RICARDO VARGAS

EMPLOYEES HAVE NO INCENTIVE TO DISCLOSE HOW PRODUCTIVE AI MAKES THEM

Ricardo Viana Vargas, Ph.D.

João Henrique Jacinto

Antonio Nieto-Rodriguez

Published In

London School of Economics Business Review (Adapted Version)
London - United Kingdom – November 2026

Introduction

An AI productivity trial can report 26 minutes saved per person per day and still leave the hardest management question unanswered: which of those minutes became value, and who felt safe enough to say so?

Most AI programs never ask the second half of that question. In the programs we see, the tools work, people save time and early satisfaction is high. Then leadership looks for the gain in enterprise performance and struggles to find it. The time was saved. It just never appeared anywhere the organization could see it.

We call this the invisible AI dividend: productivity a company has paid for and may never capture, because it cannot see it. Our argument is that the dividend stays invisible for reasons that are more behavioral than technical, and that the usual management response, which is to measure harder and to measure each person, makes the problem worse. We build the argument on the largest Microsoft 365 Copilot deployment in UK government to date, add evidence from controlled studies in the private sector, and then set out a framework that leaders can apply to their own programs.

What the UK government's Copilot trials show, and what they cannot

Between 30 September and 31 December 2024, the UK government ran what the Government Digital Service described as "the largest deployment of M365 Copilot in UK government to date."1 Some 20,000 employees in 12 organizations received access to Microsoft 365 Copilot, and the evaluation survey collected 7,115 responses. Participants reported saving an average of 26 minutes a day, which the report notes would add up to 13 days over a full working year. Satisfaction was high: 82 percent said they would not want to return to their pre-Copilot working conditions.

The report is open about how these figures were produced. The time savings were self-reported. Respondents chose a range, from "none or less than 5 minutes" up to "more than an hour," and the analysts used the midpoint of each range, counting the top range as 60 minutes. The report is equally open about what it could not establish: "due to experimental constraints it was not possible to identify how time saved was spent." More than 70 percent of users agreed that the tool had increased the time they spent on more strategic activities, but that was also the users' own account.

The Department for Work and Pensions (DWP) ran a more demanding evaluation

of its own Copilot trial.2 More than 3,500 DWP staff had access to the tool. In February and March 2025, the department collected survey responses from 1,716 users and 2,535 non-users and compared the two groups with regression analysis. It estimated that users saved 19 minutes a day across eight routine tasks, among them writing emails, summarizing information, searching for existing research and transcribing meetings, with a 95 percent confidence interval of 17 to 22 minutes. Users also reported higher job satisfaction and higher quality of work, and both results were statistically significant ($p < 0.01$). Asked what they did with the time, 89 percent said they had reallocated it to other work tasks.

These are careful evaluations, and their authors name their own limits. At DWP, licenses were allocated on a first-come, first-served basis, so early adopters may differ from everyone else. No measures were collected before the tool was introduced, so there is no baseline. And every outcome, including the reallocation of time, rests on what staff chose to report.

For our argument, these limits are the main finding. Both evaluations, with thousands of respondents, point to the same place. The gain is real, and at DWP it holds up against a comparison group. Yet in both cases, where the time went is known only to the people who saved it. The organization that paid for the tool can see adoption dashboards and survey answers. It cannot see which minutes turned into faster service, better decisions or lower cost, and which were quietly absorbed.

A second point is easy to miss. Both figures came from evaluation surveys designed to assess the tool rather than the people using it. In a trial, reporting a gain costs the respondent very little. The harder test comes after the trial, when AI becomes part of normal work and its effects start to show up in targets and performance reviews. That is the moment when the incentive to disclose can reverse.

The gain is real, and it is not unique to one product

The pattern is not specific to one product or to the public sector. In a study of 5,179 customer support agents, Brynjolfsson, Li and Raymond found that access to a generative AI assistant raised the number of issues resolved per hour by 14 percent on average, and by 34 percent for novice and low-skilled workers.3 In a field experiment with 758 Boston Consulting Group consultants, Dell'Acqua and colleagues found that, on tasks inside the AI's capabilities, consultants using AI completed 12.2 percent more tasks, finished 25.1 percent more quickly and produced results of more than 40 percent higher quality than a control group.4

The same experiment carries a warning that matters for measurement. On a task

chosen to fall outside the AI's capabilities, consultants using AI were 19 percentage points less likely to produce a correct solution. The authors call this a "jagged technological frontier": AI improves some tasks and makes others worse, even when the tasks look similar. A single figure of hours saved per person hides that unevenness. An organization needs to see where AI improves the work and where it quietly degrades it, and self-reported time savings show neither.

Read together, these studies define the problem. Under controlled conditions, with firm data or independent graders, the gain can be measured. In day-to-day operations those conditions rarely exist, and the only reliable witness to the gain is often the person who made it.

Why the gains go quiet

The first person to know that AI saved time is usually the person who saved it. That person then decides, consciously or not, what disclosure will mean. Revealing the gain might bring more work, tighter targets or a question about whether the role needs the same hours. Staying silent is a reasonable response to those incentives, and it requires no bad faith.

A second reason needs no threat from management at all. In many workplaces, visible effort still stands in for value. Admitting that AI helped produce a piece of work can make the work look less earned, even when the judgment behind it is the person's own. The gain stays private because revealing it carries a social risk, whatever the official policy says.

This is already happening at scale. Microsoft and LinkedIn's 2024 Work Trend Index, based on a survey of 31,000 people in 31 markets, found that 75 percent of knowledge workers were using AI at work and that 78 percent of AI users were bringing their own tools.5 Among people using AI at work, 52 percent were reluctant to admit using it for their most important tasks, and 53 percent worried that using it on important tasks made them look replaceable. Much of this use happens outside any official program. Leaders are not deciding whether concealment will become a problem. They are deciding how to deal with one that already exists.

The instinctive response to invisible productivity is to measure more, and to measure each person. In AI programs, that instinct backfires. When productivity data is attached to individuals, the number gains a potential defendant. It can be used to reward people, and also to compare them, to raise their targets or to justify cuts. Each person then has a rational reason to understate what AI makes possible. Social scientists have described a version of this for decades. Goodhart

observed that a statistical regularity tends to collapse once pressure is placed on it for control purposes,6 and Strathern's widely quoted paraphrase states it more simply: when a measure becomes a target, it ceases to be a good measure.7

AI adds a sharper variant. In the classic case, people change their behavior to hit a number that already exists. With AI productivity, the distortion can happen before any number is recorded, because the only person who knows about the gain decides not to report it. The number is never created.

This is the AI disclosure paradox. The harder a company pushes to measure individual AI productivity, the stronger the reason it gives each person to keep that productivity out of view. The gains leaders most need to see become the ones employees have the least reason to reveal.

Disclosure-safe measurement: a framework for leaders

Measuring less is not the answer. Leaders need to change what is measured, who can see it and what happens afterwards, so that telling the truth about AI gains becomes the sensible choice for the people who hold them. We call this disclosure-safe measurement. It has four parts, and they work as a set. Applied alone, each part tends to fail.

1. Measure the workflow, not the person

Start with the unit of analysis. The individual question, how many hours each analyst saved with AI, invites concealment because the answer points at someone. The workflow question points at no one. How long does it now take from a request to a finished output? How many revision cycles happen before approval? What share of outputs come back for rework? Has quality at the next handoff gone up or down? None of these questions singles out a person, and all of them show whether AI has changed the work (Figure 1).

Take the everyday tasks measured in the DWP evaluation: writing emails, summarizing information, searching for research and transcribing meetings. Asking each person how many minutes AI saved on these tasks produces the kind of self-reported figure the trials already have. Asking how the workflows those tasks feed have changed would produce evidence that does not depend on anyone's willingness to speak up. Useful measures would include the time to answer a query, the time to prepare a briefing and the rework rate on drafted correspondence. These examples illustrate the method. They are not findings from the evaluation.

Workflow measurement also deals with the jagged frontier. If AI degrades quality

on some tasks, rework and downstream error rates will show it, while surveys of hours saved will not. A workflow lens reports the gain and the damage in the same place.

This work suits the PMO, the transformation office or the AI center of excellence. These groups already watch how work moves across functions, and they can do so without turning the data into a verdict on individuals.

	Individual lens *measures the person*	Workflow lens *measures the process*
Question asked	How many hours did each person save with AI?	How has the work itself changed?
Typical measures	Self-reported hours saved; license activity per person	Cycle time, revision cycles, rework rate, quality at the next handoff
Who the number points at	An individual	**A process**
What reporting risks for the employee	More work, tighter targets, looking replaceable	Little: no one is singled out
What leaders end up seeing	Only what people choose to report	**Where AI improves the work and where it degrades it**

Figure 1. Two ways to measure AI productivity. Source: the authors.

2. Put a firewall between AI data and performance appraisal

Workflow metrics stay safe only if people believe the data will not be turned back on them. That belief needs a rule, written down and published before any data is collected. The rule should say that AI workflow data is reported at team or process level and never below a minimum group size, that it is not used in individual appraisal, ranking or staffing decisions, and who has access to it.

The same logic tells leaders what to stop doing. Individual AI usage leaderboards, license activity used as a proxy for personal productivity, and appraisal forms that ask people how many hours AI saved them all attach the number to a person. Each one makes concealment the rational choice.

The trials suggest why this matters. Their surveys evaluated a tool rather than the people using it, and people answered. A firewall keeps that condition in place after the trial ends and AI becomes part of normal work.

3. Write the reinvestment contract before asking for disclosure

Measurement is only half the design. People reveal gains when they trust what happens next, and that trust has to exist before the disclosure. A reinvestment contract makes the trade explicit. Before anyone is asked to report gains, leaders state in the open how recovered capacity will be used, and they name the use precisely enough that the team can later check whether the promise was kept (Figure 2).

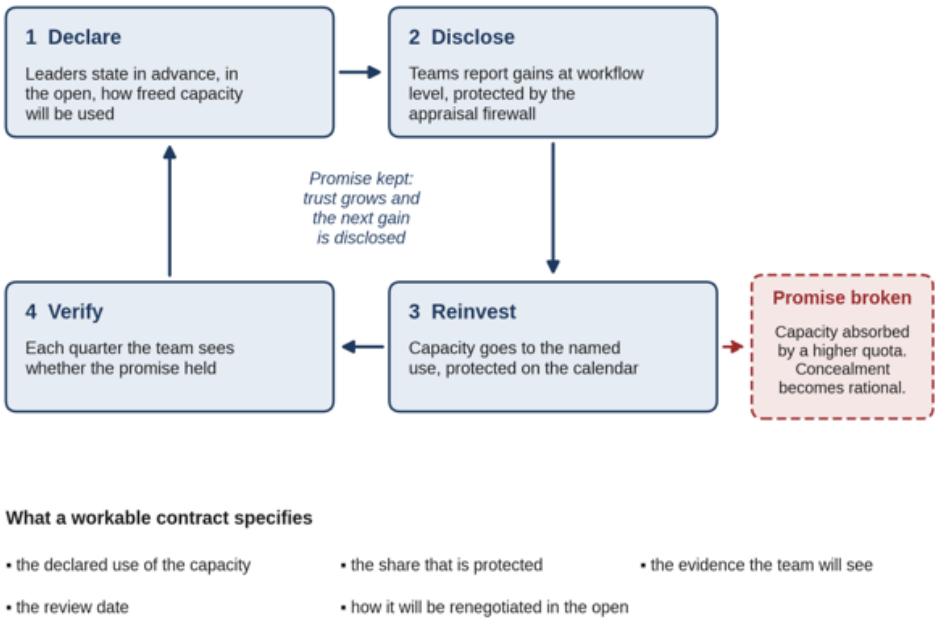


Figure 2. The reinvestment contract. Source: the authors.

What makes such a contract credible is the cost of breaking it, more than the wording of the commitment. A vague pledge that saved time "will support higher-value work" reassures no one, because it can be dropped without anyone noticing. A specific and visible commitment is much harder to drop. In practice, a workable contract has five elements: the declared use of the freed capacity, the share of that capacity that is protected (for example, on the calendar), the evidence the team will see, the date of the review, and how the contract will be renegotiated in the open if circumstances change.

Consider a hypothetical team whose members each save about 19 minutes a day, the average in the DWP evaluation. That is roughly an hour and a half per person per week. Leadership declares that this time will go to named work the team can verify, for example deeper review of complex cases and earlier legal checks, protected on the calendar instead of absorbed by a higher volume target. A quarter

later, the team can see whether the promise held. If it did, disclosure becomes the rational choice. If the capacity simply disappeared into a higher quota, the team learns the opposite lesson, and it will not need to learn it twice.

The DWP finding that 89 percent of users reallocated saved time to other work shows why this step matters. Reallocation already happens, but privately, as a series of individual choices the organization cannot see. The contract turns that private reallocation into a declared one.

4. Trace the gain to strategy through benefits realization

Minutes saved are an input. The benefit is what the organization does with them. The last part of the framework connects the task to the workflow and the workflow to a strategic outcome, which is the discipline of benefits realization. For each workflow in scope, the program should record in its benefits register the strategic objective that the reinvested capacity serves, the metric that will show progress and the executive who owns the result.

This raises the stakes of AI measurement. The question is no longer how many minutes were saved. It is whether the capacity AI releases is helping the strategy land. The PMO or transformation office is the natural owner of this link, because it already connects projects to strategy and can keep individual data out of the picture.

Putting the framework to work: the first 90 days

The framework needs no new technology. It needs a small number of management decisions, taken in the right order. A practical sequence for an executive sponsor looks like this:

- **Days 1 to 15.** Choose two or three workflows where AI is already in use, including unofficial use. Define three or four workflow metrics for each one, and record a baseline before anything changes. The DWP evaluation shows what is lost without one: with no measures taken before the tool arrived, its authors noted, it is difficult to assess changes over time with confidence.
- **Days 15 to 30.** Publish the firewall rule and the reinvestment contract, signed by the sponsor, before asking anyone to report gains.
- **Days 30 to 75.** Collect workflow data at team level. Where a view of AI use is needed, use anonymous or aggregated pulse surveys. Do not build individual usage rankings.
- **Days 75 to 90.** Review the results with the teams involved. Did the workflow

metrics move? Was the contract honored? What went into the benefits register? Decide whether to renew the contract, adjust it or extend it to more workflows.

At the end of the 90 days, a sponsor should be able to answer four questions. Which workflows changed, and by how much? Where did AI help, and where did it hurt? Where did the freed capacity go? And do the teams involved believe it is safe to report the next gain?

Limits of the approach

The framework has limits that leaders should weigh. It is built from public evidence and from practice, not from a controlled test of the framework itself. The natural next step is field research that compares units with and without disclosure-safe measurement.

Some roles need individual measurement, for example in safety-critical or regulated work. The framework does not remove that need. It asks that the reporting of AI gains be kept separate from it.

Workflow metrics can be gamed, like any metric. Pairing speed measures with quality measures, and reviewing the set of metrics every quarter, reduces that risk.

Finally, reinvestment contracts cost leaders some flexibility. That cost is intended. A promise that is free to break does not change anyone's behavior.

The harder part is human

The next phase of AI adoption will depend less on how capable the tools are than on how organizations handle disclosure. The companies that pull ahead will not be the ones that extract the most individual productivity claims. They will be the ones that make productivity safe to reveal before they try to make it easy to measure.

The hard part was supposed to be the technology. It is turning out to be the people: persuading those who hold the gains that telling the truth about them will be rewarded rather than punished. Get that wrong, and the dashboard keeps improving while the dividend stays hidden, held quietly and rationally by the people who earned it. Get it right, and the invisible dividend finally has a reason to come into view.

References

1. Government Digital Service. (2025, June 2). Microsoft 365 Copilot experiment: Cross-government findings report. GOV.UK. https://www.gov.uk/government/publications/microsoft-365-copilot-experiment-cross-government-findings-report

2. **Department for Work and Pensions.** (2026, January 29). An evaluation of DWP's Microsoft 365 Copilot trial. GOV.UK. https://www.gov.uk/government/publications/an-evaluation-of-dwps-microsoft-copilot-365-trial

3. **Brynjolfsson, E., Li, D., & Raymond, L. R.** (2023). Generative AI at work (NBER Working Paper No. 31161). National Bureau of Economic Research. https://www.nber.org/papers/w31161

4. **Dell'Acqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R.** (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013). https://www.hbs.edu/faculty/Pages/item.aspx?num=64700

5. **Microsoft & LinkedIn.** (2024, May 8). 2024 Work Trend Index annual report: AI at work is here. Now comes the hard part. https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part

6. **Goodhart, C. A. E.** (1975). Problems of monetary management: The U.K. experience. In Papers in monetary economics (Vol. 1). Reserve Bank of Australia, Sydney.

7. **Strathern, M.** (1997). 'Improving ratings': Audit in the British University system. European Review, 5(3), 305-321.