

BEYOND THE RACI MATRIX

Governance for AI Agents — When the Agent Makes the Decision

A Practitioner's Perspective on Accountability Frameworks in the Age of Autonomous Systems

Giovanni Leonardi

February 2024



Governance for AI Agents — When the Agent Makes the Decision

Transformation · No. 302

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

“Every deployment of an agentic system is, in governance terms, a transfer of risk from the explicit and traceable to the implicit and diffuse.”

1. The Accountability Gap Nobody Wants to Name

For as long as organisations have deployed technology, governance has rested on a simple assumption: a human being makes the decision. Technology advises, accelerates, automates the mechanical — but the consequential judgement, the one that carries risk and requires accountability, sits with a person whose name can be written on a chart. That assumption is now under pressure in a way that most governance frameworks are not designed to absorb.

The emergence of AI agents — systems that do not merely recommend but act, that chain reasoning steps together, invoke tools, and execute decisions within parameters that are often loosely defined — represents a qualitative shift in the relationship between technology and organisational accountability. And the structural response, across almost every sector I have observed, has been silence. Not the silence of ignorance, but the more dangerous silence of organisations that sense the problem and choose not to name it, because naming it would require rethinking governance from foundations that most leaders would prefer to leave undisturbed.

2. From Advisory to Agentic: A Governance Inflection Point

The distinction matters because it is not merely technical. An AI system that produces a recommendation and waits for human approval fits comfortably within existing governance structures. The human reviewer remains the accountable party. The model is a tool, like a spreadsheet or a risk calculator — it informs, but it does not decide.

An AI agent is categorically different. When an agent is given a goal, access to data, and the ability to take actions — to send communications, to modify records, to trigger processes, to allocate resources — the governance question shifts from “who approved this recommendation?” to “who is accountable for what the agent did?” And that question, in most organisations, does not have a clear answer.

The pattern I have observed is remarkably consistent. Organisations deploy agents in low-risk, high-volume operational tasks — customer service triage, document classification, routine procurement approvals — under the assumption that the limited scope of the task limits the governance exposure. The agent works. It works well. It handles volume that human teams could not sustain. And so the scope expands, incrementally, without the governance framework expanding with it.

This is not negligence in any conventional sense. It is a structural failure of governance design. The frameworks were built for a world where delegation flows downward through a hierarchy of named individuals. When delegation flows to a system that reasons and acts but cannot be held accountable in any meaningful sense, the framework does not fail dramatically — it simply stops providing the assurance it was designed to deliver, while continuing to look as though it does.

3. Why Existing Frameworks Cannot Simply Be Extended

The instinctive response — and I have watched it play out in boardrooms and governance committees — is to treat the agent as analogous to a junior employee. Set boundaries. Define escalation paths. Monitor outputs. This analogy is comforting and deeply misleading.

A junior employee operates within a social and professional context that constrains behaviour in ways that are largely invisible. They absorb cultural norms, read the room, exercise judgement shaped by consequences they personally bear. They can be asked why they made a decision, and their answer draws on a shared framework of professional reasoning. None of this applies to an AI agent.

The agent operates within the parameters it has been given, but it does not understand those parameters in the way a human does. It cannot distinguish between the letter and the spirit of a policy. It cannot recognise that a technically compliant action is reputationally catastrophic. It processes instructions with a thoroughness and literalness that, paradoxically, creates risk precisely because it is so effective at executing what it has been told to do.

The governance gap is not that AI agents might disobey their instructions — it is that they will follow them with a precision that exposes every ambiguity, every unspoken assumption, and every gap in the policies those instructions encode.

This means that extending existing governance frameworks to cover AI agents is not a matter of adding a new row to the RACI matrix or creating an "AI Oversight" committee. It requires confronting questions that most organisations have not yet formulated clearly:

- Who is accountable when an agent takes an action that was within its defined parameters but produces an outcome that no human anticipated or would have approved?
- How does an organisation maintain an audit trail for decisions that emerge from chains of reasoning that are probabilistic rather than deterministic?

- What does "oversight" mean when the speed and volume of agent decisions make human review of individual actions impossible?
- Where does the boundary lie between the accountability of the person who configured the agent and the accountability of the person who approved its deployment?

4. The Structural Forces That Sustain the Gap

The persistence of this governance gap is not accidental. Several structural forces actively sustain it.

The velocity incentive. Organisations deploy agents because they want speed and scale. Governance is, by its nature, a friction mechanism — it exists to slow decisions down enough that risk can be assessed. There is an inherent tension between the reason agents are deployed and the governance they require, and in most organisations, velocity wins.

The diffusion of expertise. Effective governance of AI agents requires a combination of technical understanding, legal knowledge, risk management capability, and operational domain expertise that rarely exists in a single team. The technical teams understand what the agent can do but not the governance implications. The governance teams understand accountability frameworks but not the technical reality. The result is a series of partial conversations that never converge into a coherent framework.

The precedent vacuum. Unlike financial controls, data protection, or health and safety — domains where decades of incident, regulation, and case law have created well-understood governance patterns — AI agent governance has almost no precedent to draw on. Organisations are reluctant to build frameworks in the absence of regulatory clarity, and regulators are reluctant to prescribe in the absence of sufficient evidence of harm. The result is a waiting game in which the deployment of agents accelerates while the governance response remains static.

The attribution problem. When an agent produces a poor outcome, the causal chain is often genuinely ambiguous. Was the problem in the training data? The prompt design? The parameter boundaries? The approval process? The monitoring regime? The difficulty of clean attribution makes it psychologically easier — and organisationally more convenient — to treat the incident as a one-off rather than as evidence of a systemic governance failure.

5. What Would Genuine Governance Look Like?

The organisations that will navigate this transition well — and they will be a minority — are those that begin from a different starting point. Rather than asking "how do we govern the agent?", they ask "what decisions are we delegating, and what does accountability mean when the delegate is not a person?"

This reframing leads to a fundamentally different governance architecture, one built on several principles that I have seen emerging, in fragments, across the organisations that are furthest ahead:

Decision classification, not task classification. The unit of governance is not the task the agent performs but the decision it makes. An agent that classifies documents is making decisions about categorisation. An agent that triages customer complaints is making decisions about priority and routing. Governance must be designed around the nature, reversibility, and consequence of those decisions, not around the operational process they support.

Accountability by design, not by assignment. In a traditional governance model, accountability is assigned after the fact — a role is named as accountable, and that person accepts the responsibility. With AI agents, accountability must be designed into the system before deployment. This means defining, in advance, who is accountable for each class of decision the agent can make, what information they need to discharge that accountability, and what mechanisms exist for them to intervene.

Continuous assurance, not periodic review. The traditional governance rhythm — quarterly reviews, annual audits, periodic risk assessments — is fundamentally incompatible with agents that make thousands of decisions per hour. Governance must shift to continuous monitoring that can detect drift, anomaly, and emerging patterns in agent behaviour in something closer to real time.

Governance for AI agents is not a technology problem with a governance wrapper — it is a governance problem that happens to involve technology.

Explicit boundary documentation. Every agent deployment should carry documentation — not buried in technical specifications but visible at the governance level — that defines precisely what the agent can and cannot do, what decisions it is empowered to make autonomously, what triggers escalation to a human, and what the failure modes look like. This documentation must be a living artefact, updated as the agent's scope evolves.

6. The Uncomfortable Truth

The deeper challenge, the one that sits beneath all the structural and procedural questions, is philosophical. Governance has always been, at its core, about the allocation of responsibility among people who can bear it. An AI agent cannot bear responsibility. It cannot be sanctioned, retrained through consequence, or held to account in any way that the word "accountability" has traditionally meant.

This does not mean that AI agents should not be deployed — the operational case is often compelling. But it does mean that every deployment of an agentic system is, in governance terms, a transfer of risk from the explicit and traceable to the implicit and diffuse. The organisation remains accountable, but the mechanisms through which it discharges that accountability are weaker, less tested, and less understood than for any previous form of delegation.

The organisations that recognise this — that treat the governance of AI agents not as a compliance exercise but as a fundamental question about how decisions are made and who answers for them — will be better positioned than those that continue to extend

frameworks designed for a world in which every consequential decision had a human name beside it.

That world is already changing. The governance conversation has not yet caught up.