PROJECT SENTINEL

PAPER III · FIELD PAPER

The FairByDesign Doctrine

Knowability, Distributed Accountability, and Evidence-Based Governance for Autonomous Systems

A doctrine for governing systems that cannot be perfectly understood

THIS DOCUMENT CONTAINS TWO ARTIFACTS

Part A — Full Field Paper (authority / archive version)

Part B — Public Doorway Version (public-facing entry)

Contents

PART A · FULL FIELD PAPER

The Doorway

We are now deploying machines into consequential decisions faster than we can explain them. Systems shape who is flagged, who is approved, what is escalated, and what is left alone — and the organizations that run them increasingly cannot say, in detail, why. The capabilities of these systems can appear unpredictably as they grow larger, and their inner workings resist ordinary inspection. The old assumption beneath most governance — that with enough effort a system can always be understood — has quietly stopped being true. The question is no longer how to govern a system you understand. It is how to stay accountable for one you do not.

This is Paper III in the Project Sentinel series. Paper I named the *accountability gap*. Paper II built the *accountability control plane* to operate across it. This paper supplies the doctrine beneath both: the principles an organization adopts when it accepts that it will govern systems it cannot fully understand, and refuses to govern them carelessly because of it.

The Core Claim

> *Governance should not be built on the assumption of complete understanding. It should be built on the assumption of partial knowability.*

If understanding is necessarily incomplete, accountability cannot rest on any one person's full comprehension of the system. It must be distributed across the actors who shape the system's behavior, and it must be backed by evidence that can be checked independently of what those actors say about themselves. The doctrine rests on three pillars: a **Knowability Doctrine**, a **Distributed Accountability Framework**, and an **Evidence Hierarchy**.

Why the Usual Answers Are Not Enough

The familiar responses do not close the gap. “Add more transparency” assumes the inside of the system can be made legible; for the hardest cases it cannot. “Explainable AI will fix it” helps at the margin but does not manufacture understanding where the model's behavior is genuinely emergent. “Name one accountable owner” collapses under the reality that consequential outcomes are the work of many hands. “Publish principles” produces statements of intent, not demonstrations of practice. Each answer addresses a symptom while leaving the structural problem — accountability under irreducible uncertainty — untouched. The FairByDesign reframe is to stop trying to extract accountability from the

system's internals and instead design it into everything that surrounds the system: what is recorded, who owns which decision, and what can be proven afterward.

Part I — The Knowability Doctrine

The Governance Problem

A great deal of governance practice assumes sufficient understanding can always be reached with enough effort. For traditional software, that was often defensible. For modern AI it is not. These systems exhibit emergent capabilities discovered only after deployment (Wei et al., 2022); they reason across high-dimensional spaces that resist human summary; they compound unpredictably in interaction; and their internal representations often map to no concept a human reviewer would recognize (Burrell, 2016). Understanding is not merely expensive here; sometimes it is not currently obtainable at all. A doctrine that pretends otherwise fails at the moments it is needed most.

The interpretation of “emergent abilities” is itself contested: some later work argues that apparent emergence can reflect the researcher's choice of evaluation metric rather than a genuine, discontinuous jump in the model's behavior (Schaeffer et al., 2023). The governance point survives the debate either way. Whether a capability arrives in a sharp jump or along a smoother curve, an organization cannot assume that its pre-deployment understanding will fully predict deployed behavior — and it is that gap between expectation and behavior, not the shape of the scaling curve, that governance must absorb.

Principle 1 — Not All Aspects of a System Are Equally Knowable

Knowability is not a single property a system has or lacks. Opacity itself comes in distinct forms — deliberate secrecy, a reviewer's lack of technical fluency, and the intrinsic character of large models at the scale required to use them (Burrell, 2016). The doctrine sorts the resulting states of knowledge into three categories, and treats them as a reusable handle.

Procedurally knowable

Observable directly, without inference: inputs, outputs, logs, access records, tools invoked. This is the floor of any serious program, because it is the only category that does not depend on interpretation.

Inferentially knowable

Not directly observable but estimable from evidence: risk exposure, behavioral drift, performance trends, emerging response patterns. Real knowledge, but conditional — only as good as its methods, and always to be stated with its uncertainty attached.

Epistemically limited

Not, at present, fully understandable: internal representations, latent abstractions, the actual pathway to an emergent behavior (Burrell, 2016). The honest move is to mark the boundary, not to claim insight that is not well founded.

The Governance Requirement

From this follows the inversion of how organizations often behave: **governance obligations increase as knowability decreases.** Where a system is procedurally knowable, observation may suffice. Where it is only inferentially knowable, the organization owes more monitoring, more margin, disciplined uncertainty. Where it is epistemically limited, it owes the strongest controls of all — constraints on deployment, tighter intervention authority, conservative assumptions — precisely because it cannot see what it is constraining.

> *Uncertainty is never to be treated as evidence of safety.*

The absence of an observed problem is not the presence of safety. A system whose behavior cannot be reconstructed is not a safe system; it is an unaccountable one. Confusing “we have not seen a failure” with “the system is safe” substitutes the limits of observation for a property of the world.

Part II — The Distributed Accountability Framework

The Governance Problem

When something goes wrong, organizations look for the single owner — the one name on the line. For modern AI that instinct fails. Where a harm is the work of *many hands*, its immediate causes rarely converge on a single locus of decision-making (Nissenbaum, 1996), and the problem has intensified with systems assembled from components and toolkits built by many parties (Cooper et al., 2022). Autonomy sharpens it further: when a learning system's behavior cannot in principle be predicted by its makers, traditional responsibility ascription strains into a *responsibility gap* in which neither manufacturer nor operator can straightforwardly be held to account (Matthias, 2004).

Principle 2 — Accountability Must Be Distributed but Traceable

Accountability, in its established sense, is a relationship between an actor and a forum: the actor must explain and justify its conduct, the forum can question and judge, and the actor may face consequences (Bovens, 2007). Distribution without traceability severs that relationship and becomes mere diffusion — the most common

way accountability disappears. The framework distributes ownership across roles while keeping, for each consequential decision, a chain a forum can actually follow.

The accountability map

Standing roles across the lifecycle — roles, not necessarily individuals; a small organization may concentrate several in one person. What matters is that each is held by someone identifiable:

- **Developers** — how the system is constructed.
- **Product owners** — its intended use.
- **Security teams** — protecting it and preventing misuse.
- **Risk teams** — assessing exposure.
- **Legal teams** — legal and regulatory obligations.
- **Audit functions** — independent verification.
- **Executives** — authorizing deployment.
- **Operators** — day-to-day execution.

The allocation rule

For every critical decision, four roles must be assigned and named:

- a **Decision Owner**, who makes the call;
- an **Evidence Owner**, responsible for the record of the decision and its basis;
- an **Oversight Owner**, responsible for reviewing it; and
- an **Intervention Authority**, with the standing and the means to stop or reverse it.

> *No critical decision should exist without all four. A decision missing any one of these is not governed; it is merely made.*

The separation is load-bearing. Merge Decision and Oversight and the review vanishes. Merge Evidence into Decision and the only record is kept by the person with the strongest interest in how it reads. Leave Intervention unassigned and the organization can recognize a problem and still be unable to act.

Accountability failure modes

- **Orphaned decisions** — consequential choices with no identifiable owner.
- **Shared-responsibility ambiguity** — everyone nominally responsible, so no one is.
- **Invisible approvals** — authorized somewhere, by someone, with no durable record.
- **Missing intervention authority** — the problem is seen and named, but no one is empowered to stop it.

Part III — The Evidence Hierarchy

The Governance Problem

Many programs run on declarations: the organization states it has governance, points to a binder, and treats the statement as the fact. A declaration describes an intention; it does not demonstrate a practice. Declarations are not evidence.

Principle 3 — Judge Claims by the Quality of the Evidence Behind Them

A mature program knows where each of its claims sits on a scale of confidence, from weakest to strongest.

Level 1 — Assertion. "We have governance." Nothing behind it but the statement. Where claims start, not where they should end.

Level 2 — Documentation. Policies and procedures. Shows what should happen, not what did. The gap between the two is where most failures live.

Level 3 — Operational records. Logs, approvals, configurations, decision records — the first level that demonstrates activity rather than intention. Now reflected in law: the EU AI Act requires high-risk systems to enable automatic recording of events across their lifetime (European Parliament & Council of the European Union, 2024, Art. 12).

Level 4 — Independent verification. Audit findings, external review, independent testing. Internal audit frameworks that yield a documented trail at each development stage are built to supply exactly this (Raji et al., 2020); the evidence no longer depends on the word of the interested party.

Level 5 — Reconstruction capability. The highest confidence: for a significant outcome, the organization can answer the full set of accountability questions.

- What happened?
- Why did it happen?
- Who authorized it?
- What controls existed at the time?
- Could intervention have occurred, and did it?

The Reconstruction Doctrine

This matters because, once deployed, emergent issues can otherwise become difficult or impossible to trace to their source (Raji et al., 2020). The hierarchy resolves into one test:

> *If a significant outcome cannot be reconstructed, governance over that outcome cannot be demonstrated — and what cannot be demonstrated should not be assumed to exist.*

This is the evidentiary counterpart to the rule about uncertainty. There: unseen problems are not absent problems. Here: an undemonstrable claim of control is not control. Reconstruction is where the three pillars meet — it needs the records knowability makes possible and the named owners distributed accountability supplies.

Part IV — The Reconstruction Spine: Governing Under Adversarial Conditions

The doctrine is tested hardest where systems meet adversaries. A security incident is, by its nature, a reconstruction problem: every investigation asks what happened, why, under whose authority, what controls were in force, and whether anyone could have intervened. That is Level 5 of the Evidence Hierarchy stated in the language of incident response. Security is therefore not a separate concern bolted onto the doctrine; it is the place the doctrine is proven or found wanting.

The surface to be governed widens as systems gain autonomy. Contemporary catalogues of language-model risk place prompt injection, excessive agency, and supply-chain compromise among the most serious exposures (OWASP, 2025), and agentic, tool-using systems amplify each of them: a single manipulated input can chain across tools and act in the world before any human notices. These attacks aim precisely at the regions the Knowability Doctrine marks as epistemically limited — the parts of a system whose behavior cannot be fully anticipated. An adversary need not understand the model's internals any better than its operators do; they need only find an input the operators did not foresee.

This forces governance from a one-time, pre-deployment approval toward continuous, runtime oversight. A model signed off once cannot account for an agent that improvises at machine speed. The field is already building in this direction: recent work proposes a runtime control plane that operationalizes the NIST AI Risk Management Framework for autonomous agents, with continuous, verifiable governance and explicit accountability hooks (Huang et al., 2025). That is Paper II's accountability control plane in its security register — and it requires that the Intervention Authority of Part II be able to halt or reverse a running system, not merely disapprove of it after the fact.

What the doctrine specifies, at its own level, is the evidentiary record such governance depends on. Call it the **Reconstruction Spine**: the minimum

continuous record an autonomous system must produce to remain governable under adversarial conditions. Its elements are:

- **System and actor identity** — which system or agent acted, under whose credentials and authority.
- **Model, prompt, and configuration version** — the exact model, system prompt, tools, and settings in force at the moment of action.
- **Input and provenance** — the input and its source, with untrusted or external content marked as the injection surface it is.
- **Tool-call record** — every external action invoked, with parameters and results: the agent's actual reach into the world.
- **Authorization trail** — what policy permitted the action, and who approved anything high-impact or irreversible.
- **Human override trail** — where oversight intervened, and where it could have but did not.
- **Control events** — which guardrails fired, throttled, or blocked, and which stayed silent.

These are not new controls invented for this paper; they are the operational records (Level 3) and verification trails (Level 4) of the Evidence Hierarchy, named in the form a security investigation actually needs. An organization that maintains the spine can reconstruct an attack; one that does not will discover, mid-incident, that it instrumented performance but not accountability.

> *An autonomous system that cannot be reconstructed under attack is not a secured system; it is an unobserved one.*

The Data Layer: Where Knowability Becomes Operational

One part of the spine deserves to be named on its own, because it is where the whole doctrine becomes operational. The data layer is one of the few parts of an autonomous system that can usually be governed more directly than the model's internal reasoning. Even where model behavior is epistemically limited, an organization can still ask what data the system was allowed to access, where that data came from, whether it was current, whether its use was authorized, whether it contained sensitive or protected information, whether it was retrieved from a trusted source, and whether that use was recorded.

In practice the data layer spans training and validation data, retrieval sources and vector stores, prompts, tool-accessed records, outputs, logs, feedback loops, and the downstream records the system itself changes. Each carries its own governance burden — lineage, quality, access rights, retention, consent basis, bias, security classification, and evidentiary preservation — and these are not optional niceties: for

high-risk systems the EU AI Act makes data and data governance an explicit legal obligation (European Parliament & Council of the European Union, 2024, Art. 10).

This is the doctrine's leverage point. A system can be epistemically limited at the model layer and still be governable at the data layer, so data lineage, provenance, access control, and retrieval-context records are not background detail — they are part of the Reconstruction Spine. Without them, an organization may be able to show that a system acted, but not what information shaped the action, and that is not enough for accountability. It is also where most adversarial pressure travels: poisoning, retrieval manipulation, unauthorized access, and sensitive-data exposure all ride through data paths.

> *If the data that shaped the outcome cannot be traced, the outcome cannot be fully reconstructed.*

This keeps Paper III what it is — a doctrine, not a control catalogue. The Reconstruction Spine is the doctrine-level specification of what a secure deployment must be able to prove; the concrete instrumentation, and a reusable reconstruction worksheet for AI-mediated decisions, belong to the operational papers and member assets that build on it. The security and data layers are now part of the doctrine, not a promise deferred to a later volume.

Part V — The FairByDesign Model

The three pillars are three faces of one model. **Knowability** determines what can be understood, and what cannot. **Distributed Accountability** determines who owns the outcome. The **Evidence Hierarchy** determines what can be proven, and at what confidence. Together they support accountability, oversight, trustworthiness, and legitimacy *without* requiring perfect understanding of the system. That is the reframe and the reason for the name: fairness, accountability, and legitimacy are not extracted from the system's internals after the fact; they are designed into the governance that surrounds it. This lifecycle posture — governance embedded throughout rather than bolted on — mirrors both voluntary frameworks and binding law (Tabassi, 2023; European Parliament & Council of the European Union, 2024).

Part VI — Control Translation: Implications by Role

For engineers. Build systems that generate evidence. Observability, logging, and reconstruction are not late add-ons; they are governance requirements from the first design decision, and automatic event recording is now a legal expectation for high-

risk systems (European Parliament & Council of the European Union, 2024, Art. 12). A system that cannot account for itself is incomplete, however well it performs.

For security teams. Treat the Reconstruction Spine as a security requirement, not a compliance afterthought. Threat-model the epistemically-limited regions of the system, instrument tool calls and authorization decisions as first-class telemetry, and ensure the Intervention Authority can halt a running agent at machine speed rather than only review it afterward. Prompt injection, excessive agency, and supply-chain compromise are the live exposures to design against (OWASP, 2025; Huang et al., 2025).

For lawyers. Define accountability obligations precisely — who owns what, under which duties, with what consequences — the answerability relation at the heart of accountability (Bovens, 2007). The framework is only as strong as the clarity of the duties attached to each role.

For auditors. Verify independently. Level 4 exists only if audit functions exercise real independence and produce a documented trail (Raji et al., 2020). An audit that merely confirms the audited party's assertions adds no confidence.

For executives. Accept accountable deployment authority. Authorization sits with executives and is not exercised without ownership. Deploying an autonomous system is an accountable act.

For regulators. Evaluate evidence, not promises. A regime that accepts assertions and documentation operates at Levels 1–2; one that demands operational records, independent verification, and reconstruction is asking the questions that separate governed systems from ungoverned ones. The risk-based NIST framework (Tabassi, 2023) and the EU AI Act's tiered, oversight-centered obligations (European Parliament & Council of the European Union, 2024, Art. 14) point this way.

The Institutional Question

Beneath the operational doctrine sits a public one. As a society delegates more of its consequential decisions to systems no one fully understands, who answers for the results — and on the strength of what evidence? The FairByDesign Doctrine is one answer: legitimacy does not require that we understand these systems completely, but it does require that someone owns each decision they drive and that the record can be reconstructed when it matters. A society that cannot meet that second condition has not delegated decisions to machines so much as misplaced its accountability in them.

Series Bridge

Paper I diagnosed the accountability gap; Paper II built the control plane to span it; Paper III supplies the doctrine that tells both what they are for. The work this paper leaves open is operational: how accountability behaves under *persistent* uncertainty, how reconstruction is engineered in practice, and how an organization measures whether its accountability is real or nominal. Those are the questions the next papers in the series take up.

Evidence Notes

A compact map of how each source supports the argument. Full bibliographic detail follows in the References.

Wei et al. (2022) — Grounds the claim that capabilities can emerge unpredictably with scale and cannot be extrapolated from smaller systems — the empirical basis for treating some aspects of a system as not knowable in advance (Part I).

Schaeffer et al. (2023) — Supplies the counter-position: apparent emergence may reflect the choice of evaluation metric rather than a discontinuous behavioral jump. Cited to keep the claim honest; the governance point holds whichever interpretation is correct (Part I).

Burrell (2016) — Supplies the typology of opacity — secrecy, technical illiteracy, and the intrinsic complexity of large models — that grounds the three knowability categories, and especially the "epistemically limited" one (Part I).

Matthias (2004) — Names the responsibility gap: when a learning system's behavior cannot in principle be predicted, traditional responsibility ascription strains — the problem distributed accountability is built to answer (Part II).

Nissenbaum (1996) — Establishes the "many hands" problem — why single-owner accountability fails for complex computerized systems (Part II).

Cooper et al. (2022) — Carries many-hands and relational accountability into modern data-driven and machine-learning systems assembled from many components (Part II).

Bovens (2007) — Provides the operative definition of accountability as an actor–forum relationship of answerability, grounding "distributed but traceable" (Part II).

Raji et al. (2020) — Supports both the accountability-gap framing and the evidence argument: emergent issues become hard to trace to their source,

and internal auditing yields the documented trails behind Levels 4–5 (Parts III, VI).

OWASP (2025) — Supplies the contemporary threat surface for language-model and agentic systems — prompt injection, excessive agency, supply-chain compromise — grounding the security surface and the Reconstruction Spine (Part IV).

Huang et al. (2025) — Evidence that runtime, control-plane governance for agentic AI is being built in practice, aligned to the NIST AI RMF with continuous, verifiable governance and accountability hooks; supports the shift from pre-deployment review to runtime oversight and ties the doctrine back to Paper II's control plane (Parts IV, VI).

Tabassi / NIST (2023) — Supports the risk-based, lifecycle, trustworthiness-oriented posture in which obligations are matched to risk (Parts V, VI).

EU AI Act / Regulation (EU) 2024/1689 (2024) — Supplies binding-law support across the doctrine: Art. 10 (data and data governance for high-risk systems) for the data layer, Art. 12 (automatic event logging) for Level 3, and Art. 14 (human oversight) for the Intervention Authority (Parts III, IV, VI).

References

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal, 13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Cooper, A. F., Laufer, B., Moss, E., & Nissenbaum, H. (2022). Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22)* (pp. 864–876). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533150

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, L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

Huang, K., Lambros, K. R., Huang, J., Mehmood, Y., Atta, H., Beck, J., et al. (2025). *AAGATE: A NIST AI RMF-aligned governance platform for agentic AI* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.25863

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

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics, 2*(1), 25–42. https://doi.org/10.1007/BF02639315

OWASP. (2025). *OWASP Top 10 for large language model applications (2025)*. OWASP Foundation. https://genai.owasp.org/llm-top-10/

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 (FAT* '20)* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)* (pp. 55565–55581). https://doi.org/10.48550/arXiv.2304.15004

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

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2206.07682

About This Series

Project Sentinel is an ongoing research series on AI accountability and the governance of autonomous systems. This Field Paper is the full authority version, with its complete argument, Evidence Notes, and references. A shorter Public Doorway Version (Part B, below) is written for open circulation and links back to this paper.

If this work is useful to you, the most valuable thing you can do is read the companion papers and bring the doctrine into a real governance conversation — a review, a control design, an audit charter. Readers who want the full archive, the implementation material, and early access to forthcoming papers can follow or join the series through Mercury Research. References above are verifiable and authoritative; if you adapt the doctrine, cite the sources, not just the slogans.

PART B · PUBLIC DOORWAY VERSION

Who Answers When the Machine Shapes the Outcome?

A public field essay on governing systems we can no longer fully understand

---

We are handing more and more consequential decisions to software — who gets flagged, who gets approved, what gets escalated. And we are doing it faster than we can explain how that software works. The newest AI systems develop abilities their own builders did not design and cannot fully predict, and their inner workings resist the kind of inspection that would let anyone say, in detail, why they did what they did.

That breaks an assumption buried in almost every governance playbook: that if you try hard enough, you can always understand the system. For the systems now being deployed, that is no longer reliably true. So here is the real question — not how to govern a machine you understand, but how to stay accountable for one you do not.

The short answer of the FairByDesign Doctrine: stop building governance on the dream of complete understanding, and build it on **partial knowability** instead. You do not need to see inside the system perfectly. You do need three things.

1. Know what you can know — and admit what you can't

Some things about a system can be observed outright: its inputs, outputs, and logs. Some can only be estimated from evidence, with the uncertainty stated honestly. And some — the deep internal workings — cannot, today, be fully understood at all. The doctrine's rule is counterintuitive but firm: the less you can know about a part of the system, the *more* governance you owe it. And one line does the most work of all:

> *Uncertainty is never evidence of safety.*

“We haven't seen it fail” is not “it is safe.” It is just the limit of what you have looked at.

2. Give every decision an owner — four owners, in fact

When AI systems cause harm, everyone reaches for the single person to blame, and there usually isn't one: modern systems are the work of many hands. The fix is not to invent a scapegoat but to assign ownership deliberately. For every critical decision, name four roles: a **Decision Owner** who makes the call, an **Evidence Owner** who keeps the record, an **Oversight Owner** who reviews it, and an **Intervention Authority** who can actually stop it. Miss any one of the four, and the decision isn't governed — it's just been made.

3. Be able to prove it later

Most governance runs on declarations: a policy binder, a statement that controls exist. But a statement of intent is not proof of practice. The strongest evidence is the ability to *reconstruct* what happened. Which leads to a test anyone can run on their own organization:

> *The Reconstruction Test: pick one consequential decision your system made recently. Can you say what happened, why, who authorized it, what controls were in place, and whether anyone could have stopped it? If not, you don't govern that decision. You only made it.*

This is also the test that matters most when a system is attacked or misused. Every security investigation is a reconstruction: what happened, who or what authorized it, which controls were in force, and whether anyone could have intervened. A system that cannot answer those questions under attack is not a secured system — it is an unobserved one. As AI agents gain the ability to use tools and act on their own, that record has to be kept continuously, at the speed the system runs, not assembled after the damage is done.

Why this matters beyond the org chart

This is not only a compliance puzzle. As a society, we are delegating real power to systems we cannot fully explain. If no one owns those decisions, and no one can reconstruct them afterward, then accountability has not been delegated to the machine — it has been misplaced in it. The point of the FairByDesign Doctrine is that legitimacy doesn't require us to understand these systems perfectly. It requires that someone answers for what they do, and that the record survives the moment it is needed.

A note on sources

The doctrine builds on established work: Jenna Burrell on the opacity of machine learning; Helen Nissenbaum and A. Feder Cooper and colleagues on the "many hands" problem; Andreas Matthias on the "responsibility gap"; Mark Bovens on what accountability actually means; Inioluwa Deborah Raji and colleagues on closing the AI accountability gap through auditing; the OWASP Top 10 for LLM Applications and recent work on runtime governance for agentic AI on the security surface; and the governance instruments now in force — the NIST AI Risk Management Framework and the EU AI Act. Full, verifiable references appear in the Field Paper.

Read the full Field Paper

This essay is the doorway. The full *FairByDesign Doctrine* Field Paper carries the complete three-pillar architecture, the five-level evidence hierarchy, the role-by-role implications, Evidence Notes mapping every claim to its source, and full APA

references. It is Paper III of the Project Sentinel series, following *The Accountability Gap* and *The Accountability Control Plane*. Read it, and bring the Reconstruction Test to your next governance review.