

BEYOND THE TOOLCHAIN

The Operating Model Is the Product: Why Human-AI Teaming Fails When It Is Treated as a Procurement Decision

A Practitioner's Perspective on Redesigning the Delivery Operating Model for Human-Agent Teams

Giovanni Leonardi

October 2025



Contents

1. Executive Summary	3
2. The Question Organisations Are Actually Asking	3
3. Why the Tooling Frame Fails	4
4. What Organisations Tried — and What the Evidence Shows	5
5. Weighing the Obvious Objection	5
6. The Four Conditions of a Working Human-AI Operating Model	6
7. The Recommendation: Design the Operating Model First	6
8. A Note on Sequencing and Patience	7

The Operating Model Is the Product: Why Human-AI Teaming Fails When It Is Treated as a Procurement Decision

Transformation · No. 425
October 2025

“Human-AI teaming is not a capability you add to an operating model; it is an operating model, and it displaces the one you have.”

1. Executive Summary

Across 2025 the phrase human-AI teaming moved from conference panels into board papers. Yet most organisations are still treating it as a procurement question — which agents to license, which platform to standardise on, how many seats to buy — when the evidence of the past eighteen months points somewhere quite different. The organisations that got durable value from autonomous and semi-autonomous agents did not buy their way there. They redesigned how work is decomposed, allocated, supervised, and accounted for. The tools followed.

The central claim of this paper is deliberately blunt. Human-AI teaming is not a capability you add to an operating model; it is an operating model, and it displaces the one you have. Treating it as a tooling decision is not merely incomplete — it is the specific mistake that connects nearly every stalled initiative I have examined.

This paper sets out the organisational conditions that produced the current confusion, examines what organisations actually tried through 2024 and 2025, separates what worked from what failed, weighs the obvious objection, and argues for a specific response: make the operating model the primary design object and treat tooling as a downstream choice that serves it. It closes with four conditions that distinguish the teams getting compounding value from those cycling through pilots, and a sequencing recommendation for leaders who must decide now rather than wait for the practice to settle.

Nearly every failed AI-teaming initiative I have examined failed for the same reason: the technology was treated as the variable and the operating model as the constant. It is the other way round.

2. The Question Organisations Are Actually Asking

When a leadership team asks how do we adopt AI agents in delivery?, the question sounds technological. It is not. Strip away the vocabulary and what remains is older and harder: who does what, who decides, who is accountable, and how do we know the work is good? Agents force that question to the surface because they insert a new kind of participant into the workflow — one that produces finished-looking work product, makes intermediate decisions, and does so at a speed and volume the surrounding controls were never designed to absorb.

The pattern that recurs is that organisations answer the technological question briskly and leave the operating-model question unasked. They stand up a platform, grant access, run a pilot, and measure adoption. Some months later the pilot has produced impressive demonstrations and no change to how the delivery organisation actually runs. The agents did the work they were pointed at; the operating model quietly routed around them.

This is not a failure of ambition or of technology. It is a category error. The organisation set out to answer a question about structure using an instrument designed to answer a question about tools, and the instrument returned a confident, precise, irrelevant answer.

3. Why the Tooling Frame Fails

Three structural forces sustain the tooling frame, and each is doing quiet damage. Each has to be named, because each is invisible from inside the frame itself.

- The budget lives in technology. Because agents are procured, the money and the mandate sit with the function that buys software. That function is not chartered to redesign how delivery teams allocate work, supervise output, or account for decisions — so it does not, and no one else is asked to.
- The unit of adoption is the individual. Access is granted person by person, so value is measured person by person: minutes saved on a task. This misses the point entirely. The gains from teaming are structural, not individual, and structural gains never show up in a per-seat productivity metric.
- The controls are invisible until they break. Review gates, sign-off chains, and quality assurance were built around human throughput. Agents do not violate these controls; they overwhelm them. A reviewer who could check the output of four people cannot check the output of four people each running five agents — and no one notices until the backlog of unreviewed work becomes the constraint.

*“Agents do not break your controls. They flood them,
which looks the same right up until it doesn't.”*

The tooling frame cannot see any of this, because all three forces are properties of the operating model, and the tooling frame does not look at the operating model. It is a well-lit search under the wrong lamp-post.

4. What Organisations Tried — and What the Evidence Shows

By late 2025 there is enough of a track record to be empirical rather than speculative. Four broad approaches were tried across the organisations I have observed. It is worth being precise about what each one assumed and what it actually produced.

Approach	What it assumed	What actually happened
Seat rollout	Value comes from individual productivity	High adoption, negligible operating impact; gains real but invisible at team level
Centre of excellence	A central team can define the right way and disseminate it	Good standards, little uptake; the centre held authority over tools, not over delivery
Task-by-task automation	Replace discrete tasks with agents one at a time	Local speed-ups, global congestion at the review and integration points
Team redesign	Rebuild how a delivery team allocates and supervises work	Slow to start, but the only approach that moved team-level outcomes

The first three share a lineage: they hold the operating model fixed and insert agents into it. The fourth changes the operating model and lets the agents follow. Only the fourth produced durable, team-level results — and it was consistently the least popular, because it is the hardest, the slowest to show a number, and the one that no single function owns.

The failure mode of the first three approaches is more insidious than doing nothing. They produce local evidence of success — a faster task, a delighted pilot user, a compelling demo — that masks the absence of system change. Leadership, reasonably, scales the thing that showed a number. The thing that showed a number is the seat rollout. So the seat rollout is scaled, and the organisation is surprised when scaling it changes nothing at the level that matters. You cannot scale your way out of a structural problem with a per-seat solution. The arithmetic of the pilot was never the arithmetic of the system.

The centre-of-excellence approach fails more poignantly, because it is often staffed by the people who understand the problem best. They write genuinely good standards. But a centre chartered to advise on tooling has no lever over how a delivery team decomposes its work or resources its supervision. It can publish the right answer and watch it be ignored, because publishing is not the same as changing who is accountable for what.

5. Weighing the Obvious Objection

There is a serious counter-argument, and a white paper that ducks it is not worth reading. The objection runs: why redesign anything? The tools are improving so quickly that next year's agents will be reliable enough to slot into the existing model without all this organisational upheaval. Wait, and the problem solves itself.

This is wrong, but it is wrong in an instructive way. Better agents make the individual output more reliable; they do nothing about the three structural forces above. A more capable agent still has its budget in technology, is still adopted per-seat, and still floods

controls built for human throughput — in fact a more capable agent floods them faster. Capability improvements make the operating-model gap wider, not narrower, because they raise the volume of work arriving at controls that were not re-rated to receive it.

Waiting for better tools to fix an operating-model problem is like waiting for a faster engine to fix a road that has no lanes. The better engine arrives, and the congestion is worse.

The objection contains one true thing worth keeping: the specific tools genuinely will change, and change fast. That is an argument for anchoring on the operating model, not against it — because the operating model is the part that is yours to design and the part that keeps paying off regardless of which agent you are running next year.

6. The Four Conditions of a Working Human-AI Operating Model

The teams that made this work were not, in my observation, using better agents than the teams that failed. They had, knowingly or not, satisfied four conditions. These four conditions are the actual design object — the thing a leader should be building.

1. Work is decomposed for teaming, not for humans. Tasks are broken down so the boundary between what an agent drafts and what a human judges is explicit and deliberate — decided in the design of the work, not left to whoever happens to hold the ticket.
2. Supervision is a designed role, not a residual one. Someone holds first-order accountability for the quality and coherence of agent output, with the protected time and the authority to exercise it — not a reviewer squeezing it in around a full day job.
3. The controls are re-rated for the new throughput. Gates, reviews, and assurance are rebuilt around the volume agents actually produce, moving from inspect-everything to sampling and exception-based review, with escalation triggers set in advance.
4. Accountability is unambiguous at the point of decision. When an agent makes an intermediate decision, it is clear beforehand whose decision it is, organisationally and legally. Ambiguity here is not a governance nicety; it is the precise thing that stops teams from letting agents do anything that carries consequence.

Notice what these have in common. Not one is a property of the agent. Every one is a property of how the team is organised, resourced, and held to account. That observation is the whole argument in miniature: the design surface that matters is the operating model, and it has been sitting unattended while everyone debated platforms.

7. The Recommendation: Design the Operating Model First

Given the evidence, the defensible position is simple to state and demanding to execute: make the operating model the primary design object, and treat tooling as a downstream choice that serves it. For a leader deciding now, that resolves into a sequence — and the

sequence is the recommendation, because every failed programme I examined inverted it.

1. Pick one real delivery team, not a lab. The unit of change is a team that ships something the organisation depends on. Sandboxes teach nothing about your controls, because sandboxes have none.
2. Redesign the work allocation before granting access. Decide on paper where agents draft and humans judge before anyone logs in. Access-first rollouts calcify the old allocation and then you are fighting habit as well as structure.
3. Name and resource the supervisor role. Give one person first-order accountability for agent-output quality, with protected time. This is the highest-leverage single move and the one most often skipped, because it costs a headcount decision rather than a licence.
4. Re-rate the controls deliberately. Move from inspect-everything to sample-and-except; set sampling rates and escalation triggers in advance, and instrument them so you can see when volume outpaces review.
5. Only then standardise tooling. Choose platforms to fit the operating model you have designed – not the reverse. The tool is the last decision, not the first.

The order is not incidental. It is the finding. Tool-first, model-never is the signature of the initiatives that produced demos and no change.

8. A Note on Sequencing and Patience

There is a reason the winning approach is the unpopular one: it shows its number last. A seat rollout produces a productivity anecdote in a fortnight. An operating-model redesign produces nothing visible for a quarter and then produces the only result that matters. Leadership teams under pressure to demonstrate momentum reliably choose the fast anecdote, and reliably regret it two budget cycles later, when the anecdotes have not compounded into anything.

The organisations that will look prescient a year from now are the ones willing, today, to be slow in the right way – to treat human-AI teaming as the operating-model question it actually is, and to resist the considerable gravitational pull of the procurement question it is disguised as. The tools will keep improving; that is exactly why they are the wrong anchor. Anchor on the operating model, because the operating model is the part that is yours to design, and the part that, redesigned well, keeps paying off no matter which agent you are running next.