

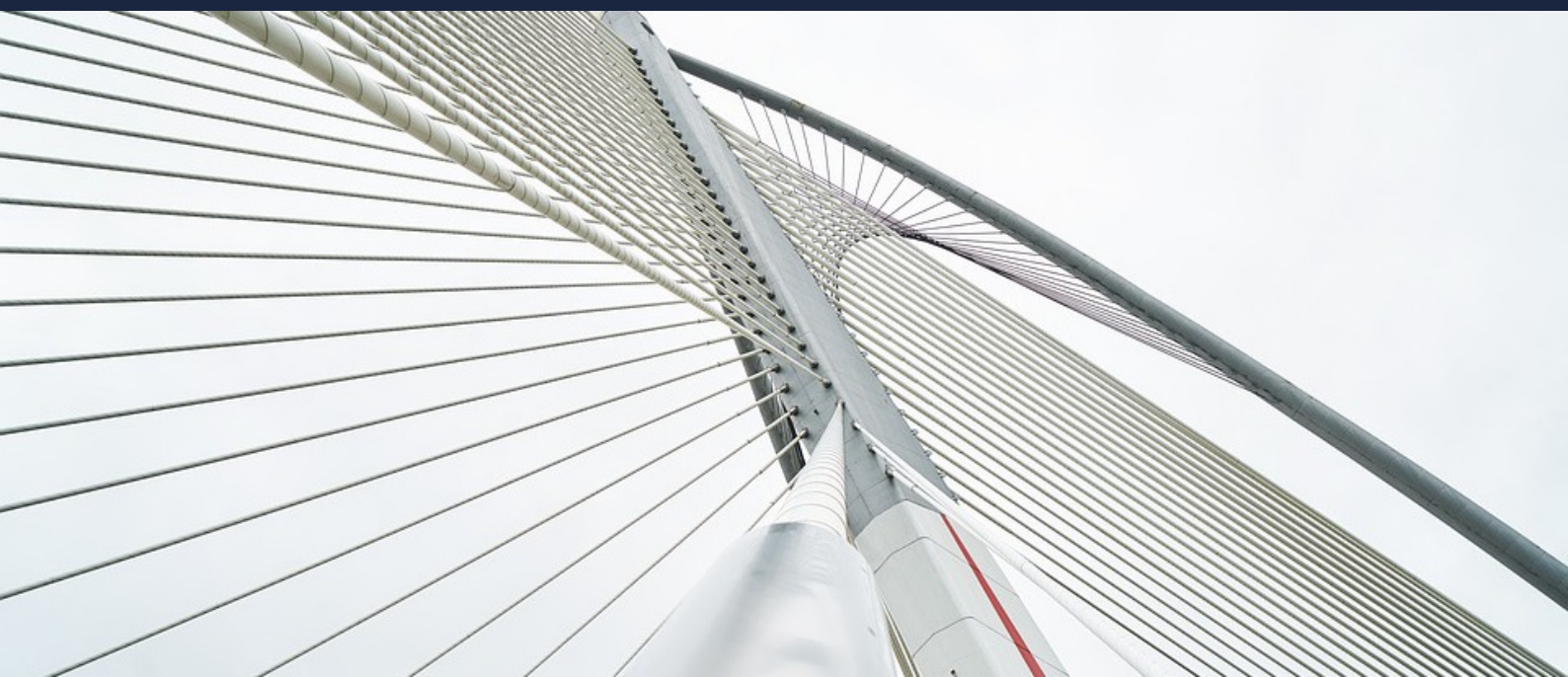
BEYOND AFTER-THE-FACT REVIEW

Governing What You Cannot See — AI Oversight in Autonomous Programmes

A Practitioner's Perspective on Holding Autonomous Delivery to Account

Giovanni Leonardi

July 2025



Contents

1. Executive Summary	3
2. The Governance Assumption That No Longer Holds	4
3. Why Traditional Oversight Fails Against Machine Speed	4
4. Three Inadequate Responses	5
5. Governing the Envelope, Not the Instance	6
6. The Four Controls of Envelope Governance	6
7. Accountability When No Human Made the Decision	7
8. What Boards and Sponsors Must Do Now	8

Governing What You Cannot See — AI Oversight in Autonomous Programmes

Leadership · No. 104
July 2025

“By the time a monthly board reviews an agent's decisions, the agent has made a hundred thousand more.”

1. Executive Summary

Autonomous agents have crossed a threshold. Where software once waited to be told what to do, a new generation of AI agents now interprets intent, plans its own sequence of actions, and executes across a delivery programme without a human approving each step. Within a single hour, an agent coordinating a workstream may take thousands of consequential decisions — reprioritising tasks, reallocating effort, escalating or suppressing risks, committing the programme to dependencies. The governance machinery we rely on was never built for this. It was built for human-paced decision-making, where the interval between a decision and its review could be measured in days or weeks, and where a steering committee that met once a month was a reasonable heartbeat for control.

This paper argues that the profession's reflex — to govern autonomous agents the way we govern people, by reviewing their decisions after the fact — is not merely inadequate but actively hazardous. By the time a monthly board reviews an agent's decisions, the agent has made a hundred thousand more. The gap between the speed of action and the speed of oversight has widened into a governance vacuum.

The answer is neither to slow the machine to human pace nor to abandon oversight as futile. It is to move the point of control. Governance must shift from the instance — the individual decision — to the envelope: the bounded space of decisions an agent may take autonomously, the conditions under which it must stop and defer to a human, and the continuous assurance that it remains inside those bounds. This paper sets out why traditional oversight fails against machine speed, examines three inadequate responses the profession is currently reaching for, and proposes a model — envelope governance — resting on four controls: bounded authority, continuous assurance, tripwires and deferral, and an unbroken chain of human accountability. It closes with the specific actions boards and programme sponsors should take now.

You cannot govern what you cannot see, and you can no longer see what an autonomous agent is doing at the speed it is doing it. Oversight must therefore move upstream — to the boundaries you set before the agent acts, not the decisions you review after.

2. The Governance Assumption That No Longer Holds

Every governance framework the profession has built rests on a single, unspoken assumption: that a human makes each decision that matters, and that another human can review it. Stage gates, approval thresholds, steering committees, change control boards, monthly reporting cycles — all of them are instruments for inserting human judgement between a proposed action and its execution, or for inspecting a human's judgement shortly after the fact. The cadence of governance was tuned to the cadence of human work.

That assumption has quietly collapsed. An autonomous delivery agent does not propose an action and wait; it acts, observes the result, and acts again, in a loop that turns over faster than any committee could convene. The decisions it makes are individually small — which task to sequence next, which of two integration approaches to attempt, whether a minor risk warrants escalation — but they are consequential in aggregate, and they accumulate at a rate that no human reviewer can track. The pattern I have observed across early adopters is consistent: the technology is deployed with enthusiasm, a governance forum is nominally made responsible for it, and within weeks that forum discovers it is being asked to approve or review decisions it cannot even enumerate, let alone understand.

This is not a failure of diligence. It is a category error. We are applying a control model designed for a world in which decisions are scarce and slow to a world in which they have become abundant and fast. The instruments still turn — the committee still meets, the report is still produced — but they no longer touch the thing they are meant to control.

3. Why Traditional Oversight Fails Against Machine Speed

Three properties of autonomous agents defeat the traditional model, and it is worth being precise about each, because the remedy follows from the diagnosis.

- **Volume.** The sheer number of decisions exceeds human reviewing capacity by orders of magnitude. Sampling — reviewing a handful of decisions and inferring the health of the rest — is the natural fallback, but sampling assumes the population is homogeneous and the errors are random. Agent behaviour is neither: a single flawed inference can propagate through thousands of downstream actions, and the one decision you did not sample is precisely the one that mattered.
- **Opacity.** Even when a specific decision is surfaced for review, its rationale is often not legible in the terms a governance forum uses. The agent did not follow a documented procedure a reviewer can check it against; it produced a judgement from a model whose reasoning cannot be fully reconstructed. Asking "why did it do

that?" frequently yields an answer that is plausible, fluent, and unverifiable.

- **Speed.** The interval between decision and consequence has collapsed. In a human programme, a questionable decision on Monday might be caught at Thursday's review before it did real damage. An agent's questionable decision is acted upon in milliseconds and has already shaped the next thousand decisions before any human is aware it was made.

Taken together, these properties mean that after-the-fact review of individual decisions cannot work. Not that it works poorly — that it cannot work at all, because the reviewer is structurally too slow, too under-informed, and too outnumbered to intervene before harm is done. Any governance model that depends on inspecting the instance is already defeated.

4. Three Inadequate Responses

Faced with this, organisations are reaching for three responses. Each is understandable, and each fails.

Response	The instinct behind it	Why it fails
Human-in-the-loop everywhere	Restore the human approval we are used to	Reintroduces human latency into every loop, destroying the value of autonomy and creating rubber-stamping when volume overwhelms the human
Ban or freeze	If we cannot govern it, we will not use it	Cedes ground to competitors and drives adoption underground, where it is ungoverned rather than un-adopted
Dashboard and trust	Build a monitoring view and hope visibility equals control	Confuses seeing with governing; a dashboard reports what has already happened and no one watches it in real time anyway

The human-in-the-loop-everywhere response is the most seductive because it feels responsible. But inserting a human approval into every agent action does not make the human a governor; it makes them a bottleneck who, under the pressure of volume, will approve without reading. We have simply moved the ungoverned decision from the machine to a human who is now decisions rubber-stamping at a rate that guarantees inattention. Worse, it obscures accountability: when something goes wrong, the human "approved" it, so the organisation believes it had control when it had only the appearance of it.

The ban-or-freeze response mistakes abstention for safety. In a competitive delivery environment, the capability does not disappear because one organisation declines to use it; it migrates to individuals who adopt it without sanction, producing exactly the ungoverned autonomy the ban was meant to prevent.

The dashboard-and-trust response is the most common and the most quietly dangerous, because it produces artefacts that look like governance. A real-time view of agent activity

is worth having, but a view is not a control. No one watches a dashboard continuously; it is consulted after an incident, at which point it is a forensic record, not an oversight mechanism.

"Seeing is not governing. A dashboard reports the past; a control shapes the future. The two are constantly mistaken for one another, and the mistake is where accountability quietly disappears."

5. Governing the Envelope, Not the Instance

If we cannot govern each decision, we must govern the space of decisions. This is the central move of the model I want to propose. Instead of asking "was this decision correct?" after the fact, governance asks, before the fact: "what is the bounded set of decisions this agent may make on its own, what must it never do, and under what conditions must it stop and hand back to a human?"

This is not a novel idea in principle. It is how we govern any capable actor whose individual actions we cannot supervise. We do not review every decision a treasury desk makes; we set position limits, define prohibited instruments, and require escalation above a threshold. We do not approve every clinical decision; we license practitioners to act within a defined scope and require referral beyond it. The envelope — the bounded authority within which autonomous action is trusted, and outside which it is forbidden or must be deferred — is the oldest instrument we have for governing at scale. What is new is the necessity of applying it to a non-human actor operating at machine speed, and the rigour that necessity demands.

The shift has a profound consequence for where governance effort is spent. Under instance governance, effort is spent downstream, reviewing outputs. Under envelope governance, effort is spent upstream, in the deliberate and demanding work of defining the envelope precisely — and then in continuously assuring that the agent remains inside it. The board's question changes from "show me the decisions" to "show me the envelope, show me that it is sound, and show me the evidence that the agent has not breached it."

6. The Four Controls of Envelope Governance

Envelope governance is not a single mechanism but four working together. Remove any one and the model fails.

1. Bounded authority. The envelope must be defined explicitly and in operational terms: the categories of decision the agent may take autonomously, the quantitative limits on each (how much effort it may reallocate, how far it may reprioritise, what it may commit the programme to), and the absolute prohibitions — the actions it must

never take under any circumstance. Vague authority is no authority; "use good judgement" is not an envelope. The discipline of writing the envelope down is itself the most valuable governance act, because it forces the organisation to decide, in advance and in daylight, what it is actually willing to delegate.

2. Continuous assurance. Because we cannot review instances, we must instrument the agent so that conformance to the envelope is checked continuously and automatically, by controls that run at machine speed alongside the agent rather than by humans after it. This means automated checks that flag when the agent approaches a limit, statistical monitoring of its decision patterns for drift, and independent verification — a second mechanism whose only job is to confirm the first is behaving. Assurance moves from periodic inspection to continuous attestation.
3. Tripwires and deferral. The envelope must specify, in advance, the conditions under which the agent must stop acting and defer to a human: novelty beyond its trained competence, low confidence in its own judgement, proximity to a hard limit, or any attempt to cross a prohibition. A tripwire is not a request for approval on a routine decision; it is a designed halt on an exceptional one. Done well, this concentrates scarce human judgement precisely where it adds value — on the genuinely novel or dangerous — instead of dissipating it across thousands of routine approvals.
4. An unbroken chain of accountability. Someone must own the envelope. Accountability cannot attach to the agent, which cannot be sanctioned, nor be allowed to evaporate into "the system decided." It must attach to the named human who set the envelope and the named human who owns the outcome. This is the control that makes the other three matter, and it is examined next.

7. Accountability When No Human Made the Decision

The hardest question autonomous delivery poses to governance is not technical but moral: when an agent makes a decision that causes harm, who is answerable? The tempting answers are all wrong. It is not the agent, which has no stake and cannot be held to account. It is not "the algorithm," which is an abstraction that conveniently belongs to no one. And it is not the human who nominally "approved" an action they could not actually evaluate — holding them accountable is both unjust and useless, because it punishes someone who was structurally unable to prevent the harm.

Envelope governance resolves this cleanly. Accountability attaches not to the decision but to the envelope. The person who defined the bounded authority is accountable for whether those bounds were appropriate. The person who owns the outcome is accountable for whether the envelope was the right one to grant. If an agent causes harm while acting inside a well-designed envelope, that is a signal the envelope was drawn wrongly, and the envelope-owner must answer for it. If it causes harm by breaching the envelope, that is a failure of the assurance controls, and their owner must answer. In every case there is a named human who is answerable, and the thing they are answerable for is something they actually controlled: the boundary, not the millisecond decision.

This is the principle that emerging AI regulation is, in its own language, beginning to reach toward — the insistence that meaningful human oversight and clear allocation of

responsibility must accompany autonomous systems. The organisations that will find that regulatory direction easy to meet are the ones that have already stopped pretending they can review the instance and have instead built genuine accountability for the envelope.

8. What Boards and Sponsors Must Do Now

The practical implication for anyone sponsoring or governing a programme that uses autonomous agents can be stated directly.

- Stop asking to review the decisions. The moment a governance forum accepts a report listing agent decisions for retrospective approval, it has accepted a fiction of control. Ask instead to see the envelope.
- Demand the envelope in writing, before deployment. No agent should act autonomously in a programme until its bounded authority, its prohibitions, and its tripwires are documented, understood, and formally owned. If no one can write the envelope down, the organisation does not understand the agent well enough to deploy it.
- Require continuous assurance, not periodic review. Insist that conformance is checked at the speed the agent operates, by automated and independent controls, and that the evidence of conformance is what is brought to the governance forum.
- Name the accountable owner of every envelope. Accountability that is not attached to a named person is not accountability. Make the envelope-owner and the outcome-owner explicit, and make clear what each will answer for.
- Govern the exceptions, personally. The tripwires are where human judgement still matters. Ensure the people who receive deferred decisions have the time, the context, and the authority to act on them — and are not so buried in false alarms that they ignore the real one.

The transition to autonomous delivery is not a reason to abandon governance; it is a reason to do governance better, and earlier, than we ever have. The organisations that treat oversight as something applied after the agent acts will find they are always one hundred thousand decisions too late. The ones that move their control upstream — to the envelope, defined in advance, assured continuously, and owned by a named human — will be the ones that can say, truthfully, that they govern what they cannot see.