

BEYOND TIME SAVED

Why Successful AI Pilots Deliver No Value — and What to Measure Instead

What a year of generative-AI trials taught us about the gap between a tool that works and a transformation that pays

Giovanni Leonardi

April 2024



Why Successful AI Pilots Deliver No Value — and What to Measure Instead

Transformation · No. 312

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

“Time does not convert pilots into returns; redesign does.”

1. The Pilot Passed. Where Did the Value Go?

The dashboard is green. Twelve weeks into the trial, adoption sits at seventy-one per cent of the invited cohort, the satisfaction score is the highest the programme has ever recorded, and the headline figure — an average of thirty-four minutes saved per user per day — has already been rounded up and pasted into three steering-committee decks. The vendor is delighted. The sponsor is delighted. Then, in the quiet part of the review, the finance business partner asks the only question that has ever really mattered: where is the money? The room goes still, because everyone in it senses that the honest answer is that no one can find it.

I have sat through a version of that meeting more often than I would like to admit over the past eighteen months. The technology varies — a drafting assistant for the bid team, a summarisation copilot for the service desk, a code companion for the engineering group — but the shape of the disappointment is identical. The pilot succeeds by every number on the slide, and the enterprise feels no different. Costs do not fall. Cycle times do not move. The annual plan is submitted with the same headcount it always carried. Something that was supposed to be transformational has been, at the level of the profit-and-loss account, completely invisible.

The reflexive explanation is that it is early, that these things take time, that value lags adoption. There is a grain of truth in it, and I will come back to that grain because it is the strongest thing the optimists have to say. But it is mostly an evasion. The deeper problem is not patience. It is that we are measuring the wrong thing, at the wrong altitude, and then being surprised when the number we are watching turns out to have almost nothing to do with the number we care about.

2. Three numbers that measure the tool, not the transformation

Look at what a typical pilot actually reports, and a pattern emerges. Almost every scorecard reduces to three figures: how many people used it, how much they liked it, and how much time each of them saved on the task in front of them. Each is easy to collect, each moves in a satisfying direction, and each is close to worthless as a measure of value realised.

Adoption tells us the tool was switched on, not that any work was redesigned around it. It is perfectly possible — common, in fact — for a workforce to enthusiastically adopt a capability and change nothing about how the work flows through the organisation. Satisfaction is worse still: it measures whether people enjoy the experience, and people reliably enjoy having a clever assistant take the tedious first draft off their hands. Neither of those sensations has a natural exchange rate into cash.

It is the third number, time saved, that does the real damage, because it looks the most like value and behaves the least like it. When we tell a user to log the minutes an assistant saved them on a document, we are asking them to estimate a counterfactual — how long would this have taken — while they are enthusiastic about a new toy and aware that the programme's continuation depends on the answer. The measurement is contaminated at the source. But even if the thirty-four minutes were real to the second, the arithmetic that follows is the actual fallacy: minutes saved, multiplied by loaded hourly cost, multiplied by number of users, annualised, and presented as a benefit. That calculation quietly assumes that a saved minute is a banked pound. It almost never is.

The central error of AI pilot measurement is the belief that time saved on a task and value delivered to the organisation are the same quantity in different units. They are not even the same kind of thing.

3. Saved time is not saved money — it has to go somewhere

Here is the mechanism the time-saved calculation ignores. When a knowledge worker saves twenty minutes on a report, those twenty minutes do not leave the building as cash. They stay inside the person's day, and they refill. Sometimes they refill with genuinely valuable work that was previously being crowded out — that is the good case, and it is real, but it is diffuse and almost never captured on any scorecard. More often they refill with more of the same activity, some of which no one needed: three more drafts of a document that will be read once, a longer analysis that changes no decision, a more polished version of an artefact whose polish was never the constraint. The freed capacity is consumed on the spot, and the profit-and-loss account never hears about it.

Consider a composite that will be recognisable to anyone who has run one of these. A service organisation pilots a summarisation and drafting assistant across a team of roughly two hundred case handlers. The trial is a genuine success on its own terms: average handling time on the targeted case type falls by eleven per cent, and the handlers, freed from the most tedious part of their write-ups, report the highest morale the function has seen in years. The programme confidently projects the eleven per cent as a headcount-equivalent saving — around twenty-two roles' worth of capacity — and books it as the business case.

Twelve months on, the headcount is unchanged, and so is the cost base. The eleven per cent was real, but it arrived as eleven per cent of each person's time, spread invisibly across two hundred people, none of whom was ever going to be the specific individual who left. Capacity released as a thin film across a large population is capacity that cannot be harvested. To convert it into money, someone would have had to consolidate the freed

fractions — redesign the queue, change the team's shape, lower the headcount plan, or deliberately load the released hours with work that was genuinely being starved. Nobody did, because the scorecard said the value had already been delivered. The measurement did not merely fail to capture the value; it actively signalled that the job was done, and so ensured the value was never captured at all.

“A pilot that measures the tool tells you the technology works. Only a measure taken at the level of the redesigned process tells you whether the enterprise will ever see the money.”

4. The strongest objection, taken seriously

The most serious counter-argument is not that the sceptics are impatient but that they are looking too early at the wrong horizon. On this view, the leading indicators — adoption, fluency, satisfaction — are exactly what you should measure in the first year, because they are the necessary precursors of value that will only materialise once the technology is embedded, once workflows mature around it, once a general-purpose capability compounds the way electricity or the spreadsheet eventually did. Demanding profit-and-loss impact from a twelve-week pilot, the argument goes, is like condemning a language because a beginner cannot yet write poetry in it.

This deserves a real answer, not a dismissal, because it is partly right. General-purpose capabilities genuinely do take time to reorganise the work around them, and some of the most important effects will be slow, compounding, and hard to attribute. But the argument smuggles in a false comfort. It treats the absence of measured value as evidence that value is quietly accumulating out of sight, when the far more likely explanation is that nothing is accumulating at all, because no one has changed the structure through which value would have to flow. Time does not convert pilots into returns; redesign does. The leading indicators are worth watching — but only if they are leading somewhere, and adoption plus satisfaction, on their own, lead nowhere in particular. The honest version of the patience argument would say: give it time, and in the meantime measure whether we are actually rebuilding the process. Almost no one measures the second half of that sentence.

5. Measure the work, not the widget

If the tool-level numbers are a mirage, what should a serious organisation watch instead? The shift is from measuring the technology to measuring the work the technology was supposed to change — and it means accepting harder, slower, less flattering numbers.

- Measure the process end to end, not the task in isolation. The relevant unit is not "minutes saved on drafting a case note" but "elapsed time and cost from case opened to case closed." Local task speed can improve while the end-to-end flow does not move at all, because the constraint was somewhere else entirely. If the whole process has not moved, no amount of task-level acceleration has created value; it has only relocated slack.
- Track freed capacity to a destination. Every claim of time saved should be required to answer the question and then what? Capacity that is not consolidated, removed, or deliberately redirected has not been realised. A benefit with no named destination is not a benefit; it is a rounding error distributed across people's calendars.
- Count decisions and outcomes changed, not artefacts produced. For work whose product is judgement rather than volume, the honest measures are things like decisions made faster or better, errors avoided, rework eliminated, revenue actions taken sooner. These are harder to instrument than a time-saved slider, which is precisely why they are more trustworthy: they cannot be improved by simply generating more output.
- Set the baseline before the tool arrives, and hold it still. Most pilots have no clean before-picture, which is what allows the after-picture to be whatever the programme needs it to be. A measured baseline of end-to-end cost and cycle time, taken before the assistant is switched on, is worth more than any quantity of in-flight enthusiasm.

None of this is exotic. It is the ordinary discipline of benefits realisation, applied to a technology exciting enough that people feel entitled to skip it. The uncomfortable truth is that the measurement problem in front of us is not new and not specific to this technology at all.

6. We have measured a wave badly before

We have run this exact experiment more than once. The organisations that spent heavily on large enterprise-software programmes and measured them by modules deployed and users trained, rather than by processes actually simplified, are the same organisations that later struggled to explain where the promised savings went. The wave of task-automation that came before this one produced glowing dashboards of processes automated and hours returned, and a great deal of that returned time evaporated for exactly the reason described here — released in fractions too thin to harvest, across teams whose shape never changed. Each time, the technology was new and the measurement error was old.

Tool-level measure	What it actually tells you	What to measure instead
Adoption rate	The tool was switched on	Whether the surrounding process was redesigned
Satisfaction score	People enjoy the experience	Whether decisions or outcomes changed

Tool-level measure	What it actually tells you	What to measure instead
Time saved per task	A contaminated local estimate	End-to-end cost and cycle time of the whole process
Projected headcount-equivalent	An unharvested fraction	Freed capacity actually consolidated to a destination

That is the pattern worth holding onto. The generative capabilities arriving now are more general, more capable, and more genuinely useful than anything in those earlier waves, and I do not doubt that the eventual value will be large. But usefulness at the task and value to the enterprise are separated by a stretch of organisational work — the consolidation of freed capacity, the redesign of the flow, the removal of the step that no longer needs a human — that no pilot metric has ever measured and no vendor slide has ever mentioned. The pilots are not lying when they say the technology works. They are simply answering a different question from the one the finance business partner asked.

The discipline, then, is almost aggressively unglamorous. Before celebrating another green dashboard, ask what work the enterprise will stop doing, or do with fewer hands, or decide differently, because of what the pilot proved. If there is no answer to that question, the value has not been deferred. It has been designed out — quietly, and with everyone's enthusiastic agreement.