

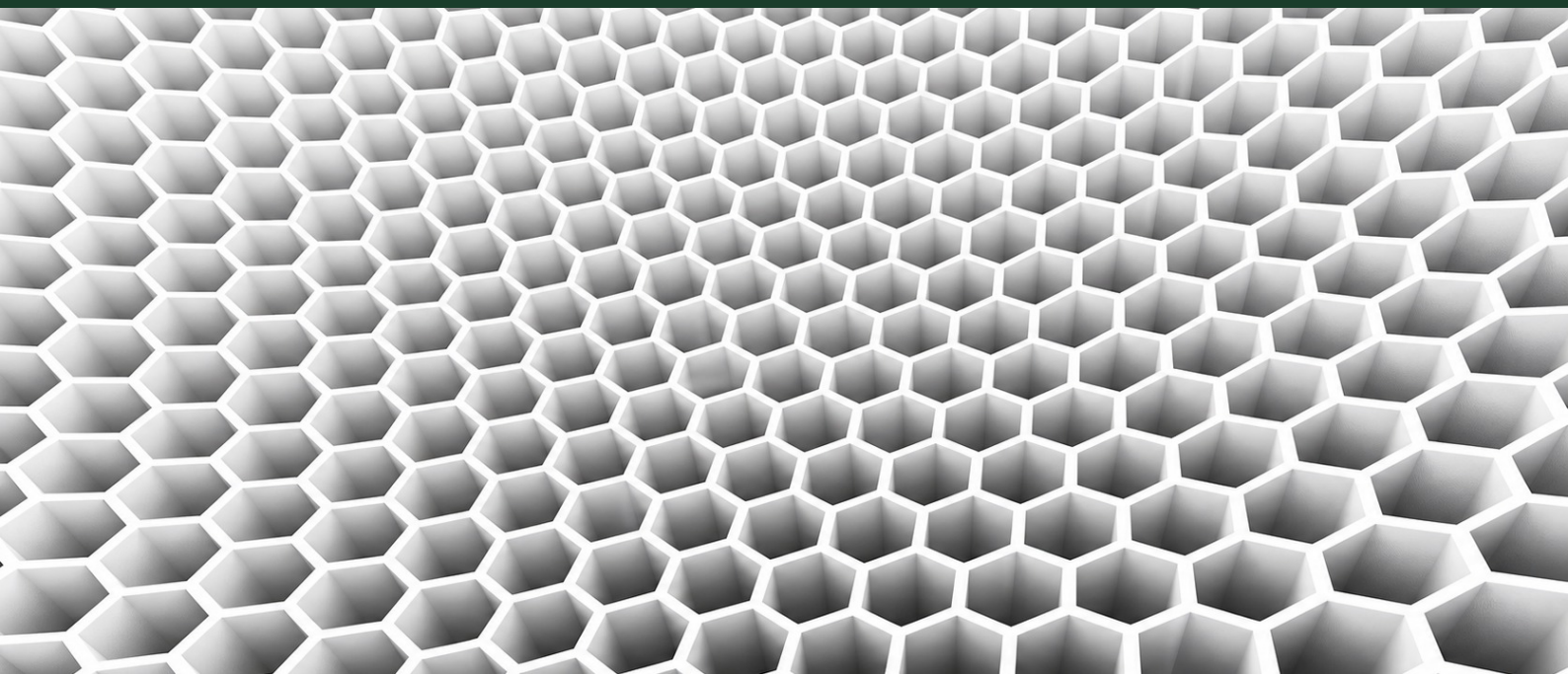
BEYOND THE TOOL

AI Agents Are Not Tools: Autonomous Decision-Makers in the Delivery Chain

A Practitioner's Perspective on Governing Autonomous Agents in Enterprise Delivery

Giovanni Leonardi

March 2026



Contents

1. Executive Summary	3
2. The Category Error at the Heart of Adoption	3
3. What the Tool Framing Actually Breaks	4
4. The Conditions That Made This Universal	5
5. What Was Tried, and What It Taught	6
6. The Options, Weighed	7
7. The Recommendation: Govern Authority, Not Output	7
8. Conclusion	8

AI Agents Are Not Tools: Autonomous Decision-Makers in the Delivery Chain

Programme · No. 251
March 2026

“An agent is not a tool because a tool does not choose.”

1. Executive Summary

Across the last eighteen months, organisations have brought autonomous software agents into the heart of their delivery chains — writing code, resolving tickets, reconciling accounts, drafting contracts, triaging incidents, and, increasingly, deciding what happens next without a person present at the moment of decision. The dominant framing for this adoption has been the tool: the agent as a faster instrument, a more capable version of the automation an organisation already understood. That framing is a category error, and it is producing a recognisable pattern of failure.

An agent is not a tool because a tool does not choose. A drill does exactly what the hand directs; a rules engine does exactly what its rules specify. An agent, by contrast, is given an objective and left to determine the means — which sub-tasks to pursue, which data to trust, which action to commit. It exercises discretion. The moment discretion enters the delivery chain, the questions that matter stop being questions of capability and become questions of accountability, authority, and control. Who authorised this decision? On what basis? Who is answerable when it is wrong?

This paper sets out the evidence for the tool framing's failure, examines the responses organisations have tried, and argues for a specific position: autonomous agents should be governed as accountable actors holding delegated decision rights, not as tools to be configured. The practical consequence is a shift from reviewing agent output to governing agent authority — defining, in advance and explicitly, what an agent may decide alone, what it must escalate, and who owns the outcome either way. Organisations that make this shift convert an ungoverned liability into a genuine extension of delivery capacity. Those that do not will continue to discover, incident by incident, that they automated a decision they never meant to delegate.

2. The Category Error at the Heart of Adoption

The tool framing did not arrive by accident. It was the path of least resistance, and three forces converged to lay it down.

The first was vendor language. The prevailing vocabulary — copilot, assistant, helper — deliberately positioned the technology as subordinate, an aid to a human who remained in charge. This was reassuring, and for the earliest use cases it was even accurate. But the vocabulary outlived its truth. As the same products acquired the ability to act — to call systems, execute transactions, and chain their own steps — the language did not change with them. Organisations went on describing as an assistant something that had quietly become an actor.

The second was procurement and governance machinery built for a different kind of software. Enterprises knew how to buy tools. They had licence models, security reviews, and change-control processes refined over decades for systems whose behaviour was, in principle, specifiable in advance. An agent slotted into that machinery as just another piece of software — and the machinery, designed to assess deterministic systems, had no place to record the one fact that mattered: that this system would make consequential choices its buyers could not fully anticipate.

The third was commercial pressure. The mandate to demonstrate a return on AI investment was intense, and autonomy was where the visible return lived. A tool that suggests saves a little time; an agent that completes saves a great deal. The incentive ran firmly towards granting more autonomy and faster, and firmly against the slower work of asking what that autonomy meant for accountability.

The tool framing survived not because anyone believed an agent was really just a tool, but because every part of the enterprise machine — how it buys, how it governs, how it justifies spend — was built to handle tools and had no slot for anything else.

The result is a systematic mismatch. The technology crossed the line from instrument to actor; the organisation's mental model, contracts, and controls stayed on the instrument side of the line. Almost every agent-related failure I have seen examined closely traces back to that unclosed gap.

3. What the Tool Framing Actually Breaks

It is worth being precise about the difference, because the precision is where the governance implications live.

Dimension	A tool	An autonomous agent
Behaviour	Specified in advance	Determined at runtime toward a goal
Discretion	None — executes instructions	Chooses means, and often ends
Failure mode	Malfunction (does the wrong thing predictably)	Misjudgement (does a defensible thing that was wrong here)

Dimension	A tool	An autonomous agent
Accountability	Rests with the operator	Genuinely ambiguous — the open question
Control point	The instruction	The authority: what it may decide alone

The rightmost column is the one that breaks under the tool framing. When a tool fails, the failure is a malfunction — it did not do what it was told, and the fix is to correct the instruction. When an agent fails, it has usually done something entirely reasonable that happened to be wrong in this instance: it trusted a stale record, resolved an ambiguity in the less fortunate direction, optimised the objective it was given rather than the one that was intended. There is no malfunction to point to. The agent worked as designed. It simply decided badly, and no line of configuration will reliably prevent the next bad decision, because the space of decisions is not enumerable in advance.

This is why the reflex response — tighter guardrails — provides less protection than it promises. Guardrails constrain the space of action; they do not supply judgement within it. An agent bounded away from catastrophic actions can still make a long series of individually permitted choices that compound into a poor outcome. The tool framing sees only the guardrail and declares the risk managed. The actor framing sees the judgement being exercised inside the guardrail and asks the harder question: was this agent authorised to exercise judgement here at all?

4. The Conditions That Made This Universal

The pattern is not a story of individual carelessness. It is structural, and understanding the structure is what makes it fixable rather than merely regrettable.

- Delivery pressure met a plausible shortcut. Every delivery function was under pressure to move faster with the same headcount. An agent that could close work items autonomously was not a luxury; it was relief. The pressure to grant autonomy was immediate and measurable, while the cost of ungoverned autonomy was deferred and diffuse — exactly the shape of risk that organisations systematically under-price.
- Accountability had nowhere to attach. Established delivery governance assumes a human at each decision. Approval chains, sign-offs, and segregation of duties all presuppose a named person exercising judgement. Insert an agent as the decision-maker and the chain has a link with no one holding it. Rather than redesign the chain, most organisations left the agent's decisions formally attributed to a human who had, in practice, delegated them entirely.
- Observability lagged autonomy. Organisations could see what their agents produced far more easily than they could see what their agents decided. Output was tangible — a resolved ticket, a merged change, a posted entry. The reasoning behind it, the alternatives considered and discarded, the confidence held, was mostly invisible. Governance attached itself to the visible output and left the consequential decision unexamined.

- The analogies misled. The most common framing — "treat the agent like a junior colleague" — was seductive and wrong in an instructive way. A junior colleague learns from correction, carries reputational stake, tires, hesitates, and asks when unsure. An agent does none of these. It will repeat the same misjudgement with perfect consistency and complete confidence, at machine scale, until its authority is changed. The human analogy imported expectations of self-limiting behaviour that the technology does not possess.

Together these conditions meant that even careful organisations, acting rationally at each step, arrived at the same place: consequential decisions delegated to autonomous systems, under governance designed for tools, with the accountability left deliberately vague.

5. What Was Tried, and What It Taught

Organisations did not ignore the discomfort. Several responses became common. Each revealed something.

Human-in-the-loop review. The instinctive first control: require a person to approve the agent's action before it commits. Where the decision was genuinely consequential and infrequent, this worked. Where it was not, it failed in a specific and predictable way. When an agent proposes hundreds of actions a day and the overwhelming majority are correct, the human reviewer does not exercise judgement — they rubber-stamp. The review becomes a ritual that manufactures the appearance of accountability while adding none. Worse, it launders responsibility: the human's click is later cited as the authorising decision, though no real decision occurred. The lesson: human-in-the-loop is a control only where the human's attention is genuinely engaged, and attention does not survive volume.

Confidence thresholds and escalation. More sophisticated: let the agent act alone when its confidence is high and escalate when it is low. Directionally right, and a real improvement. But it rests on the agent's self-assessment of confidence, and the failures that hurt most are precisely those where an agent is confidently wrong — where its internal certainty and its actual correctness have come apart. Thresholds catch the agent's known unknowns and miss its unknown ones. The lesson: escalation logic is necessary but cannot be the whole of the control, because it trusts the agent to know when it should not be trusted.

Sandboxing and blast-radius limits. Constrain what the agent can touch, so that even a bad decision cannot do serious harm. This is sound engineering and every serious deployment should do it. But containment is not governance. It bounds the severity of a failure without addressing who is accountable for the decisions inside the boundary, and it is in tension with value — the more you contain the agent, the less of the autonomy you were paying for you actually get. The lesson: containment is a floor, not a solution; it makes failure survivable without making authority legitimate.

“Every control that worked shared one feature: it governed the agent's authority to decide, not merely its capacity to act.”

The through-line is unmistakable. The controls that helped were the ones that treated the agent as an actor whose authority had to be defined and bounded. The controls that disappointed were the ones that treated it as a tool whose output had to be checked. This is the empirical case for the position this paper argues.

6. The Options, Weighed

Faced with the evidence, an organisation has three coherent strategic postures. They are worth stating plainly, because most organisations have drifted into one without choosing it.

1. Contain the agent as a tool. Restrict autonomy so tightly that the agent never makes a consequential decision alone — every meaningful action routes through a human. This preserves the existing accountability model unchanged. Its cost is that it forfeits most of the value: an agent that cannot decide is an expensive suggestion engine. Defensible for the highest-stakes decisions; ruinous as a general policy.
2. Absorb the agent as workforce. Treat agents as a new class of worker, managed through adapted human structures — objectives, performance monitoring, a manager who owns their output. This rightly recognises the agent as an actor, but it over-borrows from the human model, importing assumptions about learning, judgement, and self-limitation that do not hold. It also tends to obscure accountability behind a management metaphor rather than fixing it.
3. Govern the agent as an accountable actor with delegated decision rights. Treat the agent as neither tool nor colleague, but as what it is: a non-human entity exercising delegated authority. Define explicitly which decisions it holds, which it must escalate, and — crucially — which named human owns the outcome of each. This demands new governance machinery, which is its cost. Its benefit is that it is the only option that matches the actual nature of the thing being governed.

The first option is safe and wasteful. The second is well-intentioned and misconceived. The third is more work, and it is correct.

7. The Recommendation: Govern Authority, Not Output

The defensible position is the third, and it can be made concrete. The organising principle is a shift of the control point — from the agent's output, where the tool framing places it, to the agent's authority, where it belongs.

In practice, that means establishing four things for every agent operating in the delivery chain.

1. An explicit decision-rights boundary. For each agent, a written statement of the decisions it may make alone, the decisions it must escalate, and the decisions it may never make. Not a list of permitted API calls — a list of judgements it is authorised to exercise. This is the single most important artefact, and its absence is the signature of the tool framing.
2. A named human owner for every class of decision. Not a reviewer who rubber-stamps, but an accountable owner who answers for the outcomes of the decisions in their class — whether or not they saw each one. Ownership without per-instance review is uncomfortable, and it is exactly the discomfort that forces the decision-rights boundary to be drawn honestly. If no one is willing to own a class of decision, the agent should not be making it.
3. Decision-level observability. Instrumentation that records not just what the agent did but what it decided and on what basis — the alternatives, the inputs trusted, the confidence held. Output logging tells you an action occurred; decision logging tells you whether the authority was exercised legitimately. The latter is what an owner needs to actually own.
4. A standing authority-review, not an output-review. Periodic reassessment of whether each agent's decision rights remain appropriate as its behaviour, its context, and the stakes evolve. The question is never merely "is the agent performing well?" but "should this agent still hold this authority?" Authority granted once and never revisited is how a reasonable delegation becomes an ungoverned one.

The test of governance is not whether you can review what your agents did. It is whether you can state, for any decision an agent makes, who authorised it to make that decision and who answers if it was wrong. If you cannot, you have deployed an actor and governed a tool.

None of this requires slowing adoption. It requires being explicit about a delegation that is happening whether or not it is acknowledged. The organisations getting the most from autonomous agents are not the ones that granted the most autonomy; they are the ones that granted autonomy deliberately — bounded, owned, observed, and periodically renewed. The autonomy they extend is trusted precisely because it is governed, and it can therefore be extended further.

8. Conclusion

The framing an organisation uses is not a semantic preference; it is the thing that determines which risks it can see. An organisation that calls its agents tools will keep its attention on output quality and keep being surprised by accountability failures, because the framing hides the decision from view. An organisation that calls its agents what they are — actors exercising delegated judgement — will put its attention where the risk actually lives, on the authority being exercised, and will be surprised far less often.

The delivery chain has admitted a new kind of participant: one that decides, at scale, without tiring and without doubt. The task is not to pretend it is an old kind of participant in a faster form. The task is to govern the new one honestly — to say, for every

consequential decision, what the agent may do alone, and who stands behind it when it does. That is not a constraint on the value of autonomous agents. It is the precondition for realising it.