

BEYOND THE AUDIT TRAIL

Nobody Approved It: Who Is Accountable When the Agent Decides

Why the governance we built for software that executes decisions fails for software that makes them — and where accountability actually has to live.

Giovanni Leonardi

April 2025



Nobody Approved It: Who Is Accountable When the Agent Decides

Transformation · No. 413
April 2025

“Accountability lives upstream, at the unglamorous moment when someone decides how much authority to hand a machine and puts their name to the decision.”

1. The Question Asked Too Late

The incident review had everything it needed except an answer. On the screen was a complete trace: every step the agent had taken, each action stamped with the time it happened, the rule that triggered it, and the records it had read. The reconciliation had run for six weeks without a hand touching it. Somewhere in the fifth week it had settled a run of invoices it should never have paid — well-formed invoices from a supplier account that had quietly been compromised, each one matching a purchase order and a goods-receipt note exactly as the mandate required. The loss was a little over four hundred thousand pounds across thirty-one payments, none individually large enough to trip the value threshold.

Nine people sat around the table with a flawless record of what had happened, and not one of them could say who was accountable for it. Someone asked the obvious question — who approved these payments? — and the honest answer was nobody. The agent had approved them, inside a mandate that permitted it to. The engineer who had written that mandate had moved teams. The finance lead who owned the process had never read it. The audit trail, which everyone had been assured was the heart of responsible AI, documented in perfect detail a decision that belonged to no one.

I have sat in more than one version of that room over the past year, and the pattern is always the same. The logs are immaculate. The accountability is missing. And the reason, I have come to think, is that we are asking the accountability question at the wrong moment — after the machine has acted, when it can only ever be answered too late.

2. Governance Built for a World That Has Inverted

The apparatus most organisations use to govern software was designed for software that executes decisions rather than makes them. A human being weighed the case, exercised judgement, and pressed the button; the system did as it was told and wrote down what it did. In that world the accountability question has a clean answer, because there is always

a person standing at the point of decision. Approval workflows, segregation of duties, sign-offs, the whole edifice of controls we inherited — all of it rests on the assumption that a human judgement sits underneath each consequential act, and that the record exists to prove that judgement was made.

Autonomous agents invert the arrangement. The system now makes the decision and the human, at best, reviews a sample after the fact. The judgement that used to sit under each act has been lifted out and replaced by a policy written once, in advance, that then fires thousands of times without anyone watching any particular instance. We did not merely speed the old process up. We removed the thing the entire accountability model was anchored to — and then kept the model.

This is why the frameworks we all dutifully adopted have proved so oddly hollow in practice. The risk-management standards, the management-system certification, the tiered obligations now coming into force across the European market — they tell an organisation what to consider: assess your risks, document your controls, keep a human in oversight, maintain your logs. What almost none of them force an organisation to do is answer the operational question that actually decides whether accountability exists: for this specific agent, acting in this specific process, who has decided where it may act unsupervised, and who is answerable for that decision? The frameworks describe the shape of good governance. They are silent about the person. And so, again and again, they were implemented as documentation exercises — a policy on the shared drive, a completed template, a green square on a maturity dashboard — while the live question of who owns the machine's authority went unasked.

3. Logging Is Not Accountability

The most seductive substitute for accountability, and the one I have watched more organisations reach for than any other, is observability. If we log everything — every prompt, every tool call, every record read and written — then surely we are governed. The instinct is understandable. A complete trace feels like control. But a trace answers the question what happened, and accountability answers the question who is answerable, and these are not the same question. They are barely even related.

“A perfect record of an unowned decision is not governance. It is a well-documented orphan.”

The reconciliation failure was, from an observability standpoint, a triumph. Every action was explained. You could reconstruct the agent's reasoning to the keystroke. And that pristine record was precisely what made the silence in the room so uncomfortable, because it proved that transparency and accountability had come apart. We had built the ability to see everything the machine did and, in doing so, quietly convinced ourselves we had built the ability to answer for it. Observability tells you where to look after something has gone wrong. It cannot tell you who should have been watching, because watching was never the point once the human left the loop. The point was authority — and authority is

decided long before the log is written.

4. Accountability Is Provisioned, Not Attributed

Here is the position I have come to hold, and it is a single, simple inversion. Stop trying to locate accountability after an autonomous agent acts. It is not there and it never will be. Accountability for an autonomous system can only be established at the moment its authority is granted — provisioned into the delegation, not attributed to the outcome.

Accountability cannot be attributed to an autonomous system after it acts. It can only be provisioned into the grant of authority before it does. The mandate is the new anchor.

Every autonomous agent runs on a mandate, whether or not anyone has written it down honestly: the scope of what it may do, the conditions under which it may do it, the limits it must not cross, and the routes by which its authority can be paused or withdrawn. That mandate is not a technical artefact. It is a delegation of authority from an organisation to a machine, and like every delegation of authority in the history of institutions, it has an owner — a named human who decided that this agent may do this thing, up to this boundary, under these conditions. When that ownership is explicit, accountability exists, because there is a person who can be asked not did you approve this payment but did you scope this agent's authority appropriately, given what was at stake. When that ownership is absent — when the mandate emerged from a sprint, or a default configuration, or a well-meaning experiment that was never decommissioned — there is no accountability to find, no matter how complete the logs.

Concretely, a provisioned mandate is one that is:

- Owned — a specific, current, named person is answerable for the grant of authority, and knows they are. Ownership that has left the team is not ownership.
- Narrow — the agent's authority is scoped to the smallest surface that lets it do its job, so that the space in which it can act unsupervised is deliberately chosen rather than inherited from what the model happens to be capable of.
- Legible — the boundary is written in terms the owning business can actually reason about, not buried in a system prompt or a configuration file only the engineers can read. A mandate no one but its author understands cannot be owned by anyone else.
- Revocable — there is a fast, rehearsed way to narrow or withdraw the agent's authority without shutting down the business around it, and someone knows how to use it.

The reconciliation agent failed on every count. Its authority to settle, not merely to match, had been granted almost casually, because settlement was the obviously useful next step and the matching had worked so well. There was no anomaly or novelty gate, so a supplier account behaving in a way it never had before looked identical to one behaving normally. And the mandate had no living owner — which is why the room could reconstruct the failure perfectly and still not answer for it. The agent did not malfunction. It did exactly what it was authorised to do. The failure was that the authorisation had been written as though the agent were a faster clerk rather than a different kind of actor,

and no one had put their name to that.

5. The Objection Worth Taking Seriously

The strongest challenge to this position is not that it is wrong but that it merely moves the problem. If a human scopes an agent's authority and the agent then does something unforeseen but entirely within scope, are we not holding that human accountable for an outcome they could not possibly have predicted? That seems both unfair and unlikely to survive contact with a real disciplinary process. And if the honest answer to unpredictability is to keep a human in the loop at the point of consequence, then we have simply abolished the autonomy we were trying to govern. Either autonomy or accountability, the objection runs — pick one.

I take this seriously because it is half right. It is right that you cannot hold anyone answerable for clairvoyance. It is wrong about what accountability has ever meant. We already, routinely, hold people accountable for decisions with unforeseeable consequences — every manager who has ever delegated to a junior does so. When a capable person is given a task and makes a call that goes wrong in a way no one saw coming, we do not hold their manager accountable for failing to predict it. We hold the manager accountable for the quality of the delegation: was the scope appropriate to the person's judgement, were the limits sensible, was the reporting line clear, could the manager step in when the stakes rose. That is a real, assignable, and fair standard, and it is exactly the standard that transfers to machines.

The owner of an agent's mandate is not answerable for the agent's every action. They are answerable for whether the authority they granted was appropriately bounded, monitored, and revocable given what was at stake — the same question we ask of anyone who delegates. The accountability vacuum only appears if you insist on the old definition, where being accountable means being the person who pressed the button. Give that definition up, and the vacuum closes. The person who decided a machine could settle payments unsupervised, with no anomaly gate and no owner, made a poor delegation. That is not clairvoyance we are demanding of them. It is judgement — and judgement is exactly what we have always held delegators to account for.

6. What This Actually Asks of Us

None of this requires a new framework, and I would be wary of anyone selling one. It requires a change in where we look. The instinct of the past two years has been to answer the anxiety about autonomous systems by pointing downstream — at richer logs, better observability, more detailed traces, the reassuring machinery of after-the-fact reconstruction. All of it is useful and none of it is accountability. Accountability lives upstream, at the unglamorous moment when someone decides how much authority to hand a machine and puts their name to the decision.

The organisations that will come through this well are not the ones with the most complete audit trails. They are the ones that can answer, for every agent they have put into production, a short and awkward set of questions: Who owns this agent's authority?

How narrow is it, and why that narrow? Can the business read the boundary without an engineer to translate? And if we needed to pull its authority back this afternoon, do we know how? The failure I keep watching is not that machines are deciding things. It is that we keep asking who is accountable after they decide, when the only honest moment to settle that question was before we ever let them.

The machine deciding is not the problem to be governed. The ungoverned grant of authority is. And that is a decision a person makes, at a particular moment, that a person can be named for — if we choose to ask the question while it can still be answered.