

Show the Return: The AI ROI Measurement Playbook

A WORKPLACEAI STRATEGY PLAYBOOK

From “We Think It Is Working” to a Number the Board Will Accept

Most organizations cannot say what their AI deployments are worth. The problem is rarely the technology. It is that nobody measured the starting point, tracked the full cost, or checked what the saved time became. Here is how to fix that.

WorkplaceAI.ai · Strategy Playbook · October 2026 · [Download PDF →](#)

Key Takeaways

- The evidence points more to weak measurement than to proven failure. Dataiku’s 2026 survey of 685 CIOs found 76% lack mature AI ROI measurement, yet adoption keeps rising.
- The first step happens before launch. Without a baseline of time, cost, error rate, and volume, no later number can show improvement. Capture the baseline first, even if it is rough.
- Usage is not value. Logins, prompts, and drafts show activity. Value shows up as time returned to the business, lower cost, higher quality, faster cycles, or revenue, and only when someone checks.
- Saved time is not automatically useful time. BCG found 66% of employees saving time with AI had little or no guidance on what to do with it. Decide where the hours go.
- Count the whole bill. Model usage, platform fees, integration, review time, governance, and rework all belong in the cost side, attributed to the team that incurs them.
- Grade every figure by its evidence: Measured, estimated, or assumed. A board can accept honest uncertainty, but it cannot accept a number that falls apart under one question.

- Decide in advance what happens at each result: Scale it, fix it, or stop it. A measurement that cannot change a decision is a report, not a control.

Contents

- | | |
|--|--|
| 1. The Measurement Gap | 10. Stage 7: Calculate Net Value, and Grade the Evidence |
| 2. The Principle: Baseline, Attribute, Net Out | 11. Stage 8: Report and Decide |
| 3. The Measurement Stack | 12. Measuring Agents |
| 4. Stage 1: Set the Baseline Before Launch | 13. An Illustrative Walkthrough |
| 5. Stage 2: Choose Metrics That Mean Something | 14. Where This Breaks |
| 6. Stage 3: Attribute the Change | 15. The Pre-Launch Scorecard |
| 7. Stage 4: Follow the Saved Time | 16. A 90-Day Roadmap |
| 8. Stage 5: Count the Full Cost | 17. Conclusion: Make the Return Visible |
| 9. Stage 6: Net Out Quality and Rework | |

Every AI deployment arrives with a promise: Time saved, cost cut, quality raised, revenue grown. A year later, many leaders cannot say which of those happened. They know the tool is in use. They are less sure it is paying for itself.

This playbook is about closing that gap. It does not require a data science team or a new platform. It requires a few habits applied in the right order, starting before the deployment goes live.

The Measurement Gap

Several 2026 surveys describe the same pattern from different angles. Adoption is wide. Proof is thin.

Source	What it found
Dataiku and The Harris Poll, 685 CIOs, July 2026	76% say their organization lacks mature ROI measurement across all or most AI initiatives. 72% cannot consistently confirm that all their agents deliver business outcomes. Only 21% have full, near-real-time visibility into AI costs by team or use case.
PwC Global CEO Survey, 4,454 CEOs	56% report no significant financial benefit from AI to date. 12% report both cost and revenue benefits, and 33% report one or the other.
McKinsey State of AI 2026, 1,719 respondents	37% attribute any EBIT impact to AI, essentially unchanged from last year. About 20% say AI operating costs, including tokens, constrain how much they use it.
Deloitte State of AI in the Enterprise 2026, 3,235 leaders	66% report productivity gains. 74% hope AI will grow revenue, but only 20% say it is doing so now.
BCG Global AI at Work 2026, nearly 12,000 employees	Of employees saving time with AI, 66% report little or no guidance on what to do with it.
BetterUp Labs and Stanford Social Media Lab, via HBR, 2025	40% of surveyed desk workers had received AI-generated “workslop” in the past month, at a cost the researchers estimated at about \$186 per employee per month.

The sources differ in what they ask and who published them, and several come from vendors with an interest in the topic. Read them as direction, not as exact measures. The direction is consistent: Leaders are under pressure to show a return and often lack the means to do it. Dataiku found 97% report rising board pressure for measurable AI ROI, and 72% expect their AI budget to be cut or frozen if targets are missed. The full analysis is in [our report on the CIO survey](#).

Measurement does not guarantee a good result. It does make a good result provable, and a bad result catchable early enough to fix.

The Principle: Baseline, Attribute, Net Out

Three disciplines separate a measured deployment from an assumed one.

Baseline

Capture how the work performs before AI touches it: Time, cost, error rate, volume, and quality. Every later claim is a comparison with this.

Attribute

Separate the change AI caused from the change that would have happened anyway. Use a comparison group, a phased rollout, or at least a stated assumption.

Net Out

Subtract the full cost and the rework from the gross gain. A saving that creates a larger problem downstream is not a saving.

Skip any one of the three and the number is unreliable. Without a baseline there is nothing to compare. Without attribution, every improvement is credited to AI. Without netting out, the gain is overstated by whatever the review, correction, and running costs consume.

The Measurement Stack

Baseline → Choose Metrics → Attribute → Follow the Time → Count Cost → Net Out Quality → Grade and Report → Decide

Baseline and Metrics

Fix the starting point and pick measures that connect to outcomes, not activity.

Attribute and Follow the Time

Show the change is AI's, and track where freed hours go.

Cost and Quality

Tally the full bill and subtract rework and downstream errors.

Grade, Report, Decide

Label the evidence, report it plainly, and act on thresholds set in advance.

Apply the stages to one deployment first. The habits transfer, and a single measured example persuades more than a framework.

Stage 1: Set the Baseline Before Launch

The most common measurement failure is also the cheapest to prevent. Teams launch, then try to work out later what changed. By then the original state is a memory, and memory flatters the new system or the old one depending on who is asked.

What to capture

A USABLE BASELINE COVERS

- **Volume:** How many units of work flow through per week or month, and how much it varies.
- **Time:** The hands-on minutes per unit, and the elapsed time from start to finish.
- **Cost:** Fully loaded labor, tools, and any outside spend per unit.
- **Quality:** The error, rework, and complaint rate, defined in a way a second person could reproduce.
- **Outcome:** The business result the work exists to produce, such as a resolved ticket, a closed deal, or a filled role.

Capture two to four weeks of data when you can, so ordinary swings are visible. If history already exists in your systems, pull it. If it does not, sample: Have a few people log their time on a week of real work. A rough baseline with its limits stated beats none.

Baseline by function

Function	Baseline measures worth capturing
Customer support	Handle time, first-contact resolution, reopen rate, cost per resolved ticket, customer satisfaction.
Finance	Cost per invoice, close duration, error and duplicate rate, exception volume.
Marketing	Hours per asset, cycle time from brief to publish, revision rounds, performance of comparable past assets.
Sales	Time on research and drafting per account, response time to leads, conversion at each stage.
HR and recruiting	Time to shortlist, time to hire, recruiter hours per hire, offer acceptance.
Engineering	Cycle time, review time, defect rate, rework after release.

Legal and contracts	Review time per contract, turnaround, issues found late.
---------------------	--

Write the baseline down with its date and method, and store it where the result will be reported. The cost calculators in [Gizmos and Gadgets](#), including [Support Ticket Cost](#) and [AP Cost](#), can help structure the baseline cost per unit. Stage 7 explains how to use the [Automation ROI Estimator](#) alongside this playbook.

Stage 2: Choose Metrics That Mean Something

Metrics fall into three tiers, and most dashboards stop at the first.

Tier	Examples	What it tells you
Activity	Active users, prompts, documents processed, agents deployed	Whether the tool is used. It says nothing about value.
Efficiency	Minutes per task, cost per unit, touchless rate, cycle time	Whether the work got cheaper or faster.
Outcome	Revenue, margin, retention, error cost, risk avoided, customer or employee results	Whether the business is better off.

Aim for an efficiency measure and an outcome measure for every deployment, and treat activity as a diagnostic. High usage with flat outcomes points to a workflow problem. Low usage with strong outcomes points to a rollout problem.

Deloitte’s finding that 66% of leaders report productivity gains while only 20% see revenue growth shows why the tiers matter. Efficiency is the easier measure to improve and the easier to claim. Outcome measures are harder, and they are the ones that persuade a board.

Pair every metric with a guardrail

A single metric invites gaming. Faster handle time can mean worse answers. Higher touchless rate can mean more silent errors. Pair each target with a measure that would reveal the damage.

Target metric	Pair it with
Handle time down	Reopen rate and customer satisfaction
Touchless rate up	Error rate on touchless items, sampled weekly
Content volume up	Edit time per asset and performance per asset
Cost per ticket down	Escalation rate and resolution quality
Time to hire down	Quality of hire and early attrition

Stage 3: Attribute the Change

Business results move for many reasons. A support team can improve because of a new AI tool, a seasonal dip in volume, a product fix, or a new manager. Crediting AI with all of it is the easiest claim and the first one a skeptical finance leader will challenge.

Four ways to isolate the effect, from strongest to weakest

CHOOSE THE STRONGEST METHOD YOU CAN RUN

- **Comparison group:** One team uses the tool and a similar team does not, over the same period. Compare the change in each.
- **Phased rollout:** Start teams on different dates, so each group serves as a comparison for the others while they wait.
- **Holdout sample:** Route a small random share of work through the old process, and compare outcomes on matched items.
- **Before and after with adjustments:** Compare periods, and list the other changes that happened in between. This is the weakest method, and it should be labeled as such.

A phased rollout is usually the most practical. It matches how organizations already deploy, and it creates a comparison for free. Decide the order of teams with measurement in mind, and avoid giving the tool first to the most enthusiastic or best-resourced group, since that skews the result.

State your assumptions

When a clean comparison is not possible, say so, and write down what you assumed. “We attribute 60% of the cycle-time improvement to the tool, because a second process change happened in the same quarter” is more credible than a claim of 100% with no reasoning. The attribution share becomes an input that others can challenge.

Stage 4: Follow the Saved Time

Time saved is the most common AI benefit claim, and the most often left unconverted. A person who saves an hour a day has not produced an hour of value. They have produced an hour of capacity. What the hour becomes decides the return.

BCG’s 2026 survey of nearly 12,000 employees found 66% of those saving time had little or no guidance on what to do with it. BCG also reports that an explicit strategy for redirecting that time lifted measured business impact by 25 percentage points, compared with roughly 5 points from better tools alone. Our [analysis of AI failure statistics](#) covers the study in more detail.

Keep a time ledger

For each deployment, record where freed hours go. There are only a few destinations, and they have very different values.

Where the hours go	Value	How to measure it
Absorbed by more volume of the same work	Real, if volume was a constraint	Throughput per person before and after
Redirected to higher-value work	Real, if the new work is defined and checked	Output from the named new work
Converted to avoided cost, such as fewer hires or less overtime	Real, if the cost actually drops	Headcount, contractor, and overtime records
Spent on checking and fixing AI output	Offsets the gain	Review and rework time, sampled
Scattered across ordinary tasks	Not measurable, so not claimable	Do not count it

Decide before launch where the freed time should go, and give managers the instruction. A team that saves ten hours a week and spends them on a named backlog produces something a report can show. A team that saves ten hours and lets them dissolve produces nothing that survives an audit.

Stage 5: Count the Full Cost

The cost side is where many ROI claims are weakest. The license or the usage bill is visible. The rest is spread across budgets and easy to miss.

Everything that belongs on the bill

INCLUDE EACH OF THESE

- **Model usage:** The per-use charges for tokens, calls, or tasks, which scale with volume and complexity.
- **Platform and tools:** Subscriptions, automation platforms, and the services around them.
- **Integration and build:** The labor to connect systems, write prompts and rules, and keep them current.
- **Review and oversight:** The time people spend checking, approving, and correcting.
- **Governance and security:** Access control, logging, testing, and the staff who run them.
- **Training and change:** The effort to bring users along, including the dip while they learn.
- **Rework and failure:** The cost of errors that reach customers, clients, or the ledger.

Agent-based systems deserve particular care, because their usage costs can run well above the first estimate once they loop, retry, and call multiple models. Our piece on [agentic AI consumption costs](#) covers why. McKinsey's 2026 survey found about 20% of respondents say AI operating costs already constrain how much they use AI.

Attribute cost to the team that incurs it

Dataiku found only 21% of CIOs have full, near-real-time visibility into AI costs by business unit, team, or use case. Without that, no team can calculate its own return, and central IT ends up carrying a bill nobody owns.

- Tag every model key, workflow, and agent to a team and a use case at creation, not afterward.

- Report cost per unit of work, such as cost per ticket resolved, alongside total spend.
- Set a budget and an alert for each deployment, so a runaway loop shows up in hours.
- Review the cost report monthly with the team that owns the deployment.

The [Build vs. Buy calculator](#) can help compare the cost of a custom build with a ready-made option before you commit.

Stage 6: Net Out Quality and Rework

A faster process that produces worse output can cost more than the slow one. The research on AI-generated “workslop” describes the pattern: Content that looks finished but lacks substance moves the work to the next person. BetterUp Labs and the Stanford Social Media Lab, in research published in Harvard Business Review, found 40% of surveyed desk workers had received it in the past month. The researchers estimated about two hours of lost time per instance and about \$186 per employee per month, calling that a back-of-the-envelope figure, and the data was self-reported.

The point for measurement is simple. Gross time saved is not the number to report. Net time saved is.

$$\text{Net gain} = \text{Gross time saved} - \text{Review time} - \text{Rework time} - \text{Downstream cost of errors}$$

How to measure quality without a large program

- **Sample and score:** Each week, pull a random set of outputs and have a second person rate them against a short rubric.
- **Track the correction rate:** How often a human changes the AI output before it is used, and by how much.
- **Follow errors downstream:** Tag complaints, reopened tickets, ledger corrections, and client escalations back to whether AI touched the work.
- **Watch the receiving end:** Ask the people who receive AI-assisted work whether it saves them time or costs them time.

The last item is the one most often skipped. The person who produced the output sees a saving. The person who receives it may see a cost. Measure both ends.

Stage 7: Calculate Net Value, and Grade the Evidence

With the baseline, attribution, time ledger, cost, and quality in hand, the calculation is straightforward.

Net value = Attributed gain (time converted to value + cost avoided + revenue added) – Full cost – Rework and error cost

Report it as a range, not a single point. Show a conservative figure, an expected figure, and the assumptions behind each. Then label every input by the quality of its evidence.

Evidence grades

Grade	Meaning	Example
Measured	Observed directly in your own data, with a baseline and a comparison	Cost per ticket fell from a recorded baseline in the pilot team versus a comparison team
Estimated	Derived from samples or partial data, with stated assumptions	Review time per item estimated from a two-week sample of ten reviewers
Assumed	Taken from a vendor claim, a benchmark, or a judgment, and not yet tested in your environment	Expected savings per hire based on a published case study

The grades do three useful things. They show leaders how much of a business case rests on proof. They show the team where to invest in measurement next. And they protect credibility, because a figure labeled as assumed cannot later be accused of overstating.

Treat vendor-supplied figures, including case studies and calculators, as assumed until tested. A claim that holds across other companies still has to hold in your workflow.

Before launch: Use the ROI Estimator as a screen

The [Automation ROI Estimator](#) is a pre-launch screen. It tells you whether a task looks worth automating on the assumptions you enter. This playbook covers what comes after, which is proving the result. Used together, the two give you an estimate before launch and a measurement after it. To keep them consistent, enter the inputs like this.

- **Set the reduction after review.** The “how much of it automation removes” input should be net of the time people spend checking and fixing the output.
- **Enter the full one-time cost.** Include the build or purchase price plus setup, integration, and training.
- **Put oversight in the running cost.** Add review time to usage and platform fees in the monthly figure.
- **Remember that hours are capacity, not cash.** The estimator values every freed hour at your hourly rate. Confirm where the time will go, using the ledger in Stage 4, before treating the saving as a return.
- **Label the result as assumed.** After launch, replace each assumption with a measured number and run the estimator again.

Rerunning the estimator with measured inputs after a quarter is a quick way to see whether the original case held.

Stage 8: Report and Decide

A measurement exists to change a decision. Define the decision rules before the results arrive, so the data cannot be argued around afterward.

Set thresholds in advance

Result at review	Decision
Net value positive, evidence measured, quality stable	Scale to the next team or process.
Net value positive but evidence mostly assumed	Keep running, and invest in measurement before scaling.
Gross gain positive, net value negative after rework and cost	Fix the workflow, the prompts, or the review step, then re-measure.
Net value flat or negative after a fair trial, with the fixes attempted	Stop, and record what was learned.

The one-page scorecard

REPORT EACH DEPLOYMENT ON ONE PAGE

- **What it does:** One sentence, and the named owner.
- **Baseline and current:** The two or three key measures, side by side, with dates.
- **Net value:** The range, with the share that is measured, estimated, and assumed.
- **Full cost:** Total and per unit, against budget.
- **Quality:** Sample score, correction rate, and any downstream incidents.
- **Decision:** Scale, fix, hold, or stop, with the reason.

Boards and CEOs are asking harder questions. In Dataiku's survey, 87% of CIOs said their CEO has told them their job security depends on AI outcomes. A plain scorecard that shows what is proven, what is estimated, and what is still a bet gives leaders something they can defend upward. It is also easier to defend than a confident number that a single question can break.

Where you cannot yet measure something, say so, and say how and when you will.

Admitting a gap with a plan is stronger than hiding it.

Measuring Agents

Agents add two difficulties. They act on their own, so the output cannot always be sampled by looking at a document. And they multiply quickly. Dataiku found 67% of CIOs estimate 51 or more agents running in production, and 47% decommissioned more than 20 this year.

Measurement has to cover each agent, not just the program.

Owner

A named person accountable for the agent's results and its failures. See the [Agent Governance Playbook](#).

Task Success

Share of runs that completed the intended outcome, checked by sample or by a downstream result.

Cost per Run

Model, tool, and platform cost per completed task, with an alert on spikes.

Escalation and Override

How often a person stepped in, and why. Rising overrides signal drift or a poor fit.

Incident Record

Unintended actions, near misses, and corrections, logged with dates and owners.

Retirement Test

A rule for decommissioning an agent whose value no longer covers its cost or risk.

Dataiku also found 72% of CIOs cannot consistently confirm that all their agents deliver business outcomes. The six items above are the minimum record that makes that confirmation possible. An agent without an owner, a cost line, and a success measure is a liability that happens to be running.

An Illustrative Walkthrough

The following example is illustrative, with invented figures. It is not a case study. It shows how the stages fit together for one deployment.

A SUPPORT TEAM ADDS AI-DRAFTED REPLIES

- **Baseline (4 weeks before launch):** The team handles an average of 1,000 tickets a month. Hands-on time is 12 minutes per ticket, the reopen rate is 8%, and fully loaded cost is \$9.60 per ticket. These figures are measured from the ticketing system.
- **Rollout design:** Two of four sub-teams start in month one, and two start in month three. The later two serve as the comparison group. Another change, a new help article, lands in month two and is noted.
- **Month two results (measured):** Hands-on time falls to 9 minutes in the pilot teams and stays near 12 in the comparison teams. The reopen rate rises to 10% in the pilot teams.
- **Cost (full):** Model usage, platform, and 6 hours of weekly review add \$1.10 per ticket. Setup and training are amortized over twelve months.
- **Net out:** A 3-minute saving is worth about \$2.40 per ticket. After \$1.10 in added cost and the cost of the 2-point reopen increase, the net is positive but smaller than the headline. The team flags the reopen rise as the item to fix.
- **Time ledger:** The freed time goes to a named backlog of knowledge-base articles. The output of that work is tracked.
- **Grade and decision:** Time and cost are measured. The value of the backlog work is estimated. Satisfaction impact is assumed until two more months of data arrive. Decision: Fix the reopen issue, then scale to the later two teams, and re-measure.

Notice what made this work. The baseline existed before launch. The rollout created a comparison. The cost side counted review time. The reopen rate caught a quality problem the time savings alone would have hidden. And the decision followed rules, not enthusiasm.

Where This Breaks

Failure	What you see	Fix
No baseline	Improvement is claimed but cannot be shown	Capture a rough baseline before launch, and date it
Crediting AI with everything	The gain matches every other change that happened	Use a comparison group or phased rollout, and state the attribution share

Counting activity as value	Usage dashboards rise while outcomes stay flat	Pair activity with an efficiency and an outcome measure
Saved time that vanishes	Hours freed but no visible output	Assign the time to named work, and keep a time ledger
Hidden costs	The ROI looks strong until review, rework, and platform costs are added	Use the full cost list, tagged to the owning team
Gaming a single metric	A target is met while quality falls	Pair every target with a guardrail measure
Vendor numbers treated as proof	A business case rests on claims never tested locally	Mark vendor figures as assumed until measured
Measurement that changes nothing	Reports pile up and decisions follow opinion	Set scale, fix, and stop thresholds before the results arrive
Agent sprawl	Agents run with no owner, cost line, or success measure	Register every agent with the six-item record

The Pre-Launch Scorecard

Before approving any AI deployment, check that these are in place. A deployment that cannot answer most of them is not ready to be measured, and therefore not ready to be judged.

ANSWER EACH BEFORE LAUNCH

- A named owner for the deployment and for its results.
- A written baseline for volume, time, cost, quality, and the outcome measure, with dates and method.
- One efficiency metric, one outcome metric, and a guardrail for each.
- A comparison method: Group, phased rollout, holdout, or a stated assumption.
- A decision on where freed time should go, and who will track it.
- A cost list covering usage, platform, integration, review, governance, and training, tagged to the team.
- A quality check: Sample size, rubric, and who scores it.
- Decision thresholds for scale, fix, and stop.
- A review date, and the person who will present the scorecard.

A 90-Day Roadmap

DAYS 1 TO 30: BASELINE AND DESIGN

- Choose one live or planned deployment, and name its owner.
- Capture the baseline, even roughly, and set the metrics and guardrails.
- Pick the comparison method and the decision thresholds.
- Tag costs to the owning team from day one.

DAYS 31 TO 60: MEASURE AND SAMPLE

- Run the deployment while logging time, cost, and quality samples weekly.
- Keep the time ledger, and confirm freed hours are going to the named work.
- Review the first results with the owner, and note what is measured and what is assumed.

DAYS 61 TO 90: NET OUT AND DECIDE

- Calculate net value as a range, with evidence grades.
- Produce the one-page scorecard and present it with a recommendation.
- Apply the decision rule, and record the outcome.
- Reuse the template for the next deployment, and start retrofitting baselines for existing ones where you can.

Retrofitting is possible but imperfect. For deployments already running, rebuild a baseline from old records, from a comparison team, or from a short sample of the manual process, and mark the result as estimated.

Conclusion: Make the Return Visible

The surveys do not show that AI has no value. They show that many organizations cannot demonstrate it, and that the people responsible feel the pressure. Weak measurement is the problem that can be fixed fastest, and it makes every other AI decision better.

Baseline → Choose Metrics → Attribute → Follow the Time → Count Cost → Net Out Quality → Grade and Report → Decide

Start with one deployment, capture the baseline before anything else, count the whole cost, and grade your own evidence. The goal is a number you can defend when someone asks how you know.

If you cannot show the baseline, you cannot show the gain.

Where to Go From Here

The Automation ROI Estimator and the cost calculators help set the starting numbers. The Agent Governance Playbook covers ownership and controls for agents, and our CIO survey report covers why measurement has become a board-level issue.

[Automation ROI Estimator →](#)

[Agent Governance Playbook →](#)

[CIO Survey Report →](#)

[Agentic AI Cost Reality →](#)

[Build vs. Buy Calculator →](#)

Sources: Dataiku, *Global AI Confessions Report: CIO Edition 2026* (The Harris Poll, 685 CIOs, fielded July 9 to 29, 2026). PwC, *2026 Global CEO Survey* (4,454 CEOs). McKinsey, *The State of AI in 2026* (1,719 respondents, published August 25, 2026). Deloitte, *State of AI in the Enterprise 2026* (3,235 leaders), as reported by Unite.AI. BCG, *Global AI at Work 2026*, as covered in WorkplaceAI's earlier analysis. BetterUp Labs and Stanford Social Media Lab workslop research, published in Harvard Business Review in September 2025, as reported by Axios and BetterUp. Several of these surveys were published by vendors or consultancies with an interest in the topic; every figure is attributed to its original publisher, and none is a WorkplaceAI study. The walkthrough uses invented figures for illustration only.

Unvarnished Reviews

© 2026 WorkplaceAI.ai · AI Automation for Every Key Business Function