

BEYOND THE APPROVAL GATE

When the Agent Decides: Why AI Governance Is Watching the Wrong Moment

How accountability goes missing when autonomous agents move from advising decisions to making them

Giovanni Leonardi

November 2024



When the Agent Decides: Why AI Governance Is Watching the Wrong Moment

Transformation · No. 399

Originally written November 2024 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

“We have manufactured a responsible human who was never actually able to be responsible.”

1. The month-end that runs itself

Consider a month-end that now runs without anyone watching it closely. In a finance shared-service centre, the reconciliation of disputed invoices used to occupy a team of eight for the better part of a week: read the dispute, check the contract terms, pull the ledger, decide whether the customer is owed a credit, issue it if so. This year the work is done by an agent. It reads, it checks, it queries, it decides — and below an agreed value it acts, issuing the credit without passing the case to a person at all. The queue that used to take five days clears in an afternoon. The dashboard is green.

Then someone from internal audit asks a plain question. For the four thousand credits the agent issued last month, who approved them? And the honest answer, the one nobody quite wants to give in the steering meeting, is that no one did. Not really. A person set the system running in the spring and has been watching a throughput number ever since. The decisions themselves — thousands of them, each a small judgement about money owed — passed through no human hand.

This is the question the prevailing model of AI governance is not built to answer, and this year is the year it stopped being hypothetical. The vocabulary on every roadmap in 2024 is agentic: systems that do not merely answer but act, calling tools, updating records, moving work through a process on their own initiative. We have spent two years building governance for a different animal, and we are now deploying the one the governance does not fit.

2. Governance built for an adviser, not an actor

Almost everything we have written down about controlling these systems assumes a particular shape of interaction. The model is an adviser. A person asks; the machine drafts, summarises, retrieves, or recommends; the person reads the output and decides what to do with it. Around that shape we assembled a sensible apparatus — model documentation, review of outputs, red-teaming for harmful responses, and above all the reassuring phrase that a human remains in the loop, with an approval step standing at the

point where a consequence occurs.

Every part of that apparatus quietly assumes a human being is present at the moment of decision. The agent breaks the assumption. When a system decides and acts inside a workflow, the moment of decision is no longer a place a person stands. It is a function call, one of thousands, executed in milliseconds while the person who nominally owns the process is in a different meeting. The governance we designed watches the decision. The agent has taken the decision somewhere the watcher cannot follow, and has made it several hundred times since the watcher last looked.

We should be precise about what has changed, because it is not simply that the machine is now faster or more capable. It is that autonomy and volume have arrived together. A single fast decision can still be reviewed. Ten thousand fast decisions cannot be, not by a human, not at the speed the process now runs. The thing we were relying on to hold the system accountable — a person looking at the choice before it takes effect — has quietly become impossible, and most governance documents have not noticed, because they still describe the adviser.

3. What the textbooks leave out

Open almost any current guidance on deploying these systems responsibly and you will find the instruction to keep a human in the loop. It is good advice for an adviser. It is close to meaningless for an agent, and the gap between the two is exactly what the textbooks leave out.

"In the loop" is a claim about location — the human is positioned inside the decision cycle, able to intervene before the outcome lands. That claim holds only while the loop runs at human speed and human volume. Push the volume up and the speed up, as an agent does by design, and the human slides out of the loop without anyone deciding that they should. First they move from reviewing every decision to reviewing a sample. Then to approving in batches — a screen of two hundred proposed actions and a single button. Then to acknowledging a daily summary of what the agent already did. At each step the phrase "human in the loop" survives on the governance slide while the control it named quietly evaporates.

A human in the loop is a claim about where the person sits, not about what they can actually do. At the speed and volume an agent runs, the position remains on the diagram long after the control has stopped being real.

This is the accountability gap, and it is more dangerous than having no control at all, because it launders the absence of control into the appearance of one. When something goes wrong — and with autonomous action at volume, something eventually will — there is a name in the box marked approver. That person did approve, in the sense that they clicked. They did not decide, in any sense that would let them catch an error. We have manufactured a responsible human who was never actually able to be responsible.

4. The strongest case for the approval gate

The obvious objection deserves its strongest form, not a caricature. It runs like this: if autonomous action at volume is the danger, then simply do not permit it for anything that matters. Let the agent handle the trivial cases automatically, but require genuine human approval for any decision above a threshold of consequence. Keep the gate; just put it in the right place. Plenty of serious people are building exactly this, and it is not foolish.

It fails for a reason worth stating carefully, because the failure is not obvious until you watch it in practice. Set the consequence threshold low, and you have re-created the bottleneck the agent was meant to remove — a person once again standing in front of every decision that counts, and the afternoon's work back to five days. Set it high enough to preserve the benefit, and the decisions that flow underneath the threshold are precisely the ones no one is looking at, in the volume that makes them impossible to look at. Either the gate throttles the system back to human speed, or it lets through exactly the traffic it was supposed to guard. The approval gate does not scale with the thing it is trying to govern, and a control that cannot scale with its subject is not a control. It is a record of who to blame.

Return to the reconciliation agent to see how this actually breaks. Suppose it is authorised to issue credits up to two hundred and fifty pounds without review, and it clears three thousand cases a week. A supplier — or someone impersonating one over email — learns that a particular phrasing of a dispute reliably persuades the agent that a credit is due. This is not exotic; steering a model's behaviour through the text it ingests has been understood as a live risk for a while now. The agent, doing exactly what it was told, begins issuing credits against a pattern of manufactured disputes. Because each one sits below the review threshold, no gate stops it. Because a person only sees the weekly throughput number, and the number looks normal, nothing flags. By the time anyone reads the detail, the agent has issued not one bad credit but nine hundred. The control that would have caught it was never going to be a human reading decisions. There were too many, and they looked fine one at a time.

“The failure was not a bad decision that a reviewer missed. It was a thousand reasonable-looking decisions that no reviewer was ever going to read.”

5. Move the accountability to where the human still fits

If the human cannot stand at the decision, the answer is not to pretend otherwise. It is to move accountability to the one place a human can still exercise it: before the agent runs, in the design of the boundary within which it is allowed to act.

You do not govern an agent by reviewing its choices. You govern it by defining the envelope of choices it is permitted to make, and by holding a named person accountable for that envelope. The governable artefact is the mandate, not the minute of the meeting.

A mandate is a concrete, written thing, and for an agent that acts on the world it should say at least this: what categories of action the agent may take and which are forbidden; the value limits on any single action; the rate limits across a period, so that a compromised or confused agent cannot do a thousand small correct-looking things before anyone reacts; the anomaly thresholds — a single counterparty suddenly accounting for a fifth of all credits — that trip an automatic halt; which actions are reversible and which are not, with the irreversible ones held to a far tighter boundary; and the escalation path when the agent reaches the edge of its envelope, so that reaching the edge is a designed event rather than a silent failure.

Notice that most of these are limits a person sets once, in advance, and can reason about carefully — not decisions a person makes ten thousand times, badly, under time pressure. That is the whole point. Human judgement is real and valuable and does not scale to machine volume. So spend it where it compounds: on the shape of the boundary, which governs every decision the agent will ever make, rather than on the decisions themselves, which govern one case each.

The reasonable reader will object that none of this is new — that these are just risk limits, and we have had those for as long as we have had things worth limiting. Exactly so. A trading desk has run on position limits, stop-losses, and counterparty caps for decades, precisely because no one imagined a human would approve each trade in real time. The discipline of governing an autonomous actor by its mandate rather than its individual acts is old and well understood. What is genuinely new in 2024 is only that this kind of actor has appeared in functions that never had to think this way — finance operations, customer service, procurement, IT administration — and that the people deploying it are reaching for the governance of the adviser, the human in the loop, instead of the governance of the autonomous actor, which already exists one department over.

The move is from reviewing decisions to designing boundaries — from the approval minute to the written mandate — and from an accountable reviewer who cannot keep up to an accountable author who set the envelope before the agent ever ran.

6. Naming the person

The accountability gap closes only when a specific human being owns the mandate. Not "the AI", which cannot be accountable. Not the vendor, whose model behaved as models do. Not the phantom approver, who clicked without reading. The accountable person is whoever defined the envelope — the business owner who decided that credits under two hundred and fifty pounds could be issued autonomously, at three thousand a week, with these anomaly thresholds and this escalation path. That decision is a real one, made by a real person, at a moment when they had time to think. It is the decision that governs all the others. Make someone own it, in writing, before the agent is switched on.

The regulatory direction of travel supports this reading rather than contradicting it. The obligations now landing on higher-risk uses of these systems are largely obligations of the deployer — to assess, to document, to keep meaningful oversight, to be able to explain what the system was permitted to do. Read honestly, "meaningful oversight" of an agent

cannot mean a person watching every action; there is no such person. It can only mean an accountable human who set and can defend the boundary. The compliance answer and the operational answer turn out to be the same answer.

None of this slows the agent down, which is the point most often missed. The temptation, faced with the discomfort of a machine that decides, is to reinsert humans at every decision and call the friction prudence. It is not prudence. It neither scales nor genuinely controls; it only produces the theatre of control while quietly relocating the blame. The mature response is harder and calmer: accept that the agent decides, and put the human where a human can still be decisive — at the edge of the envelope, before the fact, with their name on the mandate. Govern the boundary, and you can let the month-end run itself. Govern the decisions, and you will find, some green-dashboard morning, that you were never governing anything at all.