accounts of responsibility attribution, distributed agency, auditing, reviewability, provenance, and organisational accountability. Related concepts such as answerability, contestability, meaningful human control, assurance, conformity assessment, and institutional trust also cover parts of the same terrain. This paper does not claim an entirely disconnected field. Its narrower claim is that the production of interoperable, decision-level responsibility claims that are attributable, evidenced, and independently assessable remains comparatively underdeveloped across these literatures.

The contribution is therefore architectural and operational rather than philosophical. It accepts that responsibility may be attributable in principle and asks a further question: what common structure, if any, is needed to make that responsibility demonstrable and portable enough for an external party to assess and rely upon?

That claim is only as strong as its worked contrast against the closest existing proposal, so it is stated here concretely rather than left as a general assertion. Cobbe, Lee, and Singh's (2021) reviewability framework requires that an automated decision-making system be designed so that its process, inputs, and legal and organisational basis can be reconstructed and reviewed after the fact. Applied to a specific AI-assisted hiring decision, a reviewability-compliant system could show what process was followed, what data was used, and what rules governed the decision. It would not, on its own, require that any specific person formally accepted responsibility for that decision, distinguish an agent who merely reviewed the output from one who exercised authority over it, or produce a record capable of being independently recognised by a third party as evidencing that acceptance. Reviewability targets the reconstructability of the process; the Operational Responsibility Gap targets the reconstructability of the claim of responsibility for that process, including whether the named individual had the authority and practical capacity described in Section 5.4. The two are complementary rather than duplicative: a fully reviewable system could still leave the Operational Responsibility Gap open, and closing the gap does not by itself make a system reviewable in Cobbe, Lee, and Singh's sense. This is offered as one worked comparison rather than a claim that the same relationship holds against every framework discussed above; the paper does not attempt an equivalent comparison for each.

5. Governance and Demonstrable Responsibility Serve Different Functions

This section develops the paper's central distinction. Governance directs and controls organisational activity; demonstrable responsibility concerns whether responsibility for a particular event can be reconstructed and assessed. The two functions overlap, but they are not interchangeable.

5.1 Governance

Governance refers to the systems, structures, policies, and processes by which an organisation directs and controls its activities. In the AI context, governance encompasses the frameworks, principles, and risk management processes that organisations use to ensure their AI systems are trustworthy, safe, fair, and compliant with applicable regulations. Governance operates at the policy level. It answers the question: How should the organisation govern AI?

The NIST AI RMF's Govern function exemplifies this level. It "cultivates and implements a culture of risk management" and "outlines processes, documents, and organizational schemes that anticipate, identify, and manage the risks a system can pose" (NIST 2023, p. 21). ISO/IEC 42001 similarly establishes requirements for an AI management system that operates at the organisational level, defining policies, objectives, and processes for responsible AI development and use (ISO 2023).

5.2 Responsibility and Demonstrable Responsibility

Responsibility, as used in this paper, denotes an attributable relationship between an agent, an assigned duty, the authority to act, and an expectation of answerability. This relationship may exist independently of whether it can later be evidenced: an agent can hold responsibility for a decision even where the record of that responsibility has been lost, was never created, or is contested. The paper's subject is narrower than responsibility in the abstract. Demonstrable responsibility refers to the additional capacity to support a responsibility claim — an assertion that a particular agent or body held or exercised responsibility — with evidence sufficient for independent assessment by another party. A recognised responsibility claim is a responsibility claim that has passed an authorised recognition process and carries an auditable standing as a result (see Section 7).

This distinction is not merely theoretical. Consider a healthcare scenario in which an AI-assisted clinical decision support system recommends a treatment plan that results in an adverse outcome. The organisation's governance documentation may demonstrate that the system was properly validated, that clinicians were trained in its use, and that appropriate oversight structures were in place. But when a regulator or court asks who held responsibility for the specific treatment decision, what authority they were operating under, and what evidence demonstrates that responsibility was properly exercised, governance documentation alone may be insufficient — not because responsibility did not exist, but because it was not made demonstrable.

5.3 Conceptual Comparison

The distinction between governance and demonstrable responsibility can be expressed across several dimensions:

Dimension	Governance	Demonstrable responsibility
Primary level	Organisation / system	Specific decision or responsibility claim
Scope	Policies, systems, and organisational controls	Specific agent, role, or body
Core output	Framework	Attributable, evidenced responsibility claim
Main function	Direction, oversight, and risk management	Demonstration and independent assessment
Temporal focus	Ex ante and ongoing	Contemporaneous capture; retrospective assessment when challenged
Primary question	How should the organisation govern AI?	Who held and exercised relevant authority in this case?
Possible result	Compliance or organisational assurance	Evidential standing subject to applicable process
Legal effect	Depends on law and context	Does not itself determine legal liability

Table 2: Conceptual distinction between governance and demonstrable responsibility.

A responsibility record may contain multiple concurrent, bounded, overlapping, or contested responsibility claims. The architecture does not presume that responsibility is singular, nor that every event can legitimately be reduced to one responsible person. For example, one hiring event may contain distinct claims concerning a vendor's model performance, an employer's deployment controls, and a hiring manager's exercise of decision authority. Those claims may be accepted, disputed, assessed, or recognised differently.

This distinction has practical consequences. An organisation can be fully compliant with its governance obligations — having implemented the NIST AI RMF, achieved ISO/IEC 42001 certification, and established comprehensive AI policies — and still be unable to demonstrate who held responsibility for a specific AI-

assisted decision when challenged. Compliance with governance frameworks demonstrates organisational intent and process. It does not, by itself, produce decision-level evidence of responsibility.

5.4 Allocation, Agency, and Accountability

A responsibility architecture that stops at naming an individual risks a specific failure mode: a named person accepted responsibility, therefore responsibility has been successfully established. That inference is incomplete. A person may be assigned responsibility without having the authority to change the outcome, the information needed to decide, the resources or time to intervene, the competence to understand the AI system's output, or protection from organisational pressure to accept a decision they did not meaningfully control. A credible responsibility claim requires three elements together: allocation — the responsibility was clearly assigned; agency — the agent had sufficient authority, competence, information, time, and practical ability to act; and accountability — the agent was answerable for how that authority was exercised. Responsibility claims are structurally incomplete where allocation is not accompanied by meaningful agency and answerability. Naming a person without establishing their authority and capacity risks converting responsibility infrastructure into a mechanism for retrospective scapegoating rather than a mechanism for genuine accountability.

6. Why Existing Audit Trails May Be Insufficient

One possible objection is that existing audit trails — workflow management systems, version control platforms, ticketing systems, and communication logs — already provide the infrastructure for reconstructing responsibility. This section argues that they may not, while acknowledging what they do provide.

6.1 Activity versus Responsibility

Most enterprise systems record activities: who performed what action, when, and within what workflow. GitHub records commits, pull requests, and code reviews. Jira records ticket assignments, status changes, and comments. ServiceNow records incident reports and service requests. Microsoft Teams and Slack record messages and file shares. Email systems record communications. HR systems record employment data and organisational roles. These systems may provide important components of a responsibility record — timestamps, actions, identities, approvals, communications, model versions, data lineage, and assigned roles. The problem is not that they contain no useful evidence. The problem is that they do not necessarily combine that evidence into a standard responsibility claim, evidence of acceptance, evidence of authority, a coherent evidence package, an independently assessed standing, or a portable receipt for resolution. A Jira ticket may show that an engineer was assigned to review an AI model's output. It does not show that the engineer accepted responsibility for the decision in a formal sense, that the acceptance was recorded in a way that can be independently verified, or that the resulting responsibility claim can be communicated to a third party with sufficient evidence.

6.2 Fragmentation and Non-Interoperability

Existing audit trails are fragmented across multiple systems with different data models, access controls, retention policies, and formats. A single AI-assisted decision may involve artefacts distributed across a version control system (code), a model registry (model version), a feature store (input data), a monitoring dashboard (model performance), a ticketing system (review process), an email system (approval communication), and an HR system (organisational role). Reconstructing responsibility from these fragments is labour-intensive, error-prone, and dependent on the continued availability and accessibility of each system. These systems are not generally designed for interoperability in responsibility claims, and do not support the concept of a "responsibility receipt" — a portable, verifiable record that a third party can rely upon to confirm that responsibility was properly exercised for a specific decision.

6.3 Temporal Gaps

Audit trails are often designed for operational purposes (incident response, performance monitoring, compliance reporting) rather than for retrospective responsibility evidence. Retention policies may delete logs after months or years, long before a responsibility challenge arises. Access controls may change as personnel leave or organisational structures shift. The systems themselves may be decommissioned, migrated, or reorganised, making historical records difficult or impossible to retrieve. These limitations mean that even

when organisations have extensive audit trails, they may be unable to reconstruct responsibility for a specific AI-assisted decision when challenged months or years later.

6.4 Provenance Is Not Responsibility

Provenance mechanisms establish the origin, transformation, or lineage of data, models, and content. They do not by themselves establish which agent accepted operational authority for the decision in which those artefacts were used, what governance basis that agent operated under, or how their responsibility can be independently recognised by third parties. This distinction is developed further in Section 8.5.

7. The Operational Responsibility Gap

7.1 Definition

Building on the existing literature's concept of a moral responsibility gap, this paper introduces the concept of an Operational Responsibility Gap:

> *The Operational Responsibility Gap exists where relevant events can be reconstructed, but the associated responsibility claim cannot be demonstrated with sufficiently attributable, authoritative, coherent, and independently assessable evidence.*

This gap is distinct from the moral responsibility gap identified by Matthias (2004) and others. The moral responsibility gap concerns whether moral attribution is possible in principle. The Operational Responsibility Gap concerns whether responsibility can be demonstrated in practice — that is, whether an organisation can produce verifiable evidence that responsibility was properly exercised for a specific AI-assisted decision when challenged by a regulator, court, affected party, or other third party.

7.1.1 Operational, Behavioural, and Cultural Responsibility

Operational responsibility should be distinguished from behavioural and cultural responsibility. Behavioural responsibility concerns whether individuals act with ownership, integrity, and appropriate judgement. Cultural responsibility concerns whether organisational norms encourage accountability, challenge, communication, and follow-through. Operational responsibility concerns a different question: whether the movement and present condition of responsibility can be demonstrated through reconstructable records.

An organisation may possess committed staff and a strong culture of accountability while still lacking the operational infrastructure required to show who held authority, who accepted responsibility, what action occurred, what evidence supports the record, and what standing can now be relied upon. Conversely, an organisation may maintain extensive governance documentation and workflow controls without producing a complete responsibility record for a contested event.

Responsibility Infrastructure is therefore not proposed as a substitute for leadership, ethics, professional judgement, or organisational culture. Nor is it merely another checklist, workflow tool, or governance control. It addresses the evidential and interoperability layer through which responsibility becomes representable, reconstructable, independently assessable, and capable of supporting external reliance.

7.1.2 Diagnosing an Operational Responsibility Gap

Many organisations may experience an Operational Responsibility Gap without recognising it as a distinct problem. Responsibility information is often distributed across policies, workflow systems, email, messaging platforms, audit logs, meeting records, technical logs, and individual recollection. Each source may appear adequate when reviewed separately, while the organisation remains unable to reconstruct the complete responsibility chain for a defined event.

A responsibility gap scan is proposed as a diagnostic assessment through which an organisation can examine whether responsibility for a defined event was explicitly allocated, supported by authority and practical agency, accepted, evidenced, independently assessable, and capable of supporting reliance. The scan does not presume organisational failure and does not assess whether staff possess an appropriate culture of accountability. Its narrower purpose is to determine whether operational records can reconstruct responsibility without depending primarily on inference, memory, or fragmented documentation.

The paper identifies the public diagnostic dimensions relevant to that inquiry, including allocation, authority, agency, acceptance, action, evidence, assessment, reconstruction, and reliance. It does not specify a proprietary scoring model, assessment threshold, interview sequence, sector-specific question set, report template, or remediation method. Those implementation choices require separate development and empirical evaluation.

A responsibility gap scan may therefore serve as the first stage of bounded testing. Its purpose is to determine whether responsibility for a defined event can already be reconstructed through existing systems and, where it cannot, to identify the specific missing conditions before any additional intervention is introduced. The scan should not presume that Responsibility Infrastructure is required. Where existing arrangements already produce a complete, independently assessable, and sufficiently interoperable responsibility record, the hypothesis would not justify additional responsibility-record requirements.

7.1.3 Boundary, Necessity, and Alternative Explanations

The Operational Responsibility Gap is not merely a documentation, provenance, auditability, or workflow gap. Documentation may show that an event was recorded. Provenance may show where information originated and how it changed. Audit logs may show that an action occurred. Governance controls may define who should have acted. Workflow systems may coordinate activity. None of these, separately, necessarily establishes the complete responsibility condition: who held authority, what responsibility was assigned, whether it was accepted, what condition it entered before action, what action followed, what evidence supports the outcome, how the resulting claim was assessed, and what standing may now be relied upon. Responsibility Infrastructure is proposed only where those elements must be represented as one reconstructable and independently assessable chain.

Responsibility Infrastructure is not presumed to be necessary in every organisation or operational context. Where existing systems can already produce a complete, interoperable, and independently reconstructable record of authority, allocation, acceptance, action, evidence, assessment, and current standing without material inference, fragmentation, or loss of provenance, an additional responsibility infrastructure layer may not be required. The need for such infrastructure arises only where those minimum reconstructability and reliance conditions cannot be demonstrated through existing arrangements.

The prevalence and severity of this condition remain empirical questions. A responsibility gap scan is intended to test whether the condition exists before any new infrastructure is proposed, and bounded comparative pilots are needed to determine whether a distinct responsibility layer improves reconstruction, dispute resolution, or external reliance beyond better use of existing records and controls.

7.2 Conditions

The Operational Responsibility Gap arises when any of the following conditions hold. The gap is not necessarily binary: its severity may increase as assignment becomes ambiguous, agency becomes weaker, acceptance becomes implicit, evidence becomes fragmented, authority cannot be demonstrated, records become vulnerable to alteration, verification becomes conflicted, recognised standing becomes unavailable, or external reliance becomes uncertain — a dimensional view that suggests a future measurement instrument rather than a simple present/absent test (see Section 12.2).

1. **Capture failure:** The organisation did not systematically record who accepted responsibility for AI-assisted decisions, under what authority, and with what evidence of due consideration.
2. **Assignment ambiguity:** Responsibility was distributed across multiple actors (developers, operators, decision-makers, oversight bodies) without a clear record of who held primary responsibility for the specific decision.
3. **Evidence fragmentation:** The evidence needed to reconstruct responsibility is distributed across multiple systems, in non-standardised formats, with no mechanism for producing a coherent, verifiable evidence package.
4. **Verification absence:** There is no independent process for verifying responsibility claims — the organisation's own records are the sole basis for the claim, with no external attestation or independent assessment.
5. **Recognition deficit:** There is no mechanism for communicating responsibility claims to third parties in a form that they can recognise and rely upon.

A sixth, cross-cutting condition deserves separate emphasis. Formal assignment and acceptance do not by themselves eliminate the gap where the named agent lacked meaningful authority, resources, competence, information, or practical capacity to affect the outcome (see Section 5.4). A record can look complete — a name, a timestamp, a signature — while concealing weak or absent agency. Formal allocation should therefore not be treated as legitimate responsibility unless the record preserves evidence that the assigned party

possessed the authority and practical agency needed to discharge it. Without those safeguards, formalisation can become a mechanism for responsibility laundering: blame is pushed downward onto visible individuals while structural decision-makers and resource constraints remain obscured. This risk is one reason the paper treats allocation, agency, and accountability as jointly necessary rather than treating assignment alone as sufficient.

Recognised standing, referred to throughout this paper, denotes an auditable evidential status resulting from an authorised recognition process. It does not constitute a judicial or regulatory determination of liability; it refers to an evidential status that may support external reliance, subject to the applicable legal forum and standard of proof.

7.3 Relation to the Moral Responsibility Gap

The Operational Responsibility Gap is related to but distinct from the moral responsibility gap. The moral gap, as formulated by Matthias (2004) and debated by Tigard (2020), Veluwenkamp and Hindriks (2024), Vallor and Vierkant (2024), and others, asks whether moral responsibility can be attributed to any agent when an AI system causes an outcome. The Operational Responsibility Gap assumes, for the sake of argument, that moral responsibility can in principle be attributed — as the literature increasingly suggests it can (Veluwenkamp & Hindriks 2024; Tigard 2020) — and asks a different question: even if moral responsibility can be attributed, can it be demonstrated through verifiable, independently assessable evidence?

This distinction matters because the practical challenge facing organisations, regulators, and courts is not primarily philosophical. It is operational. An organisation deploying an AI-assisted hiring system does not face a purely philosophical question about whether moral responsibility can be attributed. It faces the practical question of whether it can produce evidence, when challenged, that responsibility for a specific hiring decision was properly exercised by a specific person under a specific governance authority.

For example, a conventional record of an AI-assisted hiring decision may show the model score, the applicant outcome, and the name of the hiring manager. A reconstructable responsibility record would additionally show the manager's authority, the responsibility assigned and accepted, the condition of that responsibility before action, the evidence considered, the action taken, the assessment of the resulting claim, and the standing available to an authorised relying party. The difference is not simply more logging; it is a coherent evidential chain capable of independent reconstruction.

8. Responsibility Infrastructure

8.1 Introduction

This section introduces Responsibility Infrastructure (RI) as a proposed architectural response to the Operational Responsibility Gap. RI is defined here as a proposed interoperability and evidential architecture through which responsibility claims can be captured, structured, assessed, communicated, and resolved across organisational boundaries, not as a specific software platform or product. The paper does not claim that RI solves all problems of AI governance or responsibility. Rather, it proposes RI as one possible architectural approach that complements existing governance frameworks by enabling responsibility claims to be reconstructed and independently assessed at the decision level.

The architecture should be applied proportionately to the materiality, risk, contestability, and external-reliance requirements of the event. It does not require every routine action to undergo full independent verification, recognition, registry publication, or external resolution. Those functions become relevant where a responsibility claim must support challenge, cross-organisational exchange, regulatory scrutiny, or reliance by a party beyond the organisation that created the record.

8.2 First Principles

Demonstrable responsibility, in the sense used in this paper, requires several capabilities. Rather than presenting these as a single undifferentiated list, it is more useful to group them into three functional categories, since each category answers a different question and is likely to require different institutional treatment.

The lifecycle adapts familiar issue-assess-publish-resolve patterns found in public-key infrastructure and conformity assessment. The analogy is structural: these systems separate the creation of an artefact or claim, assessment, status publication, and reliance. RI applies that pattern to responsibility claims rather than to cryptographic credentials or conformity certificates.

The analogy should not be read as borrowing PKI's institutional maturity. PKI operates within established roots of trust, root-store policies, and governance arrangements; RI presently does not. Nor can cryptography prove authority, agency, acceptance, or judgement. Whether RI can develop a credible institutional root of trust is therefore part of the bootstrapping problem discussed in Section 9, not something established by the analogy.

Verification evaluates whether the evidence supports a responsibility claim against defined criteria. Recognition determines the evidential status of a verified claim through an authorised process. Standing is the current status produced by recognition and made available for resolution or reliance. A registry publishes or resolves that standing without re-deciding the underlying merits. Reliance remains a decision for the receiving party.

8.2.1 Attribution — who was assigned responsibility, and did they accept it

1. **Capture:** The ability to record, at the time a decision is made or authorised, who is exercising responsibility, for what decision, under what authority, and with what contextual information.
2. **Assignment:** The ability to formally assign responsibility to a specific agent, role, or body, with clear specification of the scope and limits of that responsibility.
3. **Acceptance:** The ability for the assigned agent to formally accept or decline responsibility, creating a record of explicit acceptance rather than implicit attribution.

Acceptance is evidential, not conclusive. A recorded acceptance does not by itself establish meaningful responsibility where authority, competence, information, time, resources, or freedom to refuse were absent. It must therefore be assessed together with the agency conditions described in Section 5.4.

8.2.2 Execution and Evidence — what happened after acceptance, and what supports the record

4. **Response state:** The ability to represent the condition of an accepted responsibility between acceptance and final action — whether it is active, at risk, escalated, or requires intervention. Without a response-state concept, a lifecycle moves directly from acceptance to action and cannot represent a responsibility that has been accepted but is stalled, contested, or deteriorating. The detailed state taxonomy, transition conditions, escalation thresholds, and enforcement logic required by a specific implementation are outside the scope of this paper.
5. **Action:** The ability to record the decision itself, including inputs considered, alternatives evaluated, AI system outputs reviewed, and the reasoning that led to the final decision.
6. **Evidence:** The ability to assemble the captured information into a coherent, standardised evidence package that can be independently reviewed.

8.2.3 Assurance and Reliance — how the claim was assessed, and how another party may rely on it

7. **Verification:** The ability to assess the evidence package against defined criteria. Section 8.4 distinguishes first-party, second-party, and third-party verification; the degree of independence required is proportionate to the claim, its context, and the consequences of error, rather than fixed at a single standard for every claim.
8. **Recognition:** The ability for an authorised process to issue a determination concerning the claim's evidential status, conferring a "standing" that third parties may rely upon.
9. Registry: The ability to publish or resolve recognised status through a designated registry record within the relevant governance scheme, accessible to authorised third parties, with appropriate protections for privacy and commercial confidentiality. In the proposed architecture, this function may be instantiated through a Responsibility Infrastructure Registry, although the registry function described here is architectural rather than implementation-specific.
10. R-Receipt: The ability to issue a resolvable artefact that contains or references the recognised responsibility record, resolves against the registry, displays current standing, and supports external reliance.

Registry publication and R-Receipt resolution are not required for every internal responsibility record. They become relevant where a recognised claim is intended to support current, cross-organisational, or repeatable external reliance and where a resolvable status is needed.

This three-category structure — Attribution, Execution and Evidence, Assurance and Reliance — is intended to carve the lifecycle at its functional joints rather than present an arbitrary sequence: each category corresponds to a distinct question a third party would ask (who is answerable; what happened and on what basis; can the claim be relied upon), and each is plausibly the responsibility of a different institutional actor. The ten finer-grained stages remain useful as an implementation-level vocabulary, summarised in Table 3.

Category	Stages	Question answered
Attribution	Capture → Assignment → Acceptance	Who was assigned responsibility, under what authority, and did they accept it?
Execution and Evidence	Response state → Action → Evidence	What happened after acceptance, what was decided or done, and what evidence supports the record?
Assurance and Reliance	Verification → Recognition → Registry → R-Receipt	How was the claim assessed, what standing resulted, and how may another party resolve or rely upon it?

Table 3: The ten responsibility lifecycle stages grouped into three functional categories.

8.2.4 Worked example: AI-assisted recruitment decision

A candidate is rejected after an AI-assisted screening system assigns a low suitability ranking. The example below illustrates how a responsibility claim might move through the proposed lifecycle. It is hypothetical and does not prescribe a software implementation, assessment method, or legal conclusion.

Stage	Illustrative record
Capture	A defined event is opened for the final hiring decision concerning Candidate A for Role B.
Assignment	The hiring manager is assigned responsibility for the final decision within authority delegated by the employer. Separate claims may concern the vendor's model performance and the HR function's deployment controls.
Acceptance	The hiring manager records acknowledgement of the assignment. This is evidential rather than conclusive and does not establish responsibility without authority and agency.
Response state	The record shows whether the responsibility is active, awaiting further information, escalated, disputed, or otherwise unresolved.
Action	The hiring manager rejects the candidate, records the decision, and identifies the role played by the AI-generated ranking.
Evidence	The record references the model output, relevant candidate information, applicable recruitment policy, override options, communications, and reasons recorded by the decision-maker.
Verification	An independent reviewer evaluates whether the evidence supports the stated claim concerning authority, agency, acceptance, and action.
Recognition	An authorised process assigns a defined evidential status to the verified claim. Recognition does not determine whether the hiring decision was lawful or substantively correct.
Registry	Where current external reliance is required, the recognised status may be published or resolved without re-deciding the underlying merits.
R-Receipt	An authorised relying party may receive a resolvable reference to the recognised record and its current standing.

Existing systems may already contain several of these elements. The proposed intervention becomes relevant only where the elements cannot be connected into a sufficiently reconstructable responsibility claim. A Responsibility Gap Scan should therefore test the existing record first rather than presume that the full lifecycle is required.

8.3 Architectural Layer

RI is conceived as an architectural layer that sits between operational AI systems and independent recognition bodies. It does not replace governance frameworks, AI management systems, or regulatory compliance processes. Instead, it provides the infrastructure needed to translate governance-level expectations into decision-level evidence. Within this layer, distinct functions belong to distinct actors: RI defines the responsibility record and exchange architecture; implementations create and manage records; independent actors verify evidence and determine claims; a registry publishes recognised standing; and relying parties

decide what weight to give that standing. This positioning is important because it avoids implying that a single RI implementation performs every assurance function itself — a point developed further in Section 9.

The architecture can be visualised as a sequence running from governance and law (e.g., the EU AI Act, the NIST AI RMF, ISO/IEC 42001), through operational systems that execute decisions and generate events, through a responsibility record architecture that captures attribution, response state, action, and evidence, to independent assessment that verifies and determines claims, to a registry that publishes current recognised standing, and finally to external reliance, where regulators, courts, auditors, and affected parties decide what weight to give the record. This layered presentation is consistent with broader layered approaches to AI governance (Gasser & Almeida 2017).

8.4 Relationship to Audit and Accountability Mechanisms

RI is related to but distinct from existing audit and accountability mechanisms. RI is not itself an audit mechanism. Algorithmic auditing, as discussed by Raji et al. (2020) and Birhane et al. (2024), focuses on assessing AI systems for compliance with ethical principles, fairness criteria, and regulatory requirements. Audits are typically periodic, system-level assessments conducted after deployment or at specific milestones. RI, by contrast, is intended to support contemporaneous capture of responsibility information as decisions are made, producing evidence packages for individual decisions rather than system-level assessments, and supporting independent recognition of individual responsibility claims rather than organisational certification. This distinction does not diminish the importance of auditing: audits assess systems, RI evidences decisions, and the two are intended to be complementary.

Not every responsibility claim requires the same degree of independence in how it is verified. It is useful to distinguish first-party verification — internal review by the organisation making the claim; second-party verification — review by a customer, contracting body, or relying organisation with a direct stake in the outcome; and third-party verification — review by an independent external body with no stake in either the claim or the organisation making it. First-party verification may support internal assurance. Where a responsibility claim is intended to support independent external reliance, the independence of verification should be proportionate to the claim, its context, and the consequences of error, rather than assumed to require third-party review in every case. Internal records may support first-party assurance, but an interested organisation should not be the sole authority establishing the independent reliability of its own claim — a distinction developed further in Section 9.

8.5 Relationship to Provenance and Traceability

RI should be distinguished from existing provenance and traceability approaches. Model cards (Mitchell et al. 2019), datasheets for datasets (Gebru et al. 2021), and the W3C PROV standard for provenance interchange (W3C 2013) track artefacts and their lineage — the origins, transformations, and dependencies of data, models, and systems. These are valuable tools for system-level transparency. However, provenance tracks what was used and how it was produced; it does not track who accepted authority for a specific decision, what governance basis they operated under, or how their responsibility can be independently recognised by third parties. RI addresses a different layer of the accountability stack: not the lineage of artefacts, but the attribution, acceptance, evidence, and recognition of responsibility for decisions that use those artefacts.

8.6 Privacy and Security Considerations

RI records may contain sensitive personal data (including health, employment, and financial information), commercially confidential information (including model details and internal processes), and information about individual decision-makers. The design of RI must therefore incorporate privacy-preserving principles. Registries should be permissioned, with access restricted to authorised parties, and essential standing and verification metadata may be publicly resolvable while sensitive records and underlying evidence remain access-controlled and subject to selective disclosure. Data minimisation should apply: records should contain only the information necessary to evidence responsibility, not exhaustive decision documentation.

Tamper-evident, append-only structures and cryptographic commitments may support detection of unauthorised alteration after the fact. They do not establish that the original input was true, complete, authorised, or fairly assessed — integrity of the record and correctness of its content are separate properties, and a design that conflates them risks overstating what cryptographic guarantees actually provide. Selective disclosure mechanisms should allow organisations to present responsibility evidence at varying levels of detail

depending on the recipient and the purpose of the inquiry. These considerations are not merely technical; they are essential for ensuring that RI does not itself become a source of privacy harm.

Distributed-ledger technology may strengthen timestamping, provenance, replication, and resistance to retrospective alteration. Those properties may assist with record integrity, but they do not establish that a person held authority, possessed meaningful agency, accepted responsibility freely, exercised appropriate judgement, or supplied sufficient evidence. Nor does decentralised storage determine which institution is authorised to verify a claim, recognise its status, or resolve a dispute. The distinction is therefore between record integrity and responsibility validity: distributed-ledger technology may support the former, while the latter remains dependent on evidence, defined criteria, institutional judgement, and procedural safeguards. Responsibility Infrastructure does not require a blockchain and does not treat decentralisation as a substitute for legitimate assessment or recognition.

8.7 Disclosure

The author originated the Responsibility Infrastructure framework and architectural proposal presented in this paper and is affiliated with an organisation developing a commercial implementation of selected lifecycle mechanisms described in Section 8.2. That implementation treats its event models, state-transition logic, evidence-assessment methods, trust-resolution logic, and other implementation-level mechanics as proprietary. This paper discloses the problem definition, architectural functions, and minimum conceptual architecture needed for evaluation and interoperability; it deliberately does not reproduce proprietary implementation mechanics. The argument is presented as an academic proposal rather than as an evaluation of a proprietary implementation. The commercial and trademark dimension of this disclosure, including the objection that it may sit in tension with Section 9's separation argument, is addressed directly in Section 9.6. See also the full competing-interests statement in the Declarations.

9. Structural Separation

9.1 The Principle of Separation

A central design principle of the proposed RI architecture is that the party whose conduct or claim is being assessed should not be the sole authority deciding whether that claim is valid for independent reliance. The principle is familiar across institutional settings: auditors do not prepare the accounts they audit; certification bodies assess organisations they do not control; and courts separate the parties presenting claims from the adjudicator. RI adopts this as an architectural separation of functions, not as a claim to constitutional or public authority.

9.2 Application to Responsibility Infrastructure

Applied to RI, implementation creates and manages responsibility records; verification evaluates evidence against defined criteria; recognition determines standing through an authorised process; registry publication records recognised standing without re-deciding the merits; resolution communicates current standing; and reliance remains a decision for the third party. The architecture therefore rejects the proposition that successful internal execution is sufficient, by itself, to establish an independently reliable responsibility claim.

The required degree of institutional separation is proportional to the reliance being claimed. It need not always require separate legal entities: functional firewalls, independent decision-makers, recusal, published criteria, review and appeal, external accreditation, or independent adjudication may provide relevant safeguards. What matters is whether conflicts inherent in self-assessment are controlled sufficiently for the intended reliance.

9.3 Institutional Analogies

Comparable structures illustrate the principle without proving RI's legitimacy. Courts separate claimants from adjudicators; public registries generally record statuses created through legally defined processes rather than deciding every underlying merits question anew; and conformity-assessment systems distinguish the assessed organisation, the certification body, and oversight or accreditation functions. These analogies show why functional separation can matter. They do not confer equivalent authority or institutional maturity on RI.

9.4 Why Separation Matters for Trust

For RI, the practical implication is limited but important: where the same organisation deploys a system, captures responsibility information, assesses its own evidence, determines standing, and publishes the result, the resulting claim should not be treated as independently reliable merely because the internal process executed correctly. Internal operation can show procedural consistency; independent reliance requires additional grounds for trust. The stronger the reliance claimed, the stronger the case for institutional pluralism and external scrutiny.

9.5 The Limits of the Separation Argument

The separation principle does not establish that separation is sufficient for trust. Courts, registries, certification systems, and mature PKI arrangements rely on legal mandates, governance arrangements, accumulated track records, or established roots of trust that RI does not presently possess. A newly created recognition body cannot acquire legitimacy merely by declaring itself independent or authoritative. Trust would have to emerge through transparent rules, genuine procedural independence, consistent determinations, challenge mechanisms, and demonstrated reliance by external parties.

That bootstrapping problem applies directly to the author's own implementation. At the time of writing, related intellectual property is held by La Touche Holdings Ltd, while La Touche Infrastructure Ltd is developing the commercial implementation. Independent verification, recognition, and registry functions have not yet been established in practice. Outputs produced entirely within that controlled structure should therefore be treated as first-party, or at most limited second-party, assurance under Section 8.4, not as evidence that the target separation has already been achieved.

This paper does not attempt to solve the bootstrapping problem. Candidate pathways, offered as possibilities rather than proven solutions, include bounded regulatory sandboxes; independently observed pilot proceedings; arm's-length academic observation; third-party governance audits; staged reliance commitments; transparent publication of methods and decisions; and external appeal or review mechanisms. Each would need to be tested rather than assumed to work. Founding governance instruments may document intended transitions, conflict rules, recusal requirements, external participation, or transfer of authority, but such instruments are commitments about institutional design rather than evidence that independence, legitimacy, or public authority has already been achieved.

9.6 Current Ownership and the Separation Objection

The current ownership structure creates a legitimate objection. Responsibility Infrastructure, Responsibility Infrastructure Registry, and R-Receipt are marks owned by La Touche Holdings Ltd, while related implementation activity remains within affiliated entities. If the architecture argues against self-assessment, ownership and control of the field's principal identifiers cannot be treated as irrelevant. The issue is not whether trademark ownership is inherently incompatible with an open protocol; it is whether control of marks, conformance claims, implementation, verification, recognition, and registry functions becomes concentrated in a way that exceeds the independence the resulting claims can support.

The paper treats this as a present governance constraint, not as a problem solved by disclosure.

Trademark ownership and protocol availability are distinct questions. The minimum interoperability requirements are published separately as RI-PROTOCOL-CORE-01 so that independent parties can inspect and, subject to its published terms, implement the protocol without adopting a particular commercial implementation. This does not resolve the governance objection: control of marks can still influence category claims and conformance designations. It simply separates ownership of identifiers from availability of the underlying technical requirements.

The objection is strongest at the present bootstrapping stage. There is, at the time of writing, no independent recognition body, no separately operated registry function, and no accreditation function distinct from the author's affiliated entities. Sections 9.1-9.4 describe a target architecture, not the current state of the author's implementation. The distinction is substantive: a proposed separation model cannot be used as evidence that separation already exists.

The practical consequence is a limit on the reliance the current implementation can credibly support. A responsibility claim verified, recognised, and registered entirely within a structure controlled by affiliated

entities should be treated as first-party assurance, whatever labels are applied internally. Greater claims of independence would require external governance, accreditation, or genuinely third-party operation that can be observed in practice. Readers should therefore treat the identity and independence of the verifying and recognising parties as material to the weight any resulting standing can bear.

10. Relationship to Existing AI Governance

10.1 Complementary, Not Competing

RI is not presented as a replacement for the EU AI Act, the NIST AI RMF, ISO/IEC 42001, or any other governance framework. Existing governance frameworks establish what organisations should do; RI is proposed as infrastructure for demonstrating that they did it for specific decisions, with evidence that can be independently assessed. The relationship can be expressed as a layered model:

Adoption is most likely where the expected value of improved reconstruction, dispute handling, procurement assurance, insurance assessment, or regulatory engagement exceeds the additional recording burden. A responsibility record may assist a court, regulator, insurer, auditor, or other receiving party by preserving structured evidence; its admissibility, weight, and legal effect remain matters for the applicable legal and institutional process.

Layer	Function
Governance and law	Define duties, expectations, and controls
Operational systems	Execute decisions and generate events
Responsibility record architecture	Capture attribution, response state, action, and evidence
Independent assessment	Verify and determine claims
Registry and resolution	Publish current recognised standing
External reliance	Regulators, courts, auditors, and affected parties decide what weight to give the record

Table 4: Layered model of AI governance, responsibility record architecture, and recognition.

Each layer in Table 4 is complete for the function it performs. The responsibility-record layer addresses a different question from the governance, operational, assessment, registry, and reliance layers rather than identifying a defect in those layers.

A responsibility gap scan would not duplicate a culture survey, governance review, process audit, or workflow assessment. It would test whether those arrangements produce a sufficiently complete and reconstructable responsibility record for a defined operational event. The distinction is important because an organisation can demonstrate responsible intentions, extensive controls, and substantial activity while remaining unable to prove the operational chain of responsibility when a claim is challenged.

Applied to the frameworks reviewed in Section 3: RI is intended to support demonstrating that oversight required by EU AI Act Articles 12, 14, 17, and 26 was exercised in specific decisions, not merely that oversight processes exist; to evidence that NIST AI RMF governance processes were followed at the decision level; and to operationalise ISO/IEC 42001's requirement that organisations define responsibilities for AI use by capturing, evidencing, and independently verifying the resulting claims. RI does not replace regulatory obligations, ethical principles, risk management processes, algorithmic auditing, or impact assessments — organisations must still comply with, adhere to, implement, and conduct each of these respectively. RI is proposed to complement these mechanisms by adding a decision-level responsibility layer that they do not, on their own, provide.

11. Implications

This section develops three sectors in some depth — recruitment, healthcare, and public administration — chosen because each combines consequential individual impact, foreseeable external challenge, and responsibility plausibly distributed across several actors. Table 5, following, summarises four further sectors more briefly to preserve breadth without repeating the same argument seven times.

11.1 Recruitment

AI-assisted recruitment tools are used for candidate screening, interview assessment, and hiring decisions, and are a strong fit for the argument developed here because AI use may affect access to employment; because human oversight under regimes such as EU AI Act Article 26 must involve real authority rather than nominal sign-off; because candidate challenges may require reconstruction of a specific decision months after it was made; and because responsibility can be distributed across a technology vendor, an HR team, and an individual hiring manager in ways that make a single named decision-maker an incomplete answer on its own. When a candidate challenges a hiring decision, the organisation may need to demonstrate that the AI system was used appropriately, that a human decision-maker held meaningful authority and accepted responsibility for the final decision, and that the decision was made in compliance with anti-discrimination law and organisational policy. RI is proposed to support such demonstrations by evidencing allocation, agency, and acceptance across the vendor, the HR function, and the hiring manager, rather than assuming responsibility sits with whichever name appears last in the process.

11.2 Healthcare

In healthcare, AI-assisted clinical decision support systems are increasingly used for diagnosis, treatment recommendations, and resource allocation. The sector is a strong fit because clinicians, healthcare providers, and technology suppliers may hold different and overlapping responsibilities; because evidential records must protect sensitive patient data while still supporting external review; because responsibility cannot be reduced to naming the final treating clinician where an AI system's recommendation materially shaped the decision; and because professional clinical judgement and AI-system influence need to be evidentially separable rather than merged into a single undifferentiated record. Habli, Lawton, and Porter (2020) found that "none of the actors in the model robustly fulfil the traditional conditions of moral accountability for the decisions of an artificial intelligence system," a finding consistent with this paper's argument that allocation alone, without agency and evidenced acceptance, is an incomplete responsibility claim (Section 5.4). RI is proposed to support evidence that a specific clinician reviewed the AI system's recommendation, considered relevant clinical factors, and exercised judgement under a defined governance authority, while keeping the underlying sensitive data access-controlled (Section 8.6).

11.3 Public Administration

Government use of AI systems for benefits determination, immigration decisions, law enforcement, and other public functions raises particularly important accountability questions, and is a strong fit for this argument because such decisions may affect rights and entitlements directly; because reasons, review, and procedural fairness are already legally significant in this domain; because external challenge — judicial review, ombudsman complaint, tribunal appeal — is foreseeable rather than exceptional; and because institutional authority for a specific decision must often be demonstrable to a reviewing body regardless of whether AI was involved. Loi and Spielkamp (2021) analysed sixteen guideline documents about AI use in the public sector and concluded that "while some guidelines refer to processes that amount to auditing, it seems that the debate would benefit from more clarity about the nature of the entitlement of auditors and the goals of auditing" (Loi & Spielkamp 2021). RI is proposed to give public sector organisations infrastructure for producing decision-level responsibility evidence that could support this kind of external review, though the paper does not claim that the architecture has been tested in a public-sector context (Section 12.1).

11.4 Further Sectors

Table 5 summarises four further sectors where the same argument applies with sector-specific detail; each is discussed only briefly here to avoid repeating the sector-by-sector structure of Sections 11.1–11.3.

Sector	Why RI is relevant	Distinguishing feature
Financial services	Credit decisions, fraud detection, algorithmic trading, and risk assessment are subject to regulatory expectations of fairness and transparency	Regulatory examination often follows a specific disputed decision rather than a system-wide review

Sector	Why RI is relevant	Distinguishing feature
Critical infrastructure	AI-assisted decisions affecting energy, transport, and telecommunications may require incident investigation and public accountability	AISI's Loss of Oversight research suggests current oversight methods may become less effective as system complexity increases (AISI 2026)
Autonomous systems	Self-driving vehicles and autonomous industrial systems raise acute responsibility questions historically framed as the moral responsibility gap (Matthias 2004; Sparrow 2007)	Even where moral responsibility can be attributed (Veluwenkamp & Hindriks 2024), demonstrating it requires evidence of meaningful human control (Siebert et al. 2022)
LLM-supported decisions	Code generation, legal research, and medical advice increasingly incorporate LLM output into consequential decisions	The responsibility chain spans model developer, deploying organisation, and the individual who accepted the model's output

Table 5: Four further sectors, summarised.

12. Limitations and Future Research

12.1 Limitations

The proposal remains unvalidated and faces several substantial limitations, summarised here rather than repeated at each point in the argument above where they apply.

Falsifiability and comparative testing: The paper's central claims are capable of being weakened or rejected through empirical comparison. The Operational Responsibility Gap hypothesis would be weakened where an organisation's existing governance records, audit trails, provenance mechanisms, and operational documentation already permit an independent reviewer to reconstruct the relevant responsibility claim without material inference, unresolved evidential breaks, or reliance on individual recollection. The architectural hypothesis would be weakened where the introduction of minimum additional responsibility-record requirements produces no material improvement in reconstructability, assessment consistency, reconstruction time, or suitability for reliance, or where any improvement is outweighed by disproportionate administrative, privacy, or behavioural costs. The institutional-separation hypothesis would be weakened where first-party or self-verifying arrangements generate equivalent reviewer confidence and external reliance to functionally separated arrangements under comparable conditions. Pilot protocols should therefore pre-specify what counts as material improvement, disproportionate burden, and a failed test rather than determining success criteria after results are observed.

These propositions require comparative testing rather than confirmation by the architecture's originator. Appropriate designs may include within-organisation comparisons between existing arrangements and a prospective responsibility-record intervention, matched-event comparisons where sufficiently similar events exist, and independent review of the same evidence by multiple assessors. The relevant measures must be defined before the pilot begins and reported regardless of whether the outcome supports the proposal.

No empirical proof: No comparative evidence currently shows that RI reduces the Operational Responsibility Gap more effectively than existing governance, auditing, or provenance mechanisms.

No legal recognition: RI is not currently a recognised legal compliance pathway in any jurisdiction, and this paper does not argue that it should be treated as one (Section 3.1).

Administrative burden: Decision-level capture may impose cost, delay, or documentation burdens that this paper does not quantify.

Gaming and scapegoating: Organisations may use responsibility records strategically to shift blame onto lower-authority individuals; Section 5.4's allocation-agency-accountability requirement is intended to reduce, but cannot eliminate, this risk.

Privacy and surveillance: Responsibility records may expose sensitive personal or professional information about the individuals named in them, beyond the organisation being examined.

Institutional independence: Formal separation of the kind argued for in Section 9 may exist on paper without genuine independence in practice — and, as Section 9.6 makes explicit, does not yet exist in practice for the author's own implementation.

Trust bootstrapping: New recognition institutions lack an established record of reliable judgement, as discussed in Section 9.5.

Interoperability: Different sectors and jurisdictions may define authority, evidence, and responsibility differently, limiting cross-context comparability of records.

Long-term maintenance: Records may need to remain interpretable and verifiable for years or decades after the decisions they describe.

False precision: Structured records may create an impression that responsibility is singular and simple where it is in fact shared, contested, or relational, echoing the concerns raised by Moss et al. (2022) and Porter et al. (2024) discussed in Section 4.

A further limitation concerns the author's position. As disclosed in Section 8.7 and in the Declarations, the author originated the Responsibility Infrastructure framework and architectural proposal presented in this paper and is commercially involved in developing an implementation of selected lifecycle mechanisms, and, as Section 9.6 discusses directly, currently controls the marks and the only existing implementation of the architecture the paper otherwise argues should be structurally separated. This creates an obvious incentive to characterise the gap as larger, and the proposed solution as more necessary, than a disinterested observer might. Readers should weigh the architectural argument on its merits independently of the author's commercial stake.

A prototype has been developed to explore the feasibility of selected lifecycle mechanisms; no prototype results are reported or relied upon in support of the conceptual claims advanced in this article. Independent replication or critique by researchers without the author's commercial stake would materially strengthen or weaken the claims made here.

A further institutional limitation should be stated directly. This paper is self-authored and self-published by an organisation affiliated with the author while arguing that self-assertion alone should not support independent reliance. Publication establishes authorship, provenance, and the existence of the argument. It does not independently validate the argument or satisfy the institutional separation proposed in Section 9. The claims should therefore remain open to external review, criticism, replication, and empirical testing.

12.2 Future Research

The paper identifies several areas for future research:

1. **Standardisation of responsibility primitives:** Can the lifecycle stages proposed in Section 8.2 be standardised? What data models, ontologies, and interchange formats would support interoperable responsibility records across organisations and sectors?
2. **Independent recognition mechanisms:** How should independent recognition of responsibility claims operate in practice? What institutional structures are needed, and what standards should recognition bodies adhere to? The ISO/IEC 42006:2025 framework for AI management system certification bodies (ISO 2025) may provide a useful starting point, but decision-level responsibility recognition raises distinct challenges.
3. **Interoperability across organisations:** Can responsibility records be made interoperable across organisations, sectors, and jurisdictions, while preserving privacy, commercial confidentiality, and data protection?
4. **An Operational Responsibility Gap diagnostic instrument:** Section 7 treats the gap as dimensional rather than binary. Developing and validating a responsibility gap scan could allow assignment clarity, evidenced agency, acceptance, evidence coherence, independent assessability, and reliance readiness to be compared across organisations or sectors. Such an instrument would require defined and reproducible measures, testing against real events, and safeguards against turning a conceptual diagnostic into a self-validating commercial claim.
5. **Empirical validation:** What empirical evidence is needed to validate the proposed architecture? A bounded pilot exploring selected lifecycle mechanisms was in progress at the time of writing (see Declarations). Its results, and those of subsequent pilots, case studies, and comparative analyses, would help establish whether RI can practically address the Operational Responsibility Gap; no such results are reported or relied upon in this article.
6. Institutional implementation and the bootstrapping problem: How should the functional separation argued for in Section 9 be implemented in practice, and how should a newly separated recognition body accumulate third-party trust before it has an observable track record?
7. **Integration with existing frameworks:** How should RI be integrated with existing governance frameworks, regulatory requirements, and auditing practices, including the EU AI Act provisions discussed in Section 3.1?

8. **Adversarial considerations:** How can RI be designed to resist manipulation, gaming, or the responsibility-laundering risk described in Section 7.2? What security properties are needed to ensure the integrity of responsibility records?
9. **Temporal sustainability:** How should responsibility records be maintained over extended periods — potentially decades — to ensure that they remain accessible, verifiable, and meaningful when responsibility challenges arise long after the original decision?

12.3 Bounded Pilot Testing

A bounded pilot should begin with a real operational event, responsibility claim, or defined prospective process in which responsibility passes between identifiable parties and later reconstruction matters. The organisation should first attempt reconstruction using its existing governance records, workflow systems, audit trails, communications, and retained evidence. Only the minimum additional responsibility-record requirements indicated by the baseline gap should then be introduced for a comparable event or defined period. The purpose is not to demonstrate adherence to a complete proprietary implementation, but to assess whether the proposed requirements materially improve the reconstruction and independent assessment of responsibility.

Low-barrier pilot formats may include retrospective reconstruction of anonymised past events, tabletop exercises using synthetic or redacted records, prospective studies involving a single team and a small number of decisions, matched comparisons of existing case-management approaches, or independent review of records from a real incident. A pilot need not test the full RI lifecycle. It may isolate one proposition, such as whether explicit records of authority, acceptance, agency, action, and evidence reduce reviewer disagreement, unresolved evidential breaks, or reconstruction time.

A credible pilot should include, at minimum:

- a clearly defined event, responsibility claim, or operational boundary;
- a documented baseline using existing arrangements;
- evidence of authority, allocation, agency, and acceptance;
- preserved action and evidence provenance;
- an assessment independent of the implementation owner;
- a defined relying party or reliance decision; and
- predetermined measures, limitations, and failure conditions.

Third-party reliance cannot be fully evaluated until a real receiving party is asked to make a decision using the resulting claim or standing. Early pilots should therefore distinguish between potential usability, assessed by an independent reviewer, and actual reliance, demonstrated only where a separate party changes, confirms, conditions, or refuses an action on the basis of the record. Absence of actual reliance evidence should be reported as a limitation rather than treated as proof of success.

Reconstructability should be assessed for the defined responsibility claim rather than for the operational process in general. Each reviewer should record whether the claim is reconstructable, partially reconstructable, or not reconstructable, together with the missing or contested elements. Reviewer agreement should compare those independently recorded conclusions and the principal reasons supporting them. Administrative burden should include the time and resources required to create, maintain, review, and resolve the additional record.

Pilot measures should be defined prospectively. Reconstruction time should run from the point at which a reviewer receives the authorised record set to the point at which the reviewer reaches a documented conclusion or records that reconstruction is not possible. An unresolved responsibility-chain break should be recorded where a required element, including authority, allocation, acceptance, agency, action, evidence, or assessment, cannot be established from the available record without material assumption or unsupported inference.

A pilot would support the hypothesis where the resulting responsibility record can be reconstructed more completely, with less material inference and lower dependence on individual recollection, and where an independent reviewer can assess the claim more consistently than under the baseline arrangements. Relevant measures may include reconstruction time, unresolved breaks in the responsibility chain, evidence of authority and acceptance, reviewer agreement, administrative burden, privacy effects, and the ability of a third party to use the resulting determination in a defined reliance decision.

The hypothesis would be weakened where existing records already produce an equivalent responsibility chain, where the additional requirements create material burden without improving reconstruction, where independent reviewers cannot use the record consistently, or where formalisation increases blame-shifting, privacy risk, administrative gaming, or responsibility laundering. Negative or mixed results should therefore be treated as evidence about the limits, proportionality, and appropriate scope of the proposal rather than as failed adoption. Claims about external reliance remain unproven where no independent relying party actually uses the result, and any recognition model would be weakened where independent reviewers applying the same published criteria cannot reproduce materially equivalent outcomes.

This pilot logic is intentionally public at the level required for evaluation. It does not disclose proprietary scoring systems, thresholds, interview sequences, sector-specific diagnostic instruments, implementation templates, or remediation methods. Those methods require separate development, governance, and empirical validation.

To support independent participation, the evaluation method should also be published as a standalone open pilot protocol setting out the research question, event-inclusion criteria, baseline method, minimum intervention, reviewer instructions, measures, privacy and consent safeguards, conflict-of-interest disclosures, failure conditions, and rules for reporting negative or mixed results. Participation in such testing should not require adoption of a proprietary implementation, a trade mark licence, or the complete RI lifecycle.

13. Conclusion

AI governance frameworks provide essential duties, controls, and oversight. They do not necessarily produce a coherent and independently assessable account of how responsibility was allocated, accepted, exercised, and evaluated in a specific operational event. This paper defines that evidential condition as the Operational Responsibility Gap.

Responsibility Infrastructure is proposed as a bounded conceptual response: a common architecture for representing responsibility, preserving provenance, attaching evidence, supporting assessment, resolving current standing, and enabling reliance across organisational boundaries. It is not offered as a replacement for law, governance, auditing, provenance, or existing operational systems, nor as a necessary layer where those arrangements already produce complete and independently reconstructable responsibility records. Nor, as Section 9.6 makes explicit, is it currently operated through the separated institutional structure its own argument calls for; that separation is a target the author's implementation has not yet reached.

The proposal remains a testable infrastructure hypothesis. Its interoperability, legitimacy, proportionality, administrative burden, privacy effects, susceptibility to gaming, and practical value require independent implementation and empirical evaluation. The paper's central claim is therefore limited but consequential: where responsibility must support external reliance and existing arrangements cannot reconstruct the full responsibility condition, governance alone may not be enough.

The next question is empirical. Researchers, regulated organisations, independent evaluators, and critical reviewers — including privacy, labour-rights, records-management, audit, and sector specialists — should test whether the Operational Responsibility Gap appears in real events, whether existing systems can already close it, and whether the minimum additional responsibility-record requirements proposed here improve reconstruction and external reliance without introducing disproportionate burden or new forms of responsibility laundering. This paper does not ask institutions to accept Responsibility Infrastructure as a completed field. It invites them to test a bounded hypothesis, report the results, and reject, refine, or extend the architecture according to the evidence. Responsibility Infrastructure should advance only to the extent that such tests support it.

Declarations

Competing interests. The author developed the Responsibility Infrastructure framework and architectural proposal described in this paper and is affiliated with La Touche Academy Ltd, the publisher of this paper. La Touche Infrastructure Ltd is developing a commercial implementation of selected mechanisms from the lifecycle described in Section 8.2, while related intellectual property is owned by La Touche Holdings Ltd. That implementation treats its detailed event models, state-transition logic, evidence-assessment methods, trust-resolution logic, and other implementation mechanics as proprietary and confidential; this paper describes the conceptual and interoperability architecture only. The paper does not evaluate any proprietary

implementation's performance or outcomes. Section 9.6 addresses directly the relationship between this competing interest and the paper's separation argument.

Funding. The author received no external funding for this research.

Ethics approval. Not applicable. This is a conceptual paper and does not report research involving human participants, human data, or animals.

Author contributions. Sole authorship. The author conceived, drafted, reviewed, and approved the manuscript in its entirety.

Data availability. This paper is conceptual and does not present empirical results. The published Responsibility Infrastructure Protocol and associated public publication records are available through the canonical Responsibility Infrastructure website. A prototype explores selected lifecycle mechanisms described in Sections 8–10, but no prototype performance data or completed independent evaluation are reported or relied upon in support of this paper's arguments.

Use of generative AI. Generative AI tools were used to assist with drafting, restructuring, literature comparison, and language refinement during the preparation of this manuscript, beyond ordinary grammar or translation assistance. The intellectual concept, architectural claims, and final arguments are the author's own; all AI-assisted output was reviewed, edited, and verified by the author, who accepts full accountability for the manuscript's content and accuracy.

References

AISI. (2025). AISI frontier AI trends report 2025. UK AI Security Institute. https://www.aisi.gov.uk/research/aisi-frontier-ai-trends-report-2025

AISI. (2026). Loss of oversight: How AI systems may become harder to audit, monitor, and investigate. UK AI Security Institute. https://www.aisi.gov.uk/research/loss-of-oversight-how-ai-systems-may-become-harder-to-audit-monitor-and-investigate

Anthropic. (2023). Anthropic's responsible scaling policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). AI auditing: The broken bus on the road to AI accountability. arXiv preprint, arXiv:2401.14462. https://doi.org/10.48550/arXiv.2401.14462

Busuioc, M. (2020). Accountable artificial intelligence: Holding algorithms to account. Public Administration Review, 81(5), 825–836. https://doi.org/10.1111/puar.13293

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. arXiv preprint, arXiv:2102.04201. https://doi.org/10.48550/arXiv.2102.04201

Coeckelbergh, M. (2019). Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and Engineering Ethics, 26, 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Costanza-Chock, S., Harvey, E., Raji, I. D., Czernuszenko, M., & Buolamwini, J. (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). https://doi.org/10.1145/3531146.3533213

Danaher, J. (2022). Tragic choices and the virtue of techno-responsibility gaps. Philosophy & Technology, 35, 94. https://doi.org/10.1007/s13347-022-00519-1

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

European Union. (2026). Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI) (Text with EEA relevance). Official Journal of the European Union, L series, 24 July 2026. https://eur-lex.europa.eu/eli/reg/2026/1744/oj

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Google. (2023). Secure AI framework. https://blog.google/technology/safety-security/an-introduction-to-googles-secure-ai-framework/

Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. IEEE Internet Computing, 21(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: Accountability and safety. Bulletin of the World Health Organization, 98(4), 251–256. https://doi.org/10.2471/BLT.19.237487

ISO. (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. International Organization for Standardization. https://www.iso.org/standard/42001

ISO. (2024). ISO 42001 explained: What it is. International Organization for Standardization. https://www.iso.org/home/insights-news/resources/iso-42001-explained-what-it-is.html

ISO. (2025). ISO/IEC 42006:2025 — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems. International Organization for Standardization. https://www.iso.org/standard/42006

Loi, M., & Spielkamp, M. (2021). Towards accountability in the use of artificial intelligence for public administrations. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). https://doi.org/10.1145/3461702.3462631

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19) (pp. 220–229). https://doi.org/10.1145/3287560.3287596

Mökander, J., & Floridi, L. (2023). Operationalising AI governance through ethics-based auditing: An industry case study. AI and Ethics, 3(2), 451–468. https://doi.org/10.1007/s43681-022-00171-7

Moss, E., Metcalf, J., Singh, R., Tafese, E., & Watkins, E. A. (2022). A relationship and not a thing: A relational approach to algorithmic accountability and assessment documentation. arXiv preprint, arXiv:2203.01455. https://doi.org/10.48550/arXiv.2203.01455

NIST. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Nyholm, S. (2017). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. Science and Engineering Ethics, 24, 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

OECD. (2024). Recommendation of the Council on Artificial Intelligence (updated). Organisation for Economic Co-operation and Development. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OpenAI. (2025). Preparedness framework (Version 2). https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

Porter, Z., Ryan, P., Morgan, P., Al-Qaddoumi, J., Twomey, B., McDermid, J., & Habli, I. (2024). Unravelling responsibility for AI. arXiv preprint, arXiv:2308.02608. https://doi.org/10.48550/arXiv.2308.02608

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. arXiv preprint, arXiv:2001.00973. https://doi.org/10.48550/arXiv.2001.00973

Siebert, L. C., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2022). Meaningful human control:

Actionable properties for AI system development. AI and Ethics, 3, 241–254. https://doi.org/10.1007/s43681-022-00167-3

Sparrow, R. (2007). Killer robots. Journal of Applied Philosophy, 24(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Stahl, B. C. (2023). Embedding responsibility in intelligent systems: From AI ethics to responsible AI ecosystems. Scientific Reports, 13, 7336. https://doi.org/10.1038/s41598-023-34622-w

Terzis, P., Veale, M., & Gaumann, N. (2024). Law and the emerging political economy of algorithmic audits. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). https://doi.org/10.1145/3630106.3658970

Tigard, D. W. (2020). There is no techno-responsibility gap. Philosophy & Technology, 34, 589–607. https://doi.org/10.1007/s13347-020-00414-7

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

UNESCO. (2024). Recommendation on the ethics of artificial intelligence (updated). https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. Minds and Machines, 34, 21. https://doi.org/10.1007/s11023-024-09674-0

Veluwenkamp, H., & Hindriks, F. (2024). Artificial agents: Responsibility & control gaps. Inquiry, 1–22. https://doi.org/10.1080/0020174X.2024.2410995

W3C. (2013). PROV-overview. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/prov-overview/