TRANSFORMATION BOOK SERIES · No. 387

Methodology

BEYOND PILOT METRICS

Pilot Metrics Do Not Measure AI Value — A Method for Proving Operational Impact

A seven-phase method for tracing generative AI from technical performance to realised organisational value

Giovanni Leonardi

June 2024

Contents

Pilot Metrics Do Not Measure AI Value — A Method for Proving Operational Impact

Transformation · No. 387
Originally written June 2024 | Re-edited and re-rendered for the WhereBeyond Edition,

> *"An AI pilot proves that a system can perform; impact measurement proves that the organisation performs differently because of it."*

1. Executive Summary

In early 2024, many organisations can demonstrate that generative AI produces an impressive answer. Far fewer can demonstrate that it produces an improved operating result. Pilot dashboards commonly report model accuracy, user adoption, minutes saved in a test and favourable participant sentiment. These measures are useful for learning whether a prototype functions. They do not establish whether the organisation has reduced cost, improved service, released capacity, lowered risk or changed a decision that matters.

An AI pilot proves that a system can perform; impact measurement proves that the organisation performs differently because of it.

This paper sets out the AI Impact Chain Method, a practical sequence for moving from pilot metrics to defensible value evidence. The method follows seven phases: define the value boundary, reconstruct the baseline, write the measurement contract, instrument the full workflow, run a controlled deployment, separate the AI effect from surrounding change, and make an evidence-based scale decision. Each phase produces an artefact and ends in a gate. The method treats value as an operational claim with an owner, a counterfactual and a conversion mechanism—not as a property of the model.

The central discipline is simple: never count a theoretical minute as value until the workflow has converted it into a measurable consequence. Time saved but absorbed by queue growth is capacity, not cash. Recommendations generated but ignored are output, not adoption. Better answers that create additional review work may be quality improvement, but they are not productivity improvement.

Used properly, the method gives sponsors a credible basis for deciding whether to stop, redesign, contain or scale an AI intervention. It also gives finance, operations, technology

and risk teams a shared account of what changed and why.

2. The measurement failure hidden inside a successful pilot

A typical pilot ends with a slide that says users saved 22 minutes a day, 78 per cent would recommend the tool, and sampled outputs were acceptable in 91 per cent of cases. The figures feel precise. Yet the operating manager cannot answer three basic questions: Which work disappeared? What happened to the released time? Did any customer, employee or financial outcome improve?

The problem begins when the pilot defines success at the point where the model produces an output. Organisational value usually appears several steps later.

Layer	Typical question	Evidence required
Model	Can it produce a suitable response?	Quality, error pattern, latency, cost per inference
Task	Does it change how one activity is performed?	Completion time, rework, exception rate
Workflow	Does the end-to-end process improve?	Throughput, queue age, hand-offs, service level
Operating result	Is capacity, cost, quality or risk changed?	Unit economics, released capacity, loss avoided, customer outcome
Strategic result	Does the change advance a chosen priority?	Growth, resilience, control, speed of decision

A pilot can succeed at the first two layers and fail at the next three. A drafting assistant may reduce the first composition of a service response from eight minutes to three, while review rises from two minutes to seven. A coding assistant may increase local output while testing and integration queues lengthen. A knowledge assistant may attract heavy usage because employees are curious, without changing a single decision.

The AI Impact Chain Method prevents that category error by tracing evidence all the way from technical behaviour to an operating consequence.

3. The operating roles

Impact measurement cannot be delegated to the AI team alone. The work crosses technical performance, process design, financial treatment and risk. Name the roles

before defining the measures.

- Value owner: accountable for the operating outcome and for deciding what released capacity will become. Usually the process or business-unit owner.
- Measurement lead: designs the baseline, comparison and evidence plan. This may sit in transformation, analytics or performance management.
- Process owner: maps the real workflow, including workarounds, exceptions and downstream queues.
- Product and technical lead: instruments the intervention, records versions and explains model or workflow changes.
- Finance partner: agrees conversion rules for cost, revenue, capacity and avoided loss before results are known.
- Risk or control owner: defines unacceptable effects, guardrails and evidence for quality, fairness, privacy, security or regulated decisions.
- Front-line representatives: expose the work as performed rather than the work as documented.

One person may fill more than one role in a small pilot. The accountabilities must still be explicit. In particular, the technical lead should not be the sole judge of business value, and finance should not arrive after the pilot to challenge assumptions that were never agreed.

4. Phase: Define the value boundary

The first phase names the unit of change. Avoid ambitions such as “improve productivity with AI.” Define a specific population, workflow, decision and time horizon.

The input is the proposed use case and its strategic rationale. The output is a Value Boundary Canvas containing:

- the users and cases in scope;
- the first and last step of the measured workflow;
- the decision or task the AI changes;
- the intended operating outcome;
- material harms or displaced work to monitor;
- the period over which an effect should appear;
- the accountable value owner.

Write the value claim as a causal sentence:

If the intervention changes [task] for [population], then [workflow measure] should change because [mechanism], producing [operating result] within [period].

For example: “If assisted drafting reduces composition and retrieval time for routine written enquiries, then completed cases per paid hour should rise because advisers spend less time finding approved language, producing additional service capacity within eight weeks.”

The sentence is deliberately testable. It also exposes weak logic. If the promised result depends on a later staffing decision, demand increase or policy change, that dependency belongs in the claim.

Gate — boundary approval: the value owner confirms that the outcome matters, the workflow is narrow enough to observe, and the mechanism is plausible. If the team cannot write a causal sentence without vague verbs such as enable, enhance or empower, it is not ready to measure.

5. Phase: Reconstruct the baseline

A baseline is not last quarter’s average copied from a dashboard. It is a description of performance before intervention, segmented enough to reveal case mix and variability.

Start with four weeks of operational data where possible, longer where volumes are low or seasonality is material. Combine system records with direct observation and short work sampling. Capture:

- volume by case type and complexity;
- elapsed time and active handling time;
- queue age and service-level performance;
- rework, escalation and error;
- staffing hours and relevant non-labour cost;
- quality or risk outcomes;
- exceptional events affecting demand or capacity.

The baseline must use the same population and definitions planned for the deployment. Median handling time may be more informative than the mean where a small number of complex cases distort results. Averages should be segmented by work type, channel and experience level where these alter performance.

Create a Baseline Pack containing the data extract, definitions, sampling method, segmentation, observed workflow and known limitations. Freeze the definitions at the gate; otherwise teams are tempted to redefine success after seeing results.

A common objection is that the organisation has never measured the process cleanly, so a rigorous baseline will delay the pilot. The objection is often valid. The answer is not to invent precision. Use a proportionate baseline, state confidence ranges, and treat discovery of poor process data as a finding. A fast but unverifiable claim does not become more useful because the pilot moved quickly.

Gate — baseline fitness: the measurement lead and process owner agree that pre-intervention performance can be compared with later performance. Where it cannot, the deployment remains exploratory and must not carry a financial benefit claim.

6. Phase: Write the measurement contract

Before users touch the intervention, agree what will count as success, what will count as harm and how each result will be converted into value. Record this in a one-page Measurement Contract.

The contract should include one primary outcome, a small set of supporting measures and guardrails.

Measure type	Purpose	Example
Primary outcome	Tests the central value claim	Completed eligible cases per paid hour
Leading measure	Shows whether the mechanism is working	Draft preparation time
Adoption measure	Shows actual use in eligible work	Accepted assisted drafts as a share of eligible cases
Quality guardrail	Prevents speed being mistaken for value	Reopen rate within seven days
Risk guardrail	Protects material obligations	Confirmed policy or privacy breach
Economic conversion	States how operational change becomes value	Hours released × agreed loaded rate, subject to realisation

Set the success threshold before observing results. A threshold should include both magnitude and durability: for example, “at least 10 per cent improvement in completed cases per paid hour over six weeks, with no material decline in reopen rate or control

compliance."

Define the value conversion rule explicitly:

1. Productive capacity: verified hours released and redeployed to a named backlog, service improvement or growth activity.
2. Cashable saving: expenditure actually removed from a budget or avoided under an approved plan.
3. Revenue contribution: incremental margin linked to changed conversion, retention or volume, net of displacement.
4. Loss avoided: change in expected frequency or severity, using an agreed risk model and conservative assumptions.
5. Quality value: measurable reduction in defects, complaints, rework or adverse outcomes.

Never add these categories without checking overlap. If released hours are valued as capacity and also used to justify avoided hiring, the same effect may be counted twice.

Gate — measurement contract: value owner, finance partner and risk owner sign the definitions, thresholds, conversion rules and stop conditions. This gate prevents a pleasing technical result from being translated into whichever value story is convenient later.

7. Phase: Instrument the full workflow

Instrumentation should capture exposure, behaviour and outcome at case level without collecting unnecessary personal data. The minimum record links:

- an eligible case identifier;
- whether AI assistance was offered and used;
- intervention version and relevant configuration;
- task start, hand-off and completion times;
- acceptance, editing or override;
- final disposition and downstream outcome;
- error, escalation or guardrail event.

Do not instrument only the AI interface. The purpose is to see whether local acceleration creates downstream work. If the intervention drafts a response, measurement must continue through review, sending, reopening and complaint—not stop at generation.

Version changes require discipline. Changes to the model, prompt, retrieval material, workflow rules or user interface may alter the intervention. Record them in a Change Ledger with date, reason and expected effect. Material changes restart or segment the measurement period; otherwise different interventions are blended into one result.

The product lead and measurement lead should test the event record against five to ten real cases before launch. Reconstruct each journey manually. Missing hand-offs, duplicate timestamps and silent fallback paths are easier to fix before results accumulate.

Gate — observability: every primary measure and guardrail can be calculated from tested records, and the data collection is proportionate, lawful and understood by participants.

8. Phase: Run a controlled deployment

The preferred design is a concurrent comparison between similar eligible groups. Random assignment is strongest where operationally and ethically feasible. Where it is not, use matched teams, staggered introduction or within-person comparison, and document the limitations.

The Deployment Protocol must specify:

- eligibility and exclusions;
- assignment or comparison method;
- sample and duration;
- training and support;
- fallback procedure;
- stop conditions;
- permitted workflow changes;
- review cadence.

Avoid comparing enthusiastic volunteers with an unselected workforce. Volunteers often have higher digital confidence and stronger motivation, which exaggerates adoption and productivity. Also avoid measuring the first week as steady state. Early results contain training cost, novelty effects and unusual support.

Run the deployment long enough to observe learning and downstream effects. For high-volume routine work, several weeks may suffice. For low-volume decisions or risk outcomes, the correct answer may be that the pilot cannot establish impact statistically. That is a limit to disclose, not a reason to manufacture certainty.

Stop conditions should be operational, not rhetorical: a confirmed serious control breach; quality below an agreed threshold for two consecutive review periods; unexplained data loss; or a material change that invalidates the comparison.

Gate — evidence sufficiency: the measurement lead confirms that exposure, sample, duration and data quality support the intended claim. A pilot may continue for learning even if it cannot yet support a value claim.

9. Phase: Separate the AI effect

Results rarely emerge in a stable environment. Demand changes, experienced staff move teams, policy is simplified, seasonal work arrives and managers pay unusual attention to the pilot. The method therefore requires an Effect Reconciliation.

Begin with the difference between intervention and comparison performance. Then test the main alternative explanations:

- case mix changed;
- staffing or experience differed;
- another process improvement occurred;
- demand or queue conditions changed;
- measurement itself changed behaviour;
- poor cases were excluded after the fact;
- quality cost moved downstream.

Report ranges rather than false precision where attribution is uncertain. The result may be “between 7 and 11 per cent improvement, with approximately two points plausibly attributable to simplified guidance introduced in week three.” That statement is more decision-useful than claiming 12.4 per cent from a fragile model.

Measure total cost as well as benefit: licences or usage charges, integration, retrieval preparation, evaluation, support, human review, control, training and increased compute or storage. Pilot economics often exclude the effort required to keep approved knowledge current. At scale, that maintenance can become a material operating cost.

10. A worked example: from saved minutes to realised capacity

Consider a composite service operation with 60 advisers handling 2,400 written enquiries each week. The intervention retrieves approved guidance and drafts responses for

routine cases. The pilot team initially reports a 5.4-minute reduction in drafting time and calls this a 27 per cent productivity gain.

The method changes the conclusion.

The baseline shows 11.8 minutes of active handling per eligible case: 7.6 minutes drafting, 2.1 reviewing, and 2.1 recording and closing. Reopen rate is 8.5 per cent. Only 62 per cent of all enquiries are eligible.

During a six-week controlled deployment:

- drafting falls from 7.6 to 3.0 minutes;
- review rises from 2.1 to 4.0 minutes;
- recording remains 2.1 minutes;
- reopen rate rises from 8.5 to 9.1 per cent;
- assistance is used in 76 per cent of eligible cases;
- 14 per cent of generated drafts are abandoned.

The net active-time reduction is 2.6 minutes per assisted case, not 5.4. Applied only to eligible, actually assisted volume, this releases about 49 hours a week:

2,400 × 62% eligibility × 76% use × 2.6 minutes ÷ 60.

The operation then examines realisation. Thirty hours are assigned to an overdue complaints queue, reducing its oldest case from 18 days to 11. Twelve hours are absorbed by coaching and knowledge maintenance. Seven cannot be traced to a changed outcome. Finance therefore records 30 hours of productive capacity, no cashable saving, and an explicit operating cost for maintenance. Because the reopen increase remains inside the agreed guardrail but trends upward for complex cases, those cases are removed from eligibility before wider deployment.

This is a smaller number than the pilot headline, but a stronger result. It shows the mechanism, the leakage and the destination of the released capacity. It also provides a clear redesign decision rather than a ceremonial declaration of success.

> *The value is not the time the model appears to save. The value is the operating consequence the organisation can trace after friction, adoption, quality and realisation.*

11. Phase: Decide, redesign or scale

The final artefact is an Impact Decision Record. It brings together the causal claim, baseline, contracted thresholds, deployment evidence, costs, guardrails, attribution limits and realisation plan.

The decision is one of four:

1. Stop: the effect is absent, uneconomic or unsafe.
2. Redesign: the mechanism is promising but adoption, workflow, data or control prevents value.
3. Contain: value exists for a narrow case class, but broader use is not supported.
4. Scale: the outcome exceeds the threshold, guardrails hold, economics remain credible and the operating model can absorb the change.

Scaling creates a new measurement cycle, not the end of measurement. Volumes, populations, user skill and support demands change. Establish a benefits ledger with a named owner, monthly operational measures, quarterly value reconciliation and triggers for reassessment when the intervention or process changes materially.

The scale decision should state what management will do with capacity. Without that choice, efficiency becomes invisible slack. If the organisation intends a cashable saving, the workforce or supplier plan must show how it will occur. If it intends service improvement, the target queue, response time or quality outcome must be named.

12. Common failure modes and corrections

12.1 Counting usage as value

High usage may show curiosity, convenience or policy pressure. Connect use to an eligible case and then to an outcome. Adoption is a mechanism measure, never the final benefit.

12.2 Using self-reported time alone

User estimates are fast and may help form a hypothesis, but they are vulnerable to optimism and inconsistent task boundaries. Triangulate them with timestamps, case volumes and observation.

12.3 Averaging away the risk

A strong mean can conceal poor outcomes in a consequential minority. Segment by case type, user experience and affected group; examine the distribution and worst credible

failure, not only the average.

12.4 Claiming capacity as cash

Released minutes do not reduce expenditure by themselves. Record capacity until an approved action converts it into avoided cost, removed cost, additional volume or improved service.

12.5 Changing the intervention during measurement

Iteration is legitimate, especially in an immature field. Unrecorded iteration destroys attribution. Use the Change Ledger and treat material versions as separate interventions.

12.6 Letting the pilot team grade itself

The product team owns learning and delivery; it should not own the final value judgement alone. The value owner, finance partner and risk owner provide necessary independence without creating a distant committee.

13. The minimum pack a practitioner should leave behind

At the end of the cycle, the work should be reconstructable from seven artefacts:

1. Value Boundary Canvas — population, workflow, mechanism, outcome and owner.
2. Baseline Pack — definitions, segmentation, performance and limitations.
3. Measurement Contract — measures, thresholds, guardrails and conversion rules.
4. Change Ledger — intervention versions and contextual changes.
5. Deployment Protocol — comparison, duration, training and stop conditions.
6. Effect Reconciliation — observed change, alternative explanations, cost and confidence.
7. Impact Decision Record — stop, redesign, contain or scale, with the realisation plan.

These are not paperwork around the experiment. Together they are the chain of evidence that allows a decision-maker to distinguish a compelling demonstration from an operational improvement.

14. What good measurement changes

A rigorous method does more than produce a better benefit number. It changes the design of the intervention. Baseline work exposes where time is actually spent. Guardrails reveal which cases should remain human decisions. Instrumentation shows whether work has moved downstream. Realisation planning forces leaders to decide what released capacity is for.

That is why measuring AI impact should begin before the pilot, not after it. The organisation is not merely evaluating a model. It is testing a causal claim about work, behaviour and value. When that claim is explicit, measured against a credible alternative and followed through to an operating consequence, investment decisions become clearer. Some pilots will stop earlier. Some will narrow. The few that scale will do so with evidence strong enough to survive beyond the excitement of the demonstration.