

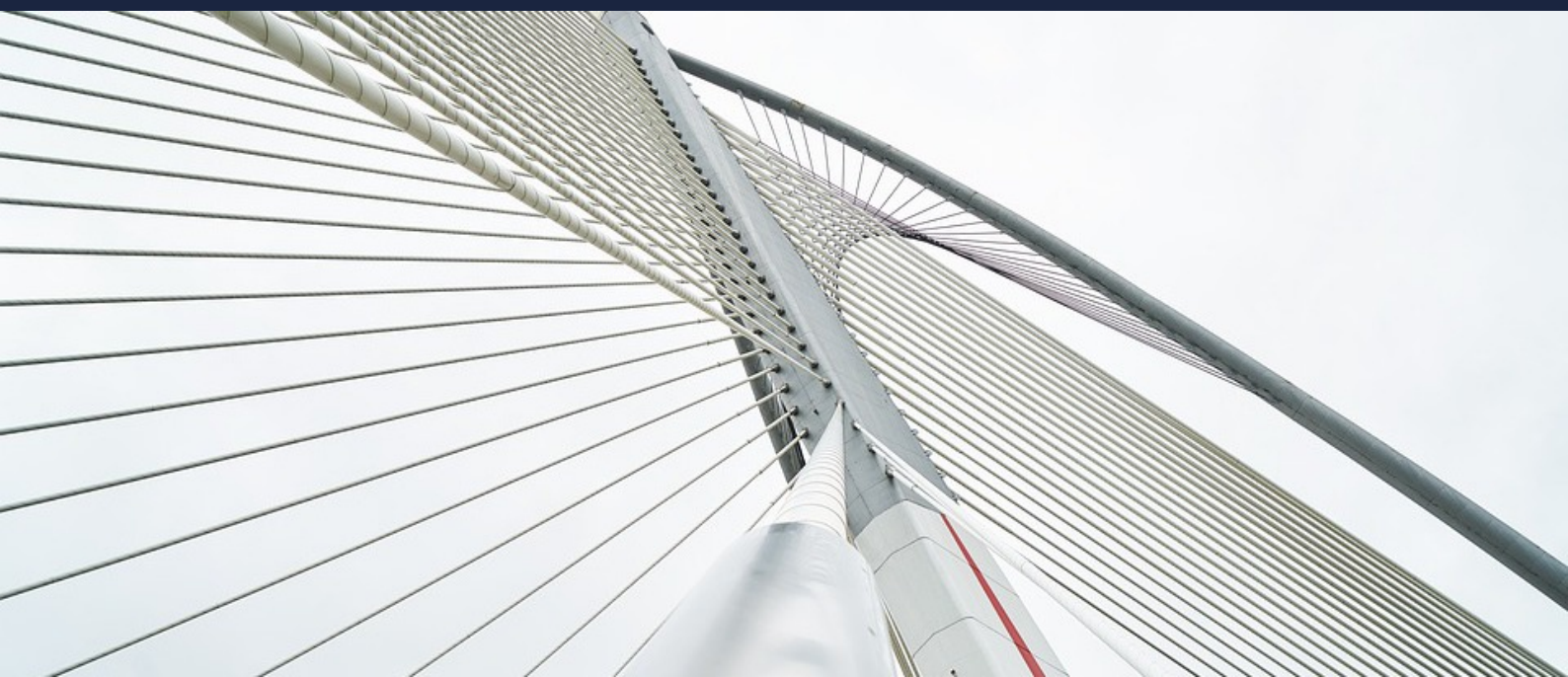
BEYOND THE GUARDRAIL

When the Assistant Becomes an Agent

A Practitioner's Perspective on Why the Transition from AI-as-Tool to AI-as-Agent Demands a Leadership Response Most Organisations Have Not Even Begun

Giovanni Leonardi

October 2024



When the Assistant Becomes an Agent

Leadership · No. 92

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

“We built governance for tools and assumed it would stretch to cover agents — but tools do not have goals, and agents do, and that difference changes everything.”

1. The Distinction That Changes Everything

There is a transition underway in how organisations use artificial intelligence that is more consequential than most leadership teams have recognised. It is the shift from AI as a co-pilot — a tool that assists, suggests, and drafts at a human's direction — to AI as a colleague: a system that observes conditions, forms intentions, and takes actions with varying degrees of independence.

This is not a difference of degree. It is a difference of kind. A co-pilot responds when prompted. A colleague acts when circumstances warrant. A co-pilot has no goals of its own. A colleague — or at least, an agentic system designed to behave like one — pursues objectives, makes trade-offs, and takes initiative within defined parameters. The organisational implications of this shift are profound, and from what I have observed, almost nobody on the delivery side of enterprise change is planning for them.

The pattern I see across organisations in late 2024 is remarkably consistent. They adopted generative AI as an assistant — drafting documents, summarising meetings, answering questions from knowledge bases. They built some guardrails: acceptable use policies, data handling rules, procurement frameworks for AI tools. And they assumed, implicitly or explicitly, that these guardrails would scale to cover whatever came next.

What came next is agents. Systems that do not wait to be asked. Systems that monitor, decide, and act. And the guardrails built for tools are not merely insufficient for agents — they are the wrong category of response entirely.

2. Why the Delivery Side Feels This First

Leadership teams tend to encounter AI as a strategic question: should we invest, where should we invest, what is the competitive landscape. This is important but abstract. The delivery side — programme managers, operational leaders, team leads, the people responsible for actually making things work — encounters AI as an operational reality. And it is on the delivery side that the co-pilot-to-colleague transition creates its most

immediate and least understood challenges.

In my experience, three challenges dominate.

The accountability gap. When a co-pilot helps a human draft a recommendation, the human is accountable for that recommendation. The accountability is clear because the human made the decision. When an agent monitors a portfolio of projects, identifies that a workstream is at risk, and automatically reallocates resources to mitigate that risk, who is accountable for the reallocation? The agent? The person who configured the agent? The person who approved its deployment? The delivery lead whose resources were moved without their direct involvement? Most organisations cannot answer this question today, and the inability to answer it is not an abstract governance concern — it is a practical delivery problem that surfaces the first time an autonomous action goes wrong.

The oversight design problem. Co-pilots are overseen by their users in real time. Every output is reviewed before it is acted upon, because the human is the actor. Agents, by design, act between human reviews. The question of how to oversee them is not a question of adding more review meetings or dashboards. It is a question of designing an entirely new kind of oversight — one that provides confidence without reintroducing the bottleneck that autonomy was meant to remove. This is an operating model problem, not a technology problem, and it is one that most organisations have delegated to their technology teams rather than their operating model designers.

The most dangerous assumption in enterprise AI today is that governance designed for tools will scale to govern agents. Tools do not pursue goals. Agents do. That difference demands a fundamentally different governance architecture.

The trust calibration challenge. Delivery teams are being asked to trust autonomous systems to make decisions that affect their work, their teams, and their outcomes. Trust in human colleagues is built through observation, shared experience, and demonstrated reliability over time. Trust in autonomous agents has no established mechanism. Organisations are discovering that asking a delivery lead to trust an agent's judgement is categorically different from asking them to trust a human colleague's judgement — and that the absence of a trust-building mechanism leads either to reflexive rejection or, worse, to uncritical acceptance.

3. What the Operating Model Needs to Become

The operating model adjustments required for the co-pilot-to-colleague transition are not incremental. They require leaders — particularly delivery leaders — to rethink several assumptions that have been stable for decades.

First, the operating model must accommodate non-human decision-makers as first-class participants. This sounds self-evident, but its implications are far-reaching. Decision-rights frameworks, RACI matrices, escalation paths, governance structures — all of these assume that every decision-maker is a person. Extending them to include agents requires not just adding a row to the RACI chart but redesigning the logic that underpins it. An agent does not attend a steering committee. It does not read the room. It does not

adjust its recommendations based on the sponsor's mood. The governance mechanisms that work for human decision-makers — mechanisms that rely on social awareness, contextual judgement, and political sensitivity — do not transfer.

Second, the operating model must make the boundary between human and agent authority explicit, reviewable, and adjustable. The co-pilot model had an implicit boundary: the human decided everything, and the AI assisted. In the colleague model, the boundary is somewhere in the middle, and its location is a design decision that has to be made, documented, reviewed, and adjusted as experience accumulates. Organisations that leave this boundary implicit — that assume it will work itself out — will find themselves managing a series of escalating incidents as agents act in ways that humans did not expect or authorise.

Third, the operating model must create feedback loops that allow the organisation to learn from agent behaviour at the same speed that agents learn from data. This is perhaps the most subtle requirement and the one I see most consistently neglected. Agents that learn and adapt are valuable precisely because they improve over time. But the organisation's understanding of what the agents are doing, and why, must keep pace. If the agents evolve faster than the organisation's oversight mechanisms, the result is a growing gap between what the agents are actually doing and what the organisation believes they are doing. That gap is where the serious failures live.

4. The Leadership Response That Is Missing

What strikes me most about the current moment is not that organisations are struggling with this transition — the transition is genuinely hard, and it is still early. What strikes me is that most leadership teams have not recognised that there is a transition to struggle with. The conversation in most boardrooms is still about AI adoption: which tools to buy, which processes to augment, how to capture efficiency gains. The conversation that is missing is about AI integration: how to redesign the operating model for a world in which some of the decisions are made by systems that have agency.

“We built governance for tools and assumed it would stretch to cover agents — but tools do not have goals, and agents do, and that difference changes everything.”

This is not a technology leadership problem. It is a delivery leadership problem, an operational leadership problem, and ultimately an organisational design problem. The leaders best placed to solve it are not the chief technology officers and AI directors — though their input is essential — but the programme directors, operations directors, and transformation leaders who understand how work actually gets done and how decisions actually get made in complex organisations.

The co-pilot era was comfortable because it preserved the fundamental operating assumption: humans decide, machines assist. The colleague era dismantles that assumption, and the organisations that acknowledge this early — that start designing the operating model, governance, and leadership practices for a world of agentic systems — will have a substantial advantage over those that wait for the first serious failure to force the conversation.

The shift from co-pilot to colleague is not coming. In many organisations, it has already arrived. The question is whether leadership will design for it or react to it. In my experience, the latter is both more common and more expensive.