WHITE PAPER · PORTFOLIO & AI GOVERNANCE

Governing AI as Operational Change

A practical standard for measuring what enterprise AI actually delivers

Marco Policani

Enterprise Portfolio · PMO · AI Operating Governance

policani.net · linkedin.com/in/marcpolicani

July 2026

The measurement gaps most programs ignore

Enterprise AI programs have a measurement problem, and it is not the one most teams think it is.

The data exists. Microsoft Copilot, OpenAI Enterprise, and comparable platforms expose seat-level telemetry by default — active users, session frequency, feature engagement, adoption curves. Most organizations can produce these reports within days of deployment. Many treat them as evidence of progress.

They are not. They are evidence of software activity.

Leadership funds AI to change recurring work: faster cycle times, more throughput, less rework, better decisions at the point of action. None of those outcomes appear in an active-user dashboard. The gap between what gets reported and what actually matters is where most AI programs quietly accumulate risk.

The scale of the problem is not theoretical. In a survey of 644 organizations, Gartner found that demonstrating AI value is the single biggest adoption barrier, cited by 49 percent of respondents — ahead of talent shortages, technical challenges, and data problems.

Over 80 percent of organizations have explored or piloted generative AI, and nearly 40 percent report deployment.[1] But widespread deployment has not produced widespread results: MIT's NANDA initiative found that 95 percent of enterprise AI programs deliver no measurable P&L impact, despite $30–40 billion in annual investment.[2] The same Gartner survey that found employees self-report saving an average of 3.6 hours per week through GenAI tools also found that not all employees see the same benefit, and that translating individual time savings into organization-level performance remains the central challenge.[3] Gartner separately projected that at least 30 percent of generative AI projects will be abandoned after proof of concept, with cost and unclear business value as the primary reasons.[4]

What these numbers describe is the same problem from different angles: AI is producing activity, and occasionally individual productivity gains, without reliably producing the organizational outcomes that justify continued investment. The argument here is straightforward: AI utilization is a workflow-performance question, not a software-activity question. What follows is a practical model for governing it that way.

Why current approaches fall short

Software-level metrics fail for five consistent reasons.

There is no denominator for real adoption. A user who opened the tool twice appears identical in most reports to a user whose weekly workflow now depends on it. Active users and meaningful users are not the same population.

Power users distort the picture. **A small group of enthusiastic early adopters can make an organization** look broadly deployed when the actual footprint is narrow. OpenAI's enterprise usage research confirms that frontier firms generate approximately seven times more messages to specialized AI tools than median enterprises — a distribution gap that averages reliably conceal.[5] Averages hide this. Distributions reveal it.

There is no baseline. If workflow performance was not measured before enablement, post-launch claims are retrospective and anecdotal. Organizations often treat this step as optional overhead. It is not.

Activity data and outcome data live in different systems. AI usage may be logged by user and date. Business outcomes live in Jira, ServiceNow, Salesforce, or a BI layer — logged by ticket, case, document, or transaction. Without a deliberate bridge between them, those signals cannot be interpreted together.

There is no accepted evidence threshold. Organizations often reject weak self-reporting but fail to replace it with anything better. This is a documented problem in organizational measurement: research has established that employees, whether consciously or not, tend to present their own performance favorably — a pattern that worsens under evaluative conditions and at scale.[6] Pre/post comparison, matched cohorts, and managerial validation are practical replacements. Most AI programs never formally adopt one.

The result is visible reporting without decision-grade signal.

A better standard for meaningful adoption

Meaningful adoption is not a utilization rate. It is a condition:

AI is in the critical path of recurring work, at sufficient scale and consistency, to produce measurable change in workflow outcomes or decision inputs — across more than a small concentration of power users.

This is the standard against which AI investment should be evaluated. It aligns measurement to what leadership actually funds. It uses data organizations already have. And it avoids the trust costs that come with intrusive monitoring.

The programs that sustain are the ones that establish this kind of discipline early. Gartner's research on AI maturity found that 45 percent of high-maturity organizations keep AI initiatives in production for three years or more, compared to 20 percent of low-maturity organizations. Trust is one of the primary differentiators: 57 percent of high-maturity organizations report that business units trust and are ready

to use new AI solutions, against just 14 percent of low-maturity organizations. Creating explicit metrics, the survey found, contributes directly to program efficacy.[7]

The three-layer evidence model

Meeting this standard requires three things working together, not any one of them in isolation.

Layer one: Platform telemetry

This is the breadth layer. Use platform-native analytics to confirm who is active, how often, and whether adoption is broad or concentrated. Microsoft's Copilot analytics architecture explicitly separates operational reporting (license assignment, adoption tracking) from strategic reporting (workflow impact, business outcomes) — a distinction that matters because mixing the two layers produces neither.[8] This data is easy to obtain and useful as a consistency check. It is not, by itself, evidence of business value.

Layer two: Workflow performance metrics

This is the value layer. The questions that matter here are operational: Did cycle time improve? Did throughput rise? Did quality hold? Did rework and escalations fall? These answers already exist in the systems organizations use to run their operations. The data does not need to be invented — it needs to be reviewed alongside the telemetry, not separately from it. Microsoft's Copilot Business Impact framework makes this integration explicit, recommending that organizations upload key operational metrics directly into the analytics layer to enable correlation analysis against usage patterns.[9]

Layer three: Explanatory validation

This is the interpretation layer. When workflow metrics move, something caused the movement. Disciplined comparison methods — pre/post analysis on the same workflow, matched cohort comparisons, manager pulse checks, quality sampling on selected deliverables — provide the structure for a defensible causal argument. These are standard quasi-experimental techniques in applied evaluation research, selected here because they are practical without requiring the experimental controls that enterprise settings rarely permit.[10] Perfect attribution is not the goal. Decision-grade evidence is.

NIST's AI Risk Management Framework (AI RMF 1.0) supports this approach in its MEASURE function, which calls for organizations to employ quantitative, qualitative, or mixed-method tools to analyze, assess, benchmark, and monitor AI impacts. The same function requires documenting aspects of AI systems' functionality and trustworthiness that cannot yet be measured reliably — a point directly relevant to the attribution problem in AI value measurement.[11]

Making it operational

The model requires four commitments before any workflow enters measurement.

Governance ownership

Three roles need to be explicit: a **Workflow Owner** who confirms the selected metrics reflect real operating performance; an **Analytics Owner** who validates source-system integrity and extraction logic; and an **AI Program Owner** who manages enablement and coordinates across teams. An executive review forum — a CIO staff meeting, transformation office, or AI steering council — reviews evidence and makes continuation decisions. These roles are a proposed minimum structure; organizations should adapt them to existing governance frameworks.

Workflow qualification

Not every AI use case warrants formal measurement. A workflow enters the program only if it has sufficient volume to produce usable data, repeats often enough to observe change in a reasonable timeframe, is already tracked in a system of record, and is important enough that leadership will act on the result. OpenAI's enterprise guidance similarly recommends focusing measurement efforts on use cases that can demonstrate clear, repeatable value — specifically a measurable change, a defined workflow, and alignment to organizational priorities — before treating an application as proven.[12]

Evidence thresholds, set in advance

No workflow should be described as value-generating without clearing a minimum standard:

- Baseline observation period: 4–8 weeks minimum (adjust to workflow cycle time — practitioner heuristic, not a validated standard)
- Post-enablement window: 4–8 weeks minimum, matched to baseline
- Primary operating metrics: at least two per workflow
- Comparison method: at least one (pre/post, matched cohort, or quality sampling)
- Telemetry check confirming breadth of use, not just aggregate activity
- Sign-off from both the Workflow Owner and Analytics Owner

The threshold must be defined before measurement begins. Programs that define it after almost always define it to fit the answer they already have.

A review cadence tied to decisions

Reporting that ends in a dashboard is not governance. Monthly operating reviews for workflow and analytics owners, combined with quarterly executive reviews, should drive four explicit dispositions for each workflow:

Status	Meaning
Scale	Evidence supports further rollout
Stabilize	Value exists but adoption consistency is uneven
Redesign	Workflow or enablement approach is flawed
Stop	No meaningful improvement against agreed metrics

These four outcomes are what separate management discipline from reporting theater.

A note on monitoring and trust

The model intentionally avoids intrusive measurement approaches — keystroke logging, screen monitoring, per-task AI attribution — not primarily for ethical reasons, but for practical ones.

Research from Cornell University's ILR School found that employees subjected to AI-driven monitoring report significantly greater loss of autonomy than those monitored by human supervisors, and are measurably more likely to exhibit resistance behaviors, complain more, perform worse, and express higher intention to quit. In four controlled experiments involving 1,195 total participants, AI-based evaluative surveillance consistently produced these effects. Framing matters: when surveillance was positioned as developmental support, the negative effects largely disappeared. When it was framed as performance evaluation, they did not.[13]

The organizational track record is instructive. In November 2020, Microsoft's Productivity Score feature exposed individual-level usage data across Microsoft 365 — showing exactly how much each named employee emailed, collaborated, and communicated. Within days of public criticism from privacy researchers, Microsoft's Corporate VP for Microsoft 365, Jared Spataro, announced the removal of all individual user names and per-person behavioral metrics from the product.[14]

Heavy-handed measurement architecture does not just create legal exposure. It damages the adoption it is supposed to track.

What this looks like in practice

Three workflow types illustrate how the model applies. These are not use-case recommendations — they are illustrations that the means of measurement are practical and generally already available.

Service and case workflows

- *Metrics:* case cycle time, throughput per agent per week, reopen rate within 7 days, escalation rate
- *Sources:* ServiceNow, Zendesk, Salesforce Service Cloud, internal BI
- *Evidence standard:* 6-week baseline vs. 6-week post-enablement period; confirm that usage is not concentrated in a small minority of agents before drawing conclusions about team-level performance

Document and proposal workflows

- *Metrics:* time to first usable draft, revision cycles, approval cycle time, rework rate
- *Sources:* document management systems, approval logs, SharePoint, PMO trackers, BI extracts
- *Evidence standard:* matched cohort comparison — teams using AI-assisted drafting against comparable teams or a prior period — paired with manager validation of whether recovered time translated to volume, quality, or speed

Requirements and internal analysis workflows

- *Metrics:* draft turnaround time, revision count, decision latency, defect leakage into downstream work
- *Sources:* Jira, Azure DevOps, PMO artifacts, governance logs
- *Evidence standard:* timing metrics should be paired with at least one quality or downstream-stability metric, so that speed gains are not mistaken for performance improvement if rework rises later

---

The shift this produces

Before this model, AI is typically governed as a software rollout: adoption dashboards, active user reports, usage curves, and generalized claims about productivity.

After this model, AI is governed as an operating intervention: workflow baselines, evidence thresholds, defined owners, and explicit decisions about where investment should expand or stop.

That shift is not about collecting more data. It is about establishing better control over what claims can be made from the data that already exists.

The underlying argument

The central management error in most AI programs is not technical. It is evidentiary.

Organizations ask the easiest question first: Are people using the tool? That question has a place — but only at the surface. The question that justifies continued investment is a different one: Has AI changed a recurring workflow in a way that warrants scaling?

The 5 percent of programs that are producing measurable returns share a common characteristic: they evaluate tools on business outcomes, not software benchmarks, and they hold vendors and internal teams accountable to operational metrics.[15] The discipline is not exotic. It is just uncommon.

If a program has delivered, it should be able to demonstrate it. If it cannot, the honest answer is that it is still in an experimentation phase — which is a legitimate stage, but a different claim than operational impact.

AI that does not change a workflow metric leadership already cares about may be promising. It is not yet meaningful adoption.

Notes and references

1. Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025, July). The GenAI Divide: State of AI in Business 2025 (v0.1). MIT NANDA (Networked Agents and Decentralized Architecture). Based on 52 structured interviews, systematic review of 300+ publicly disclosed AI initiatives, and surveys with 153 senior leaders (January–June 2025). Report PDF: https://www.artificialintelligence-news.com/wp-content/uploads/2025/08/ai_report_2025.pdf — The 80%/40% deployment figures are drawn from this report. Note: labeled v0.1 preliminary findings; the report acknowledges potential selection bias in participating organizations. ↑

2. Challapally et al. (2025), ibid. — The 95% figure refers specifically to organizations achieving no measurable P&L impact from GenAI investment. The report's primary explanations are a "learning gap" (tools that do not adapt to context or improve over time) and poor workflow integration — which is the operational measurement problem described in this document. ↑

3. Gartner, "Gartner Identifies Four Emerging Challenges to Delivering Value from AI Safely and at Scale," October 21, 2024. Survey of over 5,000 digital workers in the U.S., UK, India, Australia, and China, Q2 2024. https://www.gartner.com/en/newsroom/press-releases/2024-10-21-gartner-identifies-four-emerging-challenges-to-delivering-value-from-ai-safely-and-at-scale — The 3.6 hours/week figure is self-reported. Gartner's own analysis notes that productivity gains are not equally distributed across employees, job complexity, or experience levels. ↑

4. Gartner, "Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025," July 29, 2024. https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025 ↑

5. OpenAI, "The State of Enterprise AI: 2025 Report." https://openai.com/business/guides-and-resources/the-state-of-enterprise-ai-2025-report/ ↑

6. Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. Journal of Business and Psychology, 17(2), 245–260. https://doi.org/10.1023/A:1019637632584 — The authors document that social desirability and impression management systematically bias self-reported performance data, particularly under evaluative conditions. They find that supplementary objective data sources are required for reliable measurement. ↑

7. Gartner, "Gartner Survey Finds 45% of Organizations With High AI Maturity Keep AI Projects Operational for at Least Three Years," June 30, 2025. Survey of 432 respondents, Q4 2024. https://www.gartner.com/en/newsroom/press-releases/2025-06-30-gartner-survey-finds-forty-five-percent-of-organizations-with-high-artificial-intelligence-maturity-keep-artificial-intelligence-projects-operational-for-at-least-three-years — High-maturity organizations scored 4.2–4.5 on the Gartner AI Maturity Model (scale 1–5). The trust comparison (57% vs. 14%) is drawn directly from this press release. ↑

8. Microsoft, "Copilot Control System: Measurement and Reporting," Microsoft Learn. https://learn.microsoft.com/en-us/copilot/microsoft-365/copilot-control-system/measurement-reporting — The documentation distinguishes operational reports (Microsoft 365 admin center), strategic reports (Copilot Dashboard in Viva Insights), and advanced custom reporting. ↑

9. Microsoft, "Copilot Analytics Introduction," Microsoft Learn. https://learn.microsoft.com/en-us/viva/insights/copilot-analytics-introduction — The Copilot Business Impact Report requires organizations to upload their own operational metrics to enable correlation analysis against platform usage data. ↑

10. Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Houghton Mifflin. — The pre/post and matched cohort designs referenced here are standard quasi-experimental approaches. Their application as the minimum viable evidence standard for enterprise AI is a practitioner judgment call, appropriate where full random assignment is not feasible. ↑

11. National Institute of Standards and Technology. (2023, January). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). Tabassi, E. https://doi.org/10.6028/NIST.AI.100-1 — MEASURE function language appears in the core framework and is elaborated in the companion AI RMF Playbook. ↑

12. OpenAI, "Find and Share AI Use Cases to Show Impact," OpenAI Academy. https://academy.openai.com/public/clubs/champions-ecqup/resources/find-and-share-ai-use-cases-to-show-impact ↑

13. Schlund, R., & Zitek, E. M. (2024). Algorithmic versus human surveillance leads to lower perceptions of autonomy and increased resistance. Communications Psychology, 2, article 53. https://doi.org/10.1038/s44271-024-00102-8 PMID: 39242768; PMCID: PMC11332184. ILR School,

Cornell University. Study 1 N=107, Study 2 N=157, Study 3 N=117, Study 4 N=814; total N=1,195. Open access; data and analysis code: https://osf.io/3ztpm/ ↑

14. Spataro, J. (2020, December 1). Our commitment to privacy in Microsoft Productivity Score. Microsoft 365 Blog. https://www.microsoft.com/en-us/microsoft-365/blog/2020/12/01/our-commitment-to-privacy-in-microsoft-productivity-score/ — Spataro announced removal of all individual user names and per-person behavioral metrics following public criticism from privacy researchers, notably Wolfie Christl of Cracked Labs. ↑

15. Challapally et al. (2025), ibid. — The report describes organizations producing measurable returns as those that "evaluate tools based on business outcomes rather than software benchmarks" and "hold vendors accountable to business metrics." ↑

About the author

Marco Policani is an enterprise portfolio, PMO, and AI operating-governance leader. He builds portfolio governance systems where AI carries the analytical load and named humans own every decision — an approach documented across case studies, walkthroughs, and working governance guides on his portfolio site.

Portfolio & governance library: policani.net/governance · Full portfolio: policani.net · LinkedIn: linkedin.com/in/marcpolicani