11

WHITE PAPER · WORK, ADOPTION & JUDGMENT

Felt Productivity Is Not Measured Productivity

Why AI adoption should be governed on evidence of the work, not on how fast it feels

Marco Policani

Enterprise Portfolio · PMO · AI Operating Governance

policani.net · linkedin.com/in/marcpolicani

August 2026

EXECUTIVE SUMMARY

Govern AI on evidence of the work: a baseline, a work-output signal, and a review that keeps felt speed and delivered outcome in separate columns.

Self-reported AI speedup is a feeling, not operating evidence, and the people using AI are unreliable witnesses to their own speed in both directions. Scaling a program on that signal mistakes a feeling for evidence. This paper sets a small measurement standard: a baseline captured before the tool arrives, a work-output signal that activity cannot inflate, and a review that keeps felt speed and delivered outcome in separate columns.

01 Treat self-reported speedup as sentiment, not evidence: in a randomized trial, developers believed AI sped them up 20 percent while the measured work ran 19 percent slower.

02 Measure delivered output against a pre-tool baseline, not adoption charts, prompt counts, or survey speedup.

03 Keep felt experience and delivered outcome in two columns; a divergence between them is the finding, not an anomaly to reconcile.

04 Capture the baseline and fund the instrumentation now, because wider AI adoption is erasing the unaided comparison the standard depends on.

CONTENTS

Executive summary
The speedup you can feel and cannot find
Why a feeling cannot govern
Measure the work, not the mood
What to measure, and what to ignore
Governing on the wrong number has a cost
A review that fires on evidence
The one-page AI work-evidence standard
What this evidence does and does not establish
The distinction that governs
Notes and sources

Executive summary

Ask people how much faster AI made them and the answer is almost always a confident number. That number is a feeling, and a feeling is not operating evidence. The most useful recent finding on this is not that AI is slow, but that the

people using it are poor witnesses to their own speed, in both directions — and that honest measurement is getting harder to take, not easier.

For a portfolio leader, this is a governance problem before it is a technology problem. When a program is funded, scaled, or defended on the strength of a self-reported speedup, the organization is treating a feeling as evidence. This paper sets out a small measurement standard that keeps the feeling and the result apart: a baseline captured before the tool arrives, a work-output signal that is not an activity count, and a review that separates felt speed from delivered outcome. It also names what to stop measuring, because some of the most-quoted AI metrics measure adoption or effort, not value. The standard is a governance discipline, not a verdict on any tool.

The speedup you can feel and cannot find

Start with the finding that unsettles the confident number. In a 2025 randomized controlled trial, the research nonprofit METR had sixteen experienced open-source developers complete 246 real tasks in their own repositories, with and without AI assistance. Before the trial, the developers expected the tools to speed them up by 24 percent. After the trial, having lived through it, they still believed AI had sped them up by 20 percent. The measured result ran the other way: they took 19 percent longer with the tools than without [1].

EXHIBIT 1

Capable people misread their own speed

19% vs 20%

In METR's randomized trial, experienced developers were 19 percent slower with AI, yet afterward still believed it had sped them up by 20 percent — the gap this standard governs.

METR randomized controlled trial: 16 experienced open-source developers, 246 real tasks, July 2025.

The temptation is to read that as "AI makes people slower." It does not say that, and treating it that way would repeat the exact error the study exposes. This was sixteen expert developers, on mature codebases they knew well, using early-2025 tooling, measured at one moment. Change any of those conditions and the number can move. The durable lesson is the gap between what capable people believed about their own speed and what the clock recorded. The belief survived direct contact with the evidence.

The same divergence shows up at the level of a whole delivery system, from a different vantage point. Google's DORA program, in its 2024 Accelerate State of DevOps report, found that AI adoption "significantly increases individual productivity, flow, and job satisfaction" while it "negatively impacts software delivery stability and throughput."[3] Read those two clauses together. The people felt more productive and were more satisfied, and the delivery system they worked in became less stable and pushed less work through. Individual experience and system output pointed in opposite directions in the same dataset.

Neither of these is the last word, and both are snapshots. METR's authors have since said as much, which is the point of the next section. But together they establish the thing a governance standard has to account for: the felt experience of speed and the measured behavior of the work are separate variables, and the first is not a reliable proxy for the second.

Why a feeling cannot govern

A feeling cannot govern because the witnesses are unreliable and they do not know it. METR's own follow-up, published in February 2026, is blunt about the self-report problem: the developers' estimates of their own speedup, it notes, "can be quite unreliable."[2] That is not a claim that people lie. The internal sensation of fluency — the tool answered instantly, the draft appeared, the blank page filled — is simply a poor instrument for measuring elapsed time and reworked output. Fluency feels like progress even when it is followed by review, correction, and re-prompting that erase the saving.

The forecast-versus-actual pattern in the trial makes the mechanism visible. A 24 percent expected gain is optimism about a new capability. A 20 percent remembered gain, reported after the work was done, is the more troubling figure, because experience was supposed to correct the estimate and did not. If lived experience does not calibrate the number, then a survey asking employees how much AI helped is measuring enthusiasm, tool preference, and the natural wish to have spent one's effort well — not help.

There is a second reason self-report cannot carry a governance decision, and it is getting worse rather than better. METR's follow-up reports that the "wider adoption of AI has made it more difficult to measure task-level productivity" — the control condition is disappearing, because developers now reach for the tools reflexively even when a study asks them not to [2]. The honest reading is uncomfortable for anyone who wanted a clean number: the more normal AI use becomes, the harder it is to isolate its effect, and the more tempting it becomes to fall back on the one number that is always available and always flattering, which is what people say about themselves.

The after-estimate is the load-bearing failure, and it is worth dwelling on why. A forecast can be wrong because the future is unknown; that is forgivable and self-correcting. A memory of work already completed is supposed to be anchored to what happened. When sixteen people who have just been slowed down report that they were sped up, the instrument that failed is not prediction but recollection, and recollection is exactly what an enterprise survey samples. Worse, the people an organization asks

about a tool are disproportionately the ones who chose it, championed it, or enjoy it. The sample is not neutral, and the question invites the answer.

This is why the direction of the evidence matters more than any single figure. The durable claim is not about speed at all: self-reported speedup is a low-quality signal that systematically runs ahead of measured output, and an organization scaling AI on that signal is scaling on sentiment. A portfolio does not fund sentiment on purpose. It does so by default, when no one built the alternative.

Measure the work, not the mood

The alternative is ordinary: the discipline of measuring an operating change, applied to a category of change that arrived wrapped in enough novelty to make people forget they already knew how. NIST's AI Risk Management Framework puts measurement at the center of its lifecycle: its MEASURE function uses "quantitative, qualitative, or mixed-method" methods to analyze and monitor AI systems, so that where trade-offs arise, "measurement provides a traceable basis" for the decision [5]. Measurement here works as a control, not an after-the-fact report: it lets a leader decide whether reliance on the system should grow, hold, or stop.

So the standard this paper proposes is narrow and testable. An AI-assisted workflow is being governed on evidence when three things exist: a baseline captured before the tool changed the work, a work-output signal that reflects delivered value rather than activity, and a review that reads the felt experience and the delivered outcome as two separate records. If a program can show a rising adoption chart and a glowing internal survey but cannot produce those three things, it is being administered on impression, however disciplined the rollout looked.

Set a baseline before the tool arrives

The most common measurement failure is the one a CIO feature on AI's ROI captured in a single line from a former global CIO: "We didn't start with a baseline."[6] Without a before-picture, every after-picture is unfalsifiable. Any improvement can be attributed to the tool and any decline explained away, because there is no honest record of how the work performed when a human did it unaided. The baseline does not need to be elaborate. It needs to be captured before the workflow changes, on the same measure the review will use afterward, because a baseline reconstructed from memory after adoption inherits exactly the optimism this whole problem is made of.

Choose a work-output signal, not an activity count

A baseline is only as good as what it measures, and this is where most AI dashboards go wrong. They count activity — seats provisioned, prompts sent, tokens consumed, features accepted — because activity is easy to instrument and always goes up. The SPACE framework, from Microsoft and GitHub researchers writing in ACM Queue, was built to correct exactly this reflex: developer productivity, it

argues, is about "more than an individual's activity levels" and "cannot be measured by a single metric or dimension."[4] SPACE names five: satisfaction and well-being, performance, activity, communication and collaboration, and efficiency and flow. The useful discipline for a non-software workflow is to notice which of those an AI metric actually touches. Adoption charts and prompt counts are activity. Survey speedup is satisfaction. Only performance and, sometimes, efficiency describe whether the work came out better — and those are the two that instrumentation usually skips because they are the hardest to capture.

The lesson generalizes past software. A work-output signal answers a question about delivered value — cycle time on a real case, rework rate, error rate, coverage, throughput of finished units that met standard — and it is chosen so that it cannot be moved by simply using the tool more. Usage is an input. Output is what the portfolio funded.

Separate felt speed from delivered outcome

The third element is the one the evidence most directly demands. The review keeps two columns and never collapses them: what people report about the experience of the work, and what the work actually did against the baseline. Both are legitimate. Satisfaction and flow are real, they matter for retention and adoption, and DORA found AI genuinely raised them [3]. But they are recorded as experience, next to the output signal, not merged into it. When the two columns agree, reliance can grow with confidence. When they diverge — the team feels faster while the delivered work has not moved, or has slipped — that divergence is the most valuable thing the review produces, because it is precisely the case that a satisfaction survey alone would have scored as a win.

What to measure, and what to ignore

The hard part is deciding which numbers do not count as evidence. Adoption rate, active-user counts, prompts per employee, and self-rated speedup are all easy to collect and all point the wrong way: they measure that AI is being used and that people enjoy using it, not that the work improved. They belong on an adoption dashboard, where the question is whether a rollout reached people. They do not belong in the evidence a portfolio uses to decide whether to deepen reliance, because none of them can distinguish a workflow that got better from one that merely got busier.

The signals worth keeping share a property: they live where the value is actually created or lost. A CIO analysis of the AI ROI gap put this exactly — the governance machinery of "steering committees, approval gates and KPI dashboards" sits above the work, while "the value, or the leak, happens inside the workflow, at the level of the actual task the AI is doing." You can, the same piece warns, "instrument your governance structure perfectly and still have nothing measuring whether the agent's output was right at the point where it mattered."[7] That is the difference between a metric that governs and a metric that reassures. And it is why local task gains so often fail to reach the bottom line: as the ROI-measurement reporting found, there is frequently "no straight line between the investment

in that tool and increased revenues," often because a local saving is absorbed somewhere downstream that no one instrumented.

Two cautions bound the measurement itself. First, "that measurement is fundamentally lacking" in many organizations not because the metrics are unknown but because no one instrumented the workflow to produce them; measurement is work that has to be funded, not a byproduct that appears once a tool is bought [6]. Second, the instinct to measure harder can curdle into surveillance, and surveillance damages the very work it watches. The goal is a small number of honest output signals tied to a baseline, owned by someone accountable for the workflow — not a monitoring apparatus that measures keystrokes and erodes the trust that adoption depends on.

Governing on the wrong number has a cost

It is tempting to treat this as a measurement nicety, but the cost of governing on felt speed is real and it lands on the portfolio. The first cost is misallocated capacity. When a scaling decision is made on a survey, the tools and workflows that produce the strongest feelings attract the next round of investment, whether or not they produced the strongest results — and scarce integration, change, and review capacity follows the enthusiasm rather than the evidence. The organization compounds its bets on the wrong thing precisely because the wrong thing felt best.

The second cost is credibility, and it is slower and more corrosive. A program promoted on a promised speedup eventually meets the results it was supposed to move, and when the delivered numbers do not show the gain the survey implied, the gap does not read as a measurement error. It reads as a broken promise, and it spends the sponsor's credibility and the wider appetite for AI investment along with it. The third cost is downstream: savings claimed on felt speed are sometimes cashed in advance — in a headcount plan, a benefits case, or a budget cut justified by an efficiency no one measured. A saving that existed only as a feeling still has to be paid back in rework, escalation, and quiet backfill once the work fails to deliver it. Measuring the work is cheaper than any of these, and it is cheapest when done before the commitment, not after the shortfall.

A review that fires on evidence

Here is the shape the standard takes when a real review runs it, described as an operating pattern rather than tied to any one organization. A team adopts an AI assistant across a recurring, high-volume workflow — the kind where a small per-case saving would compound into a real number. Adoption is fast and the sentiment is excellent; the internal survey reports a large felt speedup, and the case for expanding the license writes itself. Under an impression-led process, that is where the decision gets made.

Under this standard, the survey is one column, and it is not the column that decides. The review pulls the work-output signal the team agreed on before rollout — cycle time on completed cases that met the quality bar — and compares it against the baseline captured from the pre-tool period. In the pattern I have seen hold up, the two columns disagree more often than sponsors expect: the felt speedup is genuine, and the measured cycle time on completed, standard-meeting cases has barely moved. The measurement does not, by itself, say where the saving went — only that it did not reach the delivered outcome. That is the finding. A real, local time saving stopped somewhere short of the result the portfolio funded, and only the output signal made the shortfall visible; where exactly it stopped is then the workflow owner's question to run down.

That disagreement is not a reason to abandon the tool. It is the review doing its job. It reframes the next decision from "expand because people love it" to "the workflow, not the tool, is where the value is stuck — redesign the review step, then re-measure." The felt speed told the team something true about the experience of the work. The output signal told them something true about the delivery of it. Only holding both let them act on the difference instead of funding past it.

The standard is not only a brake. When the two columns agree — the team reports a genuine speedup and the output signal on completed, standard-meeting work has moved with it against the baseline — the review produces the opposite verdict, and produces it with evidence a skeptical finance partner can inspect. That is the case for deepening reliance, expanding the license, and moving the workflow from trial to standing operation. Separating the columns does not mean distrusting every good feeling; it makes the yes as evidenced as the no, so that expansion rests on delivered work rather than on the persuasiveness of the people who liked the tool.

The one-page AI work-evidence standard

The discipline should fit on a single page and survive a change of sponsor, so that when a scaling decision arrives the evidence is already defined and only the judgment remains. A workable standard names seven things before an AI-assisted workflow is scaled.

EXHIBIT 2

A one-page standard names seven things before a workflow is scaled

1. **Named workflow and owner**
   The recurring workflow under review and the person accountable for its delivered outcome, not for the tool.
2. **Pre-tool baseline**
   The output measure captured before the tool changed the work, on the same definition the review will reuse.

3. **Work-output signal**
   A delivered-value measure — cycle time, rework, error rate, finished units — that using the tool more cannot move by itself.
4. **Felt-experience record**
   Satisfaction, flow, and self-reported speed, captured honestly and kept in their own column, never merged into output.
5. **The separation review**
   A standing check that reads the two columns side by side and treats a divergence between them as the primary finding.
6. **The reliance decision**
   An explicit outcome each cycle — deepen, hold, redesign the workflow, or stop — tied to the signal and baseline, not the adoption chart.
7. **The measurement cost**
   A funded commitment to instrument the workflow, because an unfunded standard quietly reverts to self-report.

A standard of this shape does not slow adoption; it protects it. It lets a leader say yes to deeper reliance on evidence, and lets them redirect rather than abandon a tool whose value is leaking inside the workflow. Most of all it prevents the specific failure the research warns about: scaling a capability across an organization on the strength of how fast it felt.

What this evidence does and does not establish

Honesty about the evidence is part of the standard, because a measurement discipline that overstates its own proof would be self-refuting. The METR trial is a rigorous randomized controlled trial, and it is also small and specific: sixteen expert developers, mature repositories, early-2025 tools, one point in time [1]. It does not license a general claim that AI reduces productivity, and this paper makes no such claim. DORA's finding is an association across a survey population, not a controlled experiment, and it too is a 2024 snapshot of a fast-moving practice [3]. What earns confidence is not either number in isolation but their agreement on the shape of the problem: felt experience and measured output diverge, and the divergence favors the feeling.

EXHIBIT 3

What the evidence establishes, and what it does not

METR trial
A randomized trial shows a large, persistent gap between felt speed and measured output.

DORA 2024
System-level snapshot: adoption raised satisfaction as delivery stability and throughput fell.

SPACE framework
Productivity is multidimensional; no single metric or activity count captures it.

NIST AI RMF
Measurement is an operating control that gives management decisions a traceable basis.

Not established
No universal metric, cadence, or benefit multiplier for AI; both anchors are snapshots.

The deeper caution comes from METR's own follow-up. Because wider adoption is making task-level measurement harder, the clean baseline this standard depends on is getting more expensive to capture, not less, and organizations that wait will find the unaided comparison has vanished from their own operations [2]. That is an argument for building the measurement discipline now, while a before-picture is still possible, rather than a reason to defer it until the tools "settle." No source here establishes a universal work-output metric, a required cadence, or a benefit multiplier for AI; those are properties of a specific workflow, and the standard is deliberately a method for finding them rather than a set of numbers to adopt.

The distinction that governs

The test of AI governance is not the enthusiasm of a rollout or the confidence of a survey. Any capable tool will produce both, and produce them early. The test is whether, when the felt speed and the delivered outcome disagree, the organization can see the disagreement and act on the outcome. An enterprise that measures the work — a baseline, a real output signal, and a review that keeps the feeling and the result in separate columns — is governing its AI. An enterprise that scales on how fast the work feels is governing its mood, and funding it as if it were evidence.

Felt productivity is real, and it is worth having. It is simply not the same thing as measured productivity, and only one of the two can tell a portfolio where its money went.

Notes and sources

[1] Joel Becker, Nate Rush, Beth Barnes, and David Rein, Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, METR, July 10, 2025. https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

[2] METR, We Are Changing Our Developer Productivity Experiment Design, February 24, 2026. https://metr.org/blog/2026-02-24-uplift-update/

[3] Google Cloud DORA, Accelerate State of DevOps Report 2024, Google Cloud, 2024. https://dora.dev/research/2024/dora-report/

[4] Nicole Forsgren, Margaret-Anne Storey, Chandra Maddila, Tom Zimmermann, Brian Houck, and Jenna Butler, The SPACE of Developer Productivity: There's More to It Than You Think, ACM Queue, Vol. 19, No. 1, 2021. https://www.microsoft.com/en-us/research/publication/the-space-of-developer-productivity-theres-more-to-it-than-you-think/

[5] National Institute of Standards and Technology, AI Risk Management Framework (AI RMF 1.0), Core: the MEASURE function, NIST, 2023. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[6] Maria Korolov, Why Is It So Hard to Measure the ROI of AI?, CIO, 2025. https://www.cio.com/article/4183502/why-is-it-so-hard-to-measure-the-roi-of-ai.html

[7] Pete Johnson, The AI ROI Gap Isn't a Model Problem. It's a Workflow Problem, CIO, 2025. https://www.cio.com/article/4193951/the-ai-roi-gap-isnt-a-model-problem-its-a-workflow-problem.html

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