

The Accountability Gap: Who Answers for a Single Automated Decision?

Causum

Abstract

Enterprises are learning to govern AI systems in aggregate: policies, monitoring, audit programs, model lifecycle controls. Almost nothing in the current stack governs the unit that actually causes harm — the *individual decision*. When one automated determination goes wrong, the question that follows is never statistical. It is attributive: *whose failure was this?* Was the encoded policy inadequate? Did the execution machinery err? Were the inputs incomplete? Or did a sound process reach a wrong answer that nothing could have caught? We argue that no mechanism in today’s AI stack — including our own — can answer this question, because the artifacts it retains are either nothing (a discarded context) or everything (a transcript too large to keep and too shapeless to adjudicate). We work from first principles toward what an answer would require: a record of the single decision, small enough to retain for every decision, structured enough for a party who was not present to assign responsibility. We do not yet know the minimal contents of such a record. This paper’s contribution is the problem statement, a set of desiderata, and an honest account of why the obvious candidates fail.

1 The Unit of Harm Is the Decision

The AI governance conversation is conducted in aggregates. Model risk is assessed over distributions. Monitoring detects drift across thousands of outputs. Compliance programs certify systems, not determinations. Every instrument in the current toolchain answers questions of the form “*is the system behaving acceptably?*”

But harm does not arrive in aggregate. It arrives as one denied claim, one mispriced contract, one wrong dispatch — one decision, relied upon, propagating consequences before any dashboard updates. In our work on decision-time governance [1], we argued that organizational knowledge must participate in decisions rather than observe them. That argument concerned *prevention*. This paper concerns the harder residue: when a decision has gone wrong anyway, what does it take to *answer for it*?

The question after a failure is attributive, and it has a short list of possible answers:

1. The encoded standard was wrong or incomplete — the policy’s author answers.
2. The decision machinery misapplied a sound standard — the system’s operator answers.
3. The inputs were missing, stale, or corrupt — the environment answers.
4. The standard was sound, the machinery honest, the inputs complete — and the outcome was wrong anyway. Nobody answers, and being able to *establish* that is as important as the first three.

We know of no deployed system that can distinguish these four cases for a single past decision. We include our own platform in that assessment: our evidence trails record what a decision consulted and concluded, but we cannot today hand a skeptical third party one artifact from which they could independently conclude “the standard’s author answers for this one.” The fourth verdict — *nobody* — we cannot establish at all, and we suspect it is the keystone: an attribution procedure that can never conclude “everyone did their job” will be read, correctly, as a culprit-hunting procedure, and none of its verdicts will be trusted.

2 Why the Current Artifacts Cannot Answer

Consider what survives an automated decision today, in ascending order of volume.

2.1 Nothing

Most inference leaves no artifact beyond its output. The context is discarded; the intermediate reasoning was never externalized. Attribution is not merely difficult; its subject matter no longer exists.

2.2 Logs

Operational logs record that a decision occurred, when, and what it returned. They support forensics of the *system* (was it up? was it slow?) but not of the *determination*: a log line asserts nothing about what standard the decision was held to, or whether it was honored.

2.3 Transcripts

At the opposite extreme, one may retain everything: the full prompt context, every intermediate token, every tool result. We have experimented in this direction and report two structural failures.

First, **volume**. Complete traces of nontrivial automated decisions run to hundreds of kilobytes each. At enterprise decision volumes, full retention is a storage program, not an accountability program — and in practice, a record too costly to keep is not kept, which returns us to retaining nothing. We suspect there is a deep principle here: *the economics of the record determine whether accountability exists at all*.

Second, and less obviously, **shapelessness**. A transcript records utterances, not commitments. Reading one, an auditor can see what the system said to itself; they cannot see which of those statements the outcome actually rested on, which were idle, which standard was in force, or what would count as the standard having been violated. The transcript is evidence in the way a warehouse of unsorted paper is evidence: admissible, perhaps, but not probative. Volume without structure does not answer attributive questions; it buries them.

2.4 Explanations

A popular middle course asks the model to explain its decision, and retains the explanation. We regard this as the least trustworthy artifact of the four. A generated explanation is produced by the same process whose reliability is in question, after the fact, with every

incentive toward plausibility. Recent findings that model-stated reasoning can diverge from the factors that actually drove the output [4] formalize a suspicion any auditor would arrive at unaided: self-report is not a record. It is testimony — from the party under review.

3 Desiderata for a Decision Record

Working from first principles, we can state properties any adequate answer must have. We can state them; we cannot yet construct an artifact that has them. That asymmetry is the honest condition of this paper.

- D1. Per-decision.** One record per determination, produced as a byproduct of deciding — not reconstructed on demand, which reintroduces self-report.
- D2. Bounded size.** Small enough that retaining one per decision, forever, is an unremarkable cost. Our intuition is that this bound is measured in kilobytes, not megabytes, which immediately rules out transcripts and forces a selection problem we do not know how to solve: *which* few facts about a decision carry its accountability?
- D3. Standard-bearing.** The record must identify the standard the decision was held to — the policy, the encoded knowledge, the intent — in a form that cannot be quietly revised after the fact. An attribution made against an editable standard attributes nothing: the author of a bad policy will, quite rationally, improve the policy and then dispute the history. We suspect content fingerprinting is the right primitive here, but what exactly must be fingerprinted — the standard’s text? its interpretation? the inputs? — we have not determined.
- D4. Conduct-bearing.** The record must say something checkable about *how* the determination was reached — enough that misapplication (case 2) is distinguishable from a bad standard (case 1). What is the minimal statement of conduct? We do not know. A full trace is too much; the bare output is too little; the interesting object lies somewhere between, and we lack even the vocabulary to name it.
- D5. Completeness-bearing.** The record must state what the decision could and could not see, so that case 3 (the environment answers) is detectable. An absent input and a consulted-but-unhelpful input must not read identically.
- D6. Third-party legible.** A reader who was not present — an auditor, a counterparty, a regulator — must be able to walk from the record to one of the four verdicts without trusting the producer’s narration. In the limit, we would want elements of the record to be *recomputable* by the reader rather than merely asserted; how much of a decision’s context can be made recomputable is, to our knowledge, an entirely open question.
- D7. The fourth verdict.** The record must make “nobody answers” an establishable outcome, not a default reached by exhaustion.

We note an uncomfortable interaction among D2, D4, and D6: size pushes toward summarization, conduct-bearing pushes toward detail, and third-party legibility forbids the summarizer from being the interested party. Whether these three can be satisfied simultaneously is the crux, and we do not know the answer.

4 What This Is Not

It is worth distinguishing the object we are groping toward from artifacts that already exist.

It is not an *audit log*; logs establish occurrence, not conduct against a standard. It is not a *model card* or system documentation; those describe systems in aggregate, and the gap we have named lives precisely below the aggregate. It is not an *explanation*; see above. And it is not the software supply chain’s attestation object [5], though that is the nearest neighbor we have found: build attestations bind an artifact to the process that produced it, with content digests, for consumption by a distrustful verifier. The analogy is instructive and the machinery is suggestive — but a build is deterministic and its “standard” is a recipe, whereas a decision is probabilistic and its standard is knowledge. What survives the analogy and what breaks is, we think, a research question of the first rank.

5 The Economics of Answering

Our cost analysis of ungoverned AI [2] argued that organizations pay for oversight in proportion to decision volume. The accountability gap adds a second, lumpier cost: when a consequential decision fails today, the investigation is archaeological — engineers reconstructing context from logs, screenshots, and memory, at exactly the moment trust is lowest. Interviews with practitioners suggest such reconstructions are measured in weeks. A decision record satisfying D1–D7 would convert that archaeology into a reading. We can estimate what the archaeology costs; we cannot yet build the record. That is the gap in one sentence.

6 Conclusion

Aggregate governance is necessary and maturing. Decision-level accountability does not exist. The artifacts we retain are either absent, unstructured, or self-reported, and none can carry the four-way attribution that follows real failures. We have stated desiderata for the missing object — per-decision, bounded, standard-bearing, conduct-bearing, completeness-bearing, third-party legible, capable of concluding “nobody” — and we confess we do not know its construction. Our current systems prevent classes of bad decisions; they cannot yet answer for the ones that get through. Closing that gap is, in our view, the next fundamental problem of decision-time governance, prior to and harder than better prevention. We invite correspondence from teams who believe they have any piece of it.

Contact:

Causum

www.causum.com

contact@causum.com

References

- [1] Causum. *Mars[®]: Enforceability as a Property of Data*. Causum
- [2] Causum. *Mars[®]: The Hidden Costs of Enterprise AI*. Causum

- [3] Causum. *Mars[®]: Structured Knowledge for High-Stakes Decision Intelligence*. Causum
- [4] Turpin, M., Michael, J., Perez, E., Bowman, S. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *NeurIPS*.
- [5] Torres-Arias, S., et al. (2019). in-toto: Providing Farm-to-Table Guarantees for Bits and Bytes. *USENIX Security*.
- [6] European Union. (2024). Regulation (EU) 2024/1689: Artificial Intelligence Act.
- [7] National Institute of Standards and Technology. (2023). AI Risk Management Framework 1.0.