PROJECT SENTINEL

Paper IV · Full Field Paper

The Reconstruction Principle

The Ultimate Test of AI Governance

Public Edition · Project Sentinel Research Series

June 2026

Abstract

The first three papers in the Project Sentinel series trace a single problem from diagnosis to design. Paper I named the **Accountability Gap**: the widening distance between the consequences artificial intelligence (AI) systems can produce and the mechanisms available to assign responsibility for them. Paper II proposed the **Accountability Control Plane** as the organizational layer where that responsibility can be located and exercised. Paper III set out the **FairByDesign Doctrine**, arguing that fairness and accountability must be engineered into systems rather than asserted after the fact.

This paper completes the arc with a deceptively simple question: when a framework is in place, a control plane has been stood up, and a system claims to be fair by design, how would anyone *know* whether any of it worked once an outcome goes wrong? The answer offered here is the **Reconstruction Principle**—the proposition that the ultimate evidentiary test of governance is whether significant outcomes can be reconstructed and accountability assigned. If an outcome cannot be reconstructed, governance over that outcome cannot be demonstrated. From this the paper derives a portable, framework-agnostic instrument, the **Reconstruction Test**, and a **Reconstruction Maturity Scale** that locates any system, program, or enterprise on a path from opaque to continuous. The argument is situated within the established scholarly and regulatory record, from the foundational *responsibility gap* (Matthias, 2004) to the binding traceability and oversight obligations of the European Union's Artificial Intelligence Act (European Parliament & Council of the European Union, 2024).

Keywords: AI governance, accountability, auditability, traceability, reconstruction, algorithmic transparency, oversight

Contents

Why This Matters Now

Many serious AI failures produce the same institutional silence. A decision goes wrong, a recommendation causes harm, an autonomous component acts without authorization—and in the room afterward, the first question is not how to fix it. It is more basic and more damaging: *what actually happened, and who was responsible?* Too often, no one can say. Organizations that hold certificates, frameworks, and risk registers discover that none of it answers the only question that matters once an outcome has already occurred.

This is not a hypothetical. Between 2013 and 2019, the Dutch childcare-benefits scandal exposed how algorithmic risk profiling, the discriminatory use of nationality, harsh anti-fraud policy, and fragmented institutional oversight could combine to produce large-scale harm. Tens of thousands of parents and caregivers were falsely accused of fraud, while the opacity of the risk-classification process and the surrounding administrative decision chain made accountability difficult to establish (Amnesty International, 2021). The official parliamentary inquiry into the affair concluded that fundamental principles of the rule of law had been violated (Parliamentary Interrogation Committee on Childcare Benefits, 2020). The governance failure was not only the discrimination itself; it was that, for years, no one could establish *what* had happened or *why*.

Paper III introduced reconstruction as the highest evidentiary level of the FairByDesign Doctrine. Paper IV isolates that concept, expands it into a six-question diagnostic—separating why an outcome occurred from what influenced it—and proposes a maturity scale for assessing reconstruction capability across systems and organizations. The contribution is not to invent reconstruction, but to make it a standalone, portable test.

The Core Claim

Governance serves many ends. It exists to prevent harm, protect rights, manage risk, secure systems, enable redress, preserve human agency, and support legitimate decision-making; the NIST AI Risk Management Framework, for instance, treats governance as a continuous, cross-cutting function across the AI lifecycle (National Institute of Standards and Technology [NIST], 2023). But every one of those ends shares a dependency that becomes visible only after an outcome has occurred—the ability to show what the system did and why. That dependency is the evidentiary test this paper isolates:

> *The ultimate evidentiary test of AI governance is whether significant outcomes can be reconstructed and accountability assigned. If an outcome cannot be reconstructed, governance over that outcome cannot be demonstrated.*

The consequence is sharp. A governance program that cannot survive contact with a real incident cannot demonstrate that it governed the incident at all. The intuition is consistent with the direction of intergovernmental standards, which hold that AI actors should be accountable for the proper functioning of AI systems based on their roles, context, and ability to act (Organisation for Economic Co-operation and Development [OECD], 2024).

The Governance Problem

The difficulty of attributing the behavior of autonomous, learning systems to identifiable responsible parties was named two decades ago as the *responsibility gap*: the condition in which the operator of a learning machine can no longer predict its behavior and therefore cannot straightforwardly be held answerable for it (Matthias, 2004). The intervening years added a second difficulty—opacity. Modern AI decisions are mediated by data, models, and tools whose operation resists inspection, so that the link between design and outcome is often unavailable even to the people who built the system (Mittelstadt et al., 2016).

These two difficulties compound after deployment. Once a system is in production, emergent problems can become difficult or impossible to trace back to their source (Raji et al., 2020). The result is an institutional blind spot: harm is visible, but its causes, its controls, and its owners are not. Governance that cannot close this blind spot leaves the responsibility gap exactly where it found it.

Why Existing Answers Fall Short

The conventional responses are necessary but insufficient. **Documentation** proves intent, not behavior: a policy describing a control is not evidence that the control was active when an outcome occurred. **Compliance** certifies that a process was followed, not that a specific event can be explained. **Risk scoring** ranks hazards in advance, but says nothing about reconstructing a hazard that has already materialized. And **principles**—fairness, transparency, accountability—are aspirations until they are attached to evidence that can be produced on demand.

Each of these can be present in full while the core question remains unanswerable. An organization can be documented, compliant, risk-scored, and principled, and still be unable to say what its system did, why it did it, or who authorized the conditions that produced the harm. The missing element is not more governance of the familiar kind. It is a different test.

The Reconstruction Principle

The reframe is to make reconstruction the test of governance rather than a by-product of it. The central research question becomes:

> *When an AI-enabled decision or action causes harm, can the event be reconstructed well enough to establish accountability?*

Note what this does *not* ask. It does not ask whether harm can be prevented, whether systems can be perfectly understood, or whether outcomes can be predicted. Those goals are worthy but not always achievable. Reconstruction capability, by contrast, can be deliberately designed, tested, and improved, even where complete causal explanation remains impossible—and that is precisely what makes it a usable standard. The principle holds that every *significant* AI-enabled outcome should be capable of answering

six questions. The threshold of significance is itself a governance decision: the test applies to outcomes of material financial, legal, safety, or reputational consequence, not to every trivial inference.

The Reconstruction Test

Each question pairs a category of required evidence with the failure mode that appears when that evidence is missing (Table 1).

Table 1. *The six reconstruction questions, their evidence, and their failure modes.*

Question	Required evidence	Failure mode
1. What happened?	System outputs, logs, transactions, record of actions performed	Unknown event history
2. Why did it happen?	Inputs, context, prompts, applicable policies, decision records	Outcome without rationale
3. What influenced the outcome?	Data sources, external tools, models, human inputs	Invisible influence chains
4. What controls existed?	Guardrails, approval gates, access restrictions, policy controls	Unknown governance state
5. Could intervention occur?	Escalation procedures, override authority, human review records	No meaningful stop authority
6. Who is accountable?	Ownership assignments, accountability maps, authorization records	Accountability vacuum

The questions are ordered deliberately—event, rationale, influence, control state, intervention, and ownership. Two of the six are already reflected in law. Question 1 corresponds to the automatic event-logging the EU AI Act establishes for covered high-risk systems under Article 12, and Question 5 to its human-oversight obligation under Article 14—both subject to the Regulation's phased application timetable (European Parliament & Council of the European Union, 2024). Question 3 confronts the opacity problem directly: when data, tools, models, and human inputs are not recorded, the causal chain becomes invisible and responsibility cannot be traced to its sources (Mittelstadt et al., 2016). Question 6 is the Accountability Gap of Paper I—and the responsibility gap of Matthias (2004)—observed at the level of a single event.

Two qualifications keep the test honest. Reconstruction establishes an evidenced account of the event and its governing conditions; it does not automatically prove a unique causal explanation where the available evidence supports only probabilistic inference. And a reconstruction is complete only if it is intelligible, accessible, and contestable by the forum entitled to examine it—including affected people where consequential decisions are involved. A technically faithful reconstruction that no harmed person can understand or challenge is incomplete accountability.

The Reconstruction Maturity Scale (Proposed)

Reconstruction capability is not binary. The six-level scale proposed in Table 2 locates a system, a program, or an enterprise on a path from opaque to continuous. It is an original FairByDesign diagnostic, not an empirically validated industry benchmark.

Table 2. *The proposed reconstruction maturity scale.*

Level	Name	Description
0	**Opaque**	The organization cannot answer most of the six questions.
1	**Partial**	Some records exist, but reconstruction is incomplete and unreliable.
2	**Traceable**	Events can generally be reconstructed internally.
3	**Auditable**	Independent reviewers can verify the reconstruction, not merely accept it.
4	**Forensic**	Evidence supports chain of custody, integrity verification, cross-system correlation, and incident review.
5	**Continuous**	Reconstruction is built into operations by design, available in real time rather than assembled after the fact.

The decisive shift is at Level 3. Below it, reconstruction is an internal claim—the organization asserts that it can explain its own systems. At Level 3 and above, reconstruction becomes *externally verifiable*, which is the point at which it can support accountability to regulators, courts, and the public. This is the logic of internal algorithmic auditing, which embeds a robust, verifiable process into the development lifecycle rather than relying on after-the-fact external scrutiny alone (Raji et al., 2020). The scale is a proposed diagnostic rather than an empirically validated benchmark; its purpose is to expose the difference between internal claims of traceability and independently verifiable reconstruction.

Formal use of the scale requires defined evidence thresholds, sampling rules, assessor-independence criteria, and periodic reassessment, and should also address authoritative records (as distinct from ordinary logs), retention and legal hold, chain of custody, cross-system correlation, and the availability of evidence held by third parties; these are reserved for the operational assessment instrument. As presented here it is a directional diagnostic, not a scored audit standard.

Control Translation: Reconstruction Across Stakeholders

A principle that cannot be assigned to an owner is decoration. Reconstruction is a shared capability, and each of the six questions resolves to a control owner, an evidentiary artifact, and a point of escalation. The role-based distribution below is consistent with the accountability that intergovernmental standards now assume (OECD, 2024).

- **Engineers** own traceability by design. Reconstruction that is not engineered in cannot be retrofitted with confidence, because evidence that was never captured cannot be recovered. Their artifact is the log and the data-lineage record (Questions 1–3).
- **Security teams** own evidence integrity. A reconstruction is only as trustworthy as the records it rests on; logs that can be altered, deleted, or repudiated provide the appearance of reconstruction without its substance.
- **Security and incident-response functions** own preservation and investigative readiness: activating legal or forensic holds, securing volatile evidence, correlating events across systems, and preserving chain of custody when an incident occurs.
- **Privacy and data-governance functions** own proportional evidence design: what may be captured, for which purpose, for how long, under whose access, how affected-person rights are handled, and how evidence is securely disposed of when retention is no longer justified.
- **Model-validation and quantitative-risk functions** own the characterization of evidentiary uncertainty. They assess whether the available records support causal, probabilistic, or merely descriptive conclusions, and document performance, drift, calibration, and known limits relevant to the reconstructed event.
- **Enterprise architecture, MLOps/SRE, and third-party-risk functions** own cross-system evidence availability. They ensure that system design, service interfaces, operational telemetry, contracts, and exit provisions preserve the organization's ability to reconstruct events across supplier boundaries.
- **Lawyers** own the requirement definition—translating jurisdictional obligation, including the EU AI Act's logging and oversight duties, into the concrete evidentiary fields the system must capture (Questions 4–6).
- **Independent audit functions** own or support verification, while management remains responsible for the underlying controls and evidence. Verification is the difference between Level 2 and Level 3: between an organization that believes it can reconstruct events and one whose ability has been independently confirmed.
- **Executives** own deployment authorization and remain accountable for ensuring that decision, evidence, oversight, and intervention responsibilities are clearly assigned. Executive accountability does not replace the traceable responsibilities of other actors across the lifecycle.
- **Regulators and supervisory authorities** evaluate reconstruction capability and compliance within their legal mandates; the organization itself remains responsible for establishing and maintaining readiness.

Proportionality is a design constraint, not an afterthought. Reconstruction does not justify indiscriminate collection or indefinite retention. Evidence must be proportionate to the significance and risk of the decision, limited to what is necessary for accountability, protected against secondary use, retained only as long as justified, and accessible under defined authority. A reconstruction system that proves accountability by creating an unnecessary surveillance archive is not well governed.

Two engineering disciplines make the rest dependable. First, reconstruction evidence must be protected against alteration, deletion, repudiation, and selective capture through appropriately segregated, access-controlled, time-synchronized, and integrity-verified records—with tool calls, model versions, system prompts, policy versions, and human interventions correlated under a single event identifier. In distributed deployments the relevant evidence may sit across model vendors, orchestration platforms, vector stores, identity systems, tools, clouds, and human workflows, so architecture and contracts must preserve access to it before an incident rather than after. Second, reconstruction capability should be tested before consequential use through simulated incidents, evidence-retrieval exercises, and adversarial scenarios, rather than assumed to work when an incident finally arrives. The maturity scale supplies the metric, the logging and oversight obligations supply the minimum evidentiary floor, and override authority under Question 5 must itself be logged so that the act of intervention—or its absence—can later be reconstructed.

The Regulatory Implication

AI regulation is maturing from a posture of principle toward a posture of evidence. The NIST AI Risk Management Framework—which NIST is currently revising—treats governance as a continuous, cross-cutting function and emphasizes accountability, transparency, documentation, provenance, monitoring, and clearly assigned responsibilities as important elements of trustworthy AI risk management (NIST, 2023). The European regime moves further toward enforceable obligation: the EU AI Act establishes automatic event-logging, human-oversight, and deployer log-retention obligations for covered high-risk systems, subject to its phased application timetable and the revised implementation schedule agreed through the 2026 Digital Omnibus process (European Parliament & Council of the European Union, 2024, arts. 12, 14, 26). As of June 2026, that provisional agreement—reached on 7 May 2026 and still subject to formal adoption—schedules the main Annex III high-risk obligations from 2 December 2027 and high-risk obligations for AI embedded in regulated products under Annex I from 2 August 2028.

The Reconstruction Principle offers one practical expression of where this trajectory is leading. The operative question shifts from “Did the organization claim to have governance?” to “Can the organization reconstruct the event?” This is more demanding but more honest. A claim of governance can be manufactured; a reconstruction claim becomes much harder to sustain when an incident demands independently verifiable evidence. Reconstruction is the practical bridge between governance theory and governance evidence—the transparency standard against which algorithmic accountability can be measured (Diakopoulos, 2016).

The FairByDesign Reconstruction Model

The principle integrates the series through a single chain that runs from epistemic humility to demonstrable accountability:

Knowability Doctrine

↓

Distributed Accountability Framework

↓

Evidence Hierarchy

↓

Reconstruction Capability

↓

Demonstrable Governance

The chain begins with the **Knowability Doctrine**—the acknowledgement, foreshadowed by the responsibility gap (Matthias, 2004) and the opacity literature (Mittelstadt et al., 2016), that not everything about an AI system can be known in advance, which is why governance must focus on what *can* be established after the fact. It passes through the **Distributed Accountability Framework** of the stakeholder model and an **Evidence Hierarchy** that ranks records by integrity and verifiability. These converge on **Reconstruction Capability**, which produces **Demonstrable Governance**—governance that can be shown, not merely asserted. The model converts uncertainty into accountability without pretending the uncertainty away.

The Institutional Question

Beneath the operational test lies a question of public interest. As consequential decisions are increasingly delegated to systems whose behavior their own operators cannot fully predict, a society must decide what it will accept as an answer when one of those decisions causes harm. If the answer is a certificate, a policy binder, or a risk score, accountability becomes a performance. If the answer is a reconstruction—an evidenced account of what happened, why, under what controls, and on whose authority—then accountability remains real. The Reconstruction Principle is a proposal for which answer institutions should require.

Series Bridge

Paper IV is the hinge of the series. Papers I–III move from naming the problem (the Accountability Gap), to locating responsibility (the Accountability Control Plane), to engineering it in (the FairByDesign Doctrine). Paper IV supplies the test that tells whether any of it worked. The papers that follow operationalize the test: Paper V (*Governance Under Persistent Uncertainty*) develops the Knowability Doctrine; Paper VI (*Measuring Accountability*) turns the maturity scale into metrics; Paper VII (*The Human Oversight Control Plane*) deepens Question 5; and Paper VIII (*The Demonstrable Governance Standard*) consolidates the series into a single standard. Papers I–III are earlier works in the Project Sentinel series and are referenced as such; the external reference list below is limited to independently verifiable third-party sources.

Conclusion

Three impossibilities frame the modern governance challenge. Perfect prediction may never be possible. Perfect understanding may never be possible. Perfect alignment may never be possible. A governance regime that depends on any of them is built on ground that will not hold.

Reconstruction capability remains within reach. For significant outcomes, many of the relevant conditions, influences, controls, and decisions can be preserved and reassembled if the system and organization are designed to capture them—though missing telemetry, third-party components, privacy and minimization duties, stochastic behavior, and poorly recorded human decisions all set real limits. That design choice, made in advance, is what separates governance that can be demonstrated from governance that can only be claimed. The ability to reconstruct outcomes may become the defining evidentiary test of trustworthy AI governance: the one question that, when answered, renders the rest of governance credible—and when unanswered, leaves it merely asserted.

Evidence Notes

The following notes map each external source to the specific claim it supports. Sources are cited where they genuinely carry an argument, not for decoration; each was verified against its primary or authoritative record. Papers I–III are earlier works in the Project Sentinel series and are not included in this external reference list, which is limited to independently verifiable third-party sources.

Source	What it supports in this paper
Matthias (2004)	Establishes the responsibility gap—that learning systems sever traditional responsibility ascription. Grounds The Governance Problem, Question 6, and the Knowability Doctrine.
Mittelstadt et al. (2016)	Maps algorithmic opacity across the lifecycle. Supports Question 3 (invisible influence chains) and the premise that perfect understanding is not guaranteed.
Diakopoulos (2016)	Argues for algorithmic transparency and accountability standards. Supports the regulatory shift from claimed to demonstrable governance.
Raji et al. (2020)	Defines internal algorithmic auditing (SMACTR) and shows post-deployment issues are often impossible to trace to source. Anchors Why This Matters Now and the Auditable maturity level.
Amnesty International (2021)	Documents the Dutch childcare-benefits scandal, in which algorithmic risk profiling, the discriminatory use of nationality, and fragmented oversight combined to produce large-scale harm that was difficult to reconstruct.
Parliamentary Interrogation Committee on Childcare Benefits (2020)	The official Dutch parliamentary inquiry ("Ongekend onrecht") that concluded fundamental rule-of-law principles were violated in the childcare-benefits affair. Supports the factual account in Why This Matters Now.
NIST (2023)	Treats governance as a continuous, cross-cutting function and emphasizes accountability, transparency, documentation, provenance, monitoring, and assigned responsibilities. Supports The Regulatory Implication and Control Translation.
European Parliament & Council (2024)	EU AI Act Articles 12, 14, and 26: event-logging, human-oversight, and deployer log-retention obligations for covered high-risk systems, subject to phased application. Supports Questions 1 and 5.
OECD (2024)	Establishes role-based accountability—by role, context, and ability to act—as an intergovernmental standard. Supports The Core Claim and the Distributed Accountability Framework.

References

Amnesty International. (2021). *Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal*. Amnesty International. https://www.amnesty.org/en/documents/eur35/4686/2021/en/

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56–62. https://doi.org/10.1145/2844110

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

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Organisation for Economic Co-operation and Development. (2024). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449; adopted 2019, revised 2024). OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Parliamentary Interrogation Committee on Childcare Benefits [Parlementaire ondervragingscommissie Kinderopvangtoeslag]. (2020). *Ongekend onrecht* [Unprecedented injustice]. House of Representatives of the Netherlands. https://www.tweedekamer.nl/kamerleden_en_commissies/commissies/pok

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

About Project Sentinel

Project Sentinel is an ongoing research series on AI governance and accountability. This Field Paper is the authority edition of Paper IV; a shorter public version introduces the same diagnostic for a general audience. Subscribe to the free Mercury Brief to follow Project Sentinel and receive future public field notes; Mercury Research members receive the full research archive and deeper operational materials as they are released. The subscription builds a relationship with serious readers rather than gating the doctrine—the full Field Paper, its Evidence Notes, and its references are provided in this document.