

BEYOND COPILOT ADOPTION

Human-AI Teaming Fails When AI Is Added to the Old Operating Model

What 2025's evidence says about redesigning decision loops, roles and exception capacity

Giovanni Leonardi

October 2025



Human-AI Teaming Fails When AI Is Added to the Old Operating Model

Transformation · No. 358
October 2025

"The unit of redesign is not the task; it is the decision loop."

1. The queue cleared, but the work did not

At nine o'clock on Monday, a service team opened a queue of 43 complex cases. An AI assistant had already classified 39, drafted responses for 31 and highlighted six as potential policy exceptions. The dashboard suggested that most of the morning's work had been completed before anyone arrived.

By noon, the team was behind.

Two experienced specialists had rechecked every classification because nobody knew which signals the assistant had weighted. A manager rewrote the six escalations because the summaries omitted the commercial consequence. Three draft responses were technically accurate but contradicted commitments recorded elsewhere. The assistant had accelerated several tasks, but the operating model still assumed that people would create, verify, coordinate and own the work exactly as before.

This composite pattern has become increasingly familiar during 2025. Organisations have acquired capable generative AI assistants, connected some of them to workflow tools and trained staff to prompt them. Productivity appears in isolated measures: a faster first draft, quicker retrieval, more cases screened. Yet end-to-end performance often improves far less than expected. Work is completed twice, exceptions multiply, managers become verification bottlenecks and accountability grows less clear.

The problem is not that humans and AI cannot work together. It is that most organisations are treating teaming as a feature of software rather than as a design of work.

The unit of redesign is not the task; it is the decision loop.

2. What the emerging evidence actually shows

By autumn 2025, the evidence available to practitioners is uneven but directionally consistent. Controlled experiments, internal pilots and operational deployments measure different things, so headline productivity claims should be treated cautiously. A short

drafting exercise is not the same as a regulated process; a novice completing a bounded task is not the same as an expert carrying accountability across a chain of decisions.

Even so, four patterns recur.

2.1 AI creates the most value where work can be inspected

Human-AI teams perform well when an AI contribution can be reviewed quickly against a clear standard. Summarising a known document set, generating alternative formulations, classifying routine items and checking for omissions all fit this pattern. The machine reduces the cost of producing or scanning possibilities; the person applies context and judgement.

The gain falls when verification takes as long as creation. A fluent answer may require an expert to reconstruct sources, assumptions and dependencies before it can be trusted. Apparent speed at the front of the process then becomes hidden review work downstream.

2.2 Capability gains are distributed unevenly

Less experienced staff can gain substantially when the system supplies structure, examples or a plausible starting point. This can narrow some performance gaps. It can also create a new risk: the people who benefit most may be least able to recognise a confident error.

Experienced practitioners often gain differently. They use AI to expand the option set, challenge a first view or remove routine preparation. Their advantage comes not from accepting more output, but from knowing what to reject and where to probe.

The operating implication is important. One training course and one control threshold cannot serve every level of expertise.

2.3 Task acceleration does not guarantee process acceleration

A process is a network of dependencies. If AI reduces drafting time from 90 minutes to 15 but leaves a four-day approval queue unchanged, the customer experiences almost no improvement. If it produces three times as many analyses without increasing decision capacity, it creates inventory rather than value.

Local productivity measures conceal this. Organisations count prompts, licences, documents and minutes saved while ignoring rework, hand-offs, decision latency and exception volume. The relevant evidence sits at the level of flow and outcome.

2.4 Trust is calibrated through feedback, not declared by policy

People do not become appropriately trusting because a policy tells them to “use human judgement”. They learn calibration when they see where the system succeeds, where it fails and how their interventions change the outcome.

Too little trust leads to duplicate work. Too much trust leads to automation bias. Both are symptoms of an operating model that gives users neither visible evidence nor a

disciplined feedback loop.

Human oversight is effective only when the human has the time, evidence, authority and competence to disagree.

3. The operating models organisations have tried

The first wave of adoption has produced four common models. Each solves part of the problem; none is a complete answer.

3.1 The personal copilot

Individuals choose when and how to use an assistant. This encourages experimentation and can improve personal productivity quickly. It works best for reversible, low-consequence work where the user already owns the output.

It fails as an enterprise operating model because methods, controls and learning remain private. One employee verifies every claim; another accepts the first response. Useful prompts and known failure modes disappear inside personal practice. The organisation pays for capability but cannot reliably govern or improve it.

3.2 The AI first draft

The machine produces and a person approves. This is easy to explain and often suitable for text-heavy work. It fails when “approval” becomes a ritual. High volumes turn reviewers into passive signatories, while the original reasoning remains opaque.

The model also preserves old hand-offs. The person is still positioned at the end of the line rather than shaping objectives, constraints and escalation criteria at the beginning.

3.3 Human in every loop

Every material step requires explicit human permission. This provides reassurance during a pilot and is appropriate for novel or high-impact decisions. At scale, it creates queues and diluted attention. Reviewers begin to approve routine cases quickly, leaving less capacity for the exceptions that genuinely require judgement.

Human involvement is not the same as human control. Control depends on where the person enters the loop and what decision rights they hold.

3.4 Autonomous execution with exception handling

The system completes bounded work and sends only exceptions to people. This offers the strongest potential for end-to-end improvement. It also exposes weak process definitions. If thresholds are vague, data is fragmented or escalation capacity is undersized, the agent either stops too often or acts too broadly.

This model works only when autonomy is earned through clear boundaries, observable performance and reliable intervention.

Design question	Conventional deployment	Human-AI teaming model
Primary unit	Individual task	End-to-end decision loop
Human role	User or approver	Objective setter, judge and exception owner
AI role	General assistant	Bounded contributor with defined authority
Performance measure	Output speed and usage	Flow, quality, outcome and recovery
Control point	Final review	Constraints, evidence, thresholds and escalation
Learning	Informal user practice	Shared operational feedback

4. The strongest objection: let the technology mature first

A credible objection is that operating-model redesign is premature. Models are improving rapidly, tools are changing and many use cases remain experimental. Redesigning roles and controls around today's limitations could lock an organisation into arrangements that become obsolete. A lighter approach—provide secure tools, let teams experiment and formalise only after stable patterns emerge—preserves flexibility.

This view is right about the danger of over-engineering. Organisations should not build a permanent bureaucracy around every pilot. Nor should they pretend to know the final division of labour between people and machines.

But waiting is not neutral. Informal practices are already becoming the operating model. Verification work is being absorbed without measurement, accountability is drifting to whoever presses the final button and local automations are creating dependencies. The longer these arrangements persist, the harder they are to see and redesign.

The answer is not a rigid target organisation. It is a minimum viable operating model: a small set of explicit decisions that can evolve as capability changes.

5. A more effective design: the decision-loop model

A human-AI operating model should be designed around complete decision loops. A loop begins with an objective, turns evidence into a judgement, takes an action, observes the result and learns. Tasks matter, but only as contributions to that loop.

Six design choices make the model operational.

5.1 Start with the outcome and the decision

The work should be described in business terms before tools are assigned. What outcome is required? Which decision changes that outcome? What constitutes a good result, and over what period?

This prevents a common error: automating the most visible activity rather than the constraint. Faster report production has little value if the decision forum remains monthly and overloaded.

5.2 Divide work by comparative strength

AI is useful for breadth, speed, pattern detection, retrieval and the production of alternatives. People remain essential where work requires legitimate authority, moral or commercial judgement, negotiation, empathy, contextual interpretation and responsibility for consequences.

This is not a permanent boundary. It is a design hypothesis to be tested. The division should be made explicit enough that the team can observe when capability, risk or workload changes it.

5.3 Put human judgement before, within and after the loop

Human involvement should take three different forms:

- Before: set the objective, constraints, success measures and authority envelope.
- Within: decide genuine exceptions, resolve conflicting evidence and intervene when uncertainty exceeds a threshold.
- After: sample outcomes, investigate failures, adjust boundaries and improve the process.

A single approval at the end performs none of these roles well.

5.4 Design evidence into the hand-off

When work moves from AI to a person, the recipient needs more than an answer. The hand-off should include the source material used, assumptions made, uncertainty or conflict detected, actions already taken and the precise decision required.

In the opening example, the six policy exceptions became manageable only when each escalation stated the relevant rule, the customer commitment at risk, the evidence in conflict and the latest safe decision time. Review time fell because the specialist no longer had to reconstruct the case.

5.5 Size the exception system

Exception handling is capacity, not a footnote. If a team processes 1,000 cases a week and 12 per cent require human judgement, the model creates 120 specialist decisions. At 15 minutes each, that is 30 hours of skilled work before meetings, investigation or rework.

A design that funds the AI service but not the exception capacity will either slow down or lower its standards. Thresholds, staffing and service levels must be designed together.

5.6 Build a shared learning cadence

Teams need a regular forum to examine outcomes rather than admire outputs. A useful review looks at false positives, missed exceptions, reversals, rework, user overrides, customer effects and emerging failure patterns.

The purpose is to decide something: adjust an instruction, change a threshold, improve source data, redesign a hand-off, narrow authority or remove a use case. Learning that does not change the system is only reporting.

6. How to move from pilot to operating model

A practical transition can be made in five moves.

1. Map one complete decision loop. Select a use case with meaningful volume and a measurable outcome. Trace objective, evidence, judgement, action, feedback and accountability.
2. Measure the baseline flow. Record elapsed time, active work, hand-offs, rework, exception rates and outcome quality before introducing the new design.
3. Specify roles and boundaries. Name the business owner, the human exception role, the AI contribution, prohibited actions and the conditions for escalation or suspension.
4. Run a bounded operating trial. Compare the whole loop with the baseline. Segment results by task type, risk and user expertise rather than relying on an average productivity figure.
5. Scale only what the evidence supports. Expand authority where outcomes are stable and detectable; retain recommendation or review where consequence is high or verification remains difficult.

Leadership has a distinct responsibility throughout. It must resist two convenient fictions: that buying access constitutes transformation, and that adding an approval step constitutes control. The operating model should make it clear who owns the result, how work is divided, what evidence travels with a decision and how the system learns.

7. The recommendation

Organisations should stop treating human-AI teaming as an adoption programme and establish it as an operating-model discipline. The immediate priority is not a new organisational chart. It is the redesign of selected decision loops, supported by explicit roles, measured outcomes, usable evidence and funded exception capacity.

This approach is deliberately modest. It does not assume that current model limitations are permanent, nor that autonomous work should be held back until every uncertainty disappears. It creates a structure that can absorb improvement without surrendering accountability.

The evidence from 2025 does not support the claim that AI simply replaces tasks and leaves the surrounding organisation intact. It points to a more demanding conclusion: value appears when the work system changes with the tool.

Human-AI teaming succeeds when neither side is asked to imitate the other. Machines should not be dressed as accountable colleagues, and people should not be reduced to slow approval gates. The better model gives each a distinct contribution inside a decision loop that the organisation can see, govern and improve.