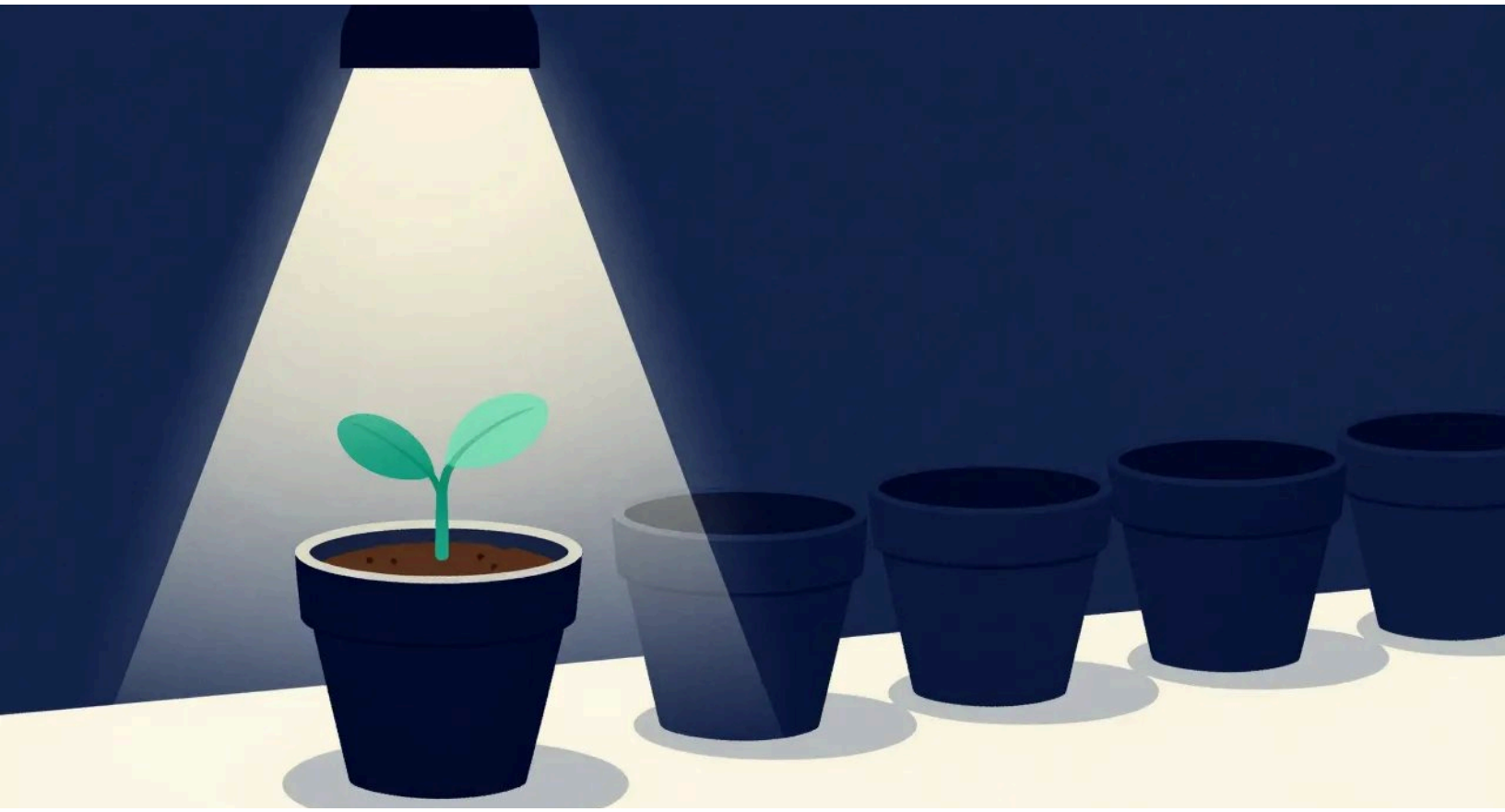


EXECUTIVE PAPER · AI

Prove One AI Use Case in Ninety Days

When the board asks for the company's AI plan, the strongest answer is one funded use case with a measured baseline, a deadline, and a stated condition for stopping it. Broad programs and pilot portfolios mostly produce activity.

AUTHOR

Chris Watkins

SERVICE

AI

FOR

CEO, CFO, Board · 12 min read

AT A GLANCE

The problem. Nearly every company now uses AI somewhere, yet in McKinsey's 2026 survey only 37% of respondents reported any contribution to EBIT, and about 6% attributed 5% or more of EBIT to it. MIT researchers found in 2025 that about 95% of enterprise generative AI pilots showed no measurable profit and loss impact.

Why it persists. Boards ask for a plan, and management answers with a program: a steering committee, a vendor platform, and a dozen pilots chosen by enthusiasm. Few of them start with a measured baseline, so none can prove a result, and none has a date on which it must stop.

The recommendation. Answer the board with one charter. Pick one high-volume process with a single accountable owner and an existing metric. Measure the baseline first, set a target, fund 90 days, compare against a control group, and write down in advance the result that ends the project. Scale what passes. Stop what fails, and report both to the board.

The Decision in Front of You

The question usually arrives as "what is our AI plan?" and it usually arrives from the board. Management usually reaches for an enterprise AI program, with a platform purchase and a center of excellence, or for a portfolio of departmental pilots to see which ones take. A third option is to fund one narrow use case, prove or disprove its value against a baseline within 90 days, and decide what comes next on that evidence.

This paper argues for the third answer. Gartner put the cost of generative AI deployment approaches at $5 million to $20 million in 2024 [6]. The larger cost is time: a year spent on a program that cannot show a result leaves the board where it started, with less patience.

BCG found in 2024 that the companies getting value from AI pursued about half as many opportunities as their peers and expected more than twice the return [5].

What's Happening

Adoption is near-universal, returns are concentrated

McKinsey's 2026 State of AI survey found that 37% of respondents attribute at least some EBIT impact to AI, about the same as the year before [2]. Only about 6% qualify as high performers, defined as attributing at least 5% of EBIT to AI, a share that has not moved since 2025 [2]. The share reporting enterprise-wide scaling rose from 38% to 44% over the same period [2]. McKinsey's related 2026 analysis found that the high performers differ mainly in practice: they redesign workflows, track KPIs for their AI solutions and have visibly committed senior leaders [16].

MIT's Project NANDA reported in 2025, from 150 leader interviews, a survey of 350 employees and a review of 300 public deployments, that about 95% of enterprise generative AI pilots produced little or no measurable P&L impact [1]. The report found more than half of generative AI budgets going to sales and marketing tools, while the largest returns came from back-office automation such as reducing outsourcing and agency spend [1]. Its lead author described the successful pattern as picking one pain point and executing well [1].

BCG's research reads the same way across three years. In 2024, 74% of 1,000 surveyed companies had yet to show tangible value from AI [5]. In 2025, 60% of 1,250 companies reported minimal revenue and cost gains despite investment [3]. By September 2026, BCG reported that nearly half of 1,330 surveyed companies were creating real value, and attributed the difference to leaders choosing where AI creates value and redesigning those workflows end to end [4].

Analysts expect many projects to be abandoned

Gartner predicted in July 2024 that at least 30% of generative AI projects would be abandoned after proof of concept by the end of 2025, citing poor data quality, inadequate risk controls, escalating costs or unclear business value [6]. In June 2025 it predicted that more than 40% of agentic AI projects would be canceled by the end of 2027, for escalating costs, unclear business value or inadequate risk controls [7]. Unclear business value appears in both forecasts. That is a scoping failure, settled before any model is chosen.

RAND interviewed 65 data scientists and engineers in 2024 and estimated that more than 80% of AI projects fail, twice the rate of IT projects without AI [8]. The first root cause on its list is that stakeholders misunderstand or miscommunicate the problem the AI is meant to solve [8].

Measured results differ from felt results

Two controlled studies show why a baseline matters. A study of 5,179 customer support agents at a software company, published in the Quarterly Journal of Economics in 2025, found an AI assistant raised issues resolved per hour by 14% on average and by 34% for novice and low-skilled agents [9]. That was a narrow tool in one process, measured against a control.

METR ran a randomized trial in 2025 with 16 experienced open-source developers on 246 real tasks [10]. The developers expected AI tools to make them 24% faster. They took 19% longer. After the study they still believed the tools had sped them up by 20% [10]. The authors caution that the result reflects early-2025 tools in one setting [10]. The lesson for executives holds anyway: people using AI cannot reliably report its effect on their own work. Only a measured baseline settles it.

Three cases from 2024 and 2025

Klarna, 2024 to 2026. In February 2024 Klarna said its AI assistant was handling two-thirds of customer service chats, doing the work of 700 agents, cutting resolution time from 11 minutes to 2, and would improve 2024 profit by an estimated $40 million [11]. The metrics were speed and cost. In May 2025 the

chief executive told Bloomberg the AI-focused path had been wrong and that customers should always be able to reach a human, and Klarna began recruiting human agents again [12]. In 2026 he said AI makes service "the cheap customer service" and that the industry had focused too much on cost [13]. The lesson: the target must include the quality measure customers care about, set before launch, alongside the cost measure.

Commonwealth Bank of Australia, 2025. In July 2025 Australia's largest bank told staff that 45 customer service roles would go, replaced by an AI voice bot on its inbound line [14]. The Finance Sector Union said call volumes were rising and that managers were offering overtime and asking team leaders to answer calls [14]. In August the bank reversed the decision, said its initial assessment was wrong and the roles were not redundant, and apologized to the employees [14]. A headcount decision was made before the measurement supported it. Done in the other order, the reversal and the public apology would not have been needed.

McDonald's, 2024. McDonald's ran an AI voice ordering test with IBM at about 100 drive-thrus from October 2021 [15]. It had set a target of better than 95% order accuracy before broader adoption [15]. In June 2024 it announced the test would end [15]. A written accuracy target gave the company a clear basis for ending the test, though it took two and a half years to reach the decision.

SIDEBAR: WHY ONE, AND WHY NINETY DAYS

One use case means one owner, one metric and one decision the board can follow. A portfolio of pilots spreads scarce engineers, data owners and executive attention across projects that each stall at the same point. Ninety days is long enough to build or buy a narrow tool and collect four to six weeks of measured results. It is short enough that the board hears an answer within one or two meetings. Some use cases need longer, and the charter should say so up front.

Why the Usual Approach Fails

The broad program buys capability before it has a problem. A platform license and a center of excellence make a visible start without a number that changes on the P&L. A year later the board asks what it got, and the honest answer is usage statistics.

The pilot portfolio fails more quietly. Each department nominates a pilot. Each pilot is small enough to approve without scrutiny, so none has a baseline or a stopping rule. Results come back as anecdotes from enthusiastic users, and METR's data shows how unreliable those are [10]. Nothing gets killed, because nothing was ever defined as failure. MIT's finding that most pilots stall describes this pattern [1].

Both approaches skip the same step: writing down, before spending, what the process does today, what it should do after, and what result would end the work. RAND's first root cause is the absence of that step [8].

The Framework: The One-Bet Charter

The One-Bet Charter is a single page the board approves. It has six fields, and the project cannot start until each is filled with a number or a name.

Exhibit 1. The One-Bet Charter: six fields, each filled before money is spent

Field	What it holds	Example
Process and owner	One process, with one business owner accountable for its P&L line	Inbound warranty claims; VP of customer operations
Baseline	The current metric, measured for at least two weeks before work starts	6.2 claims handled per agent hour; 4.1% rework rate
Target	The improvement that justifies scaling, in the same unit	12% more claims per hour with rework no higher than 4.1%
Guardrail	The quality, risk or customer measure that must not get worse	Customer satisfaction and complaint rate at or above baseline
Deadline	The date of the scale-or-stop decision	Day 90
Kill criterion	The result that stops the project, written before launch	Less than 6% gain, or any guardrail worse than baseline

Source: author's framework.

Good candidates share four traits. The process runs at high volume, so a few weeks produce enough data. It already has a metric someone reports today. It has a single owner who can change the workflow. And a wrong answer is cheap to catch, as with a draft a person reviews before it goes out. Back-office work such as claims handling, invoice exceptions, contract review and internal service desks often fits, which matches where MIT found the returns [1].

Measure against a control. Give the tool to some teams, shifts or queues and leave others alone, as the customer support study did [9]. Without a control, seasonal swings and the attention a pilot gets will be counted as AI gains. How to test the system's own accuracy is covered in a separate paper on evaluations; this charter measures the business result.

The 90 days divide into four stretches. Days 1 to 15 settle the charter and measure the baseline. Days 16 to 45 build or buy a narrow tool. Days 46 to 80 run it against a control group. Days 81 to 90 compare the result with the target, the guardrail and the kill criterion, and end in one of two outcomes: scale to the whole process, or stop and report why.

Buying is often faster than building for the first use case. MIT found that tools bought from specialized vendors and built with partners reached deployment about two-thirds of the time, and that internal builds succeeded only about a third as often [1]. The charter should require the buy option to be priced

before a build is approved.

What the decision is worth

The arithmetic below is the author's illustrative estimate. Assume a claims team of 40 people at a fully loaded $75,000 each, or $3.0 million a year. The 90-day project costs $250,000, and licenses and model usage run $60,000 a year after launch.

Exhibit 2. The kill criterion is set where payback stops making sense (author's illustrative estimate)

Measured gain	Annual capacity value	Net of $60,000 run cost	Payback on $250,000
14% (the support study's average [9])	$420,000	$360,000	8.3 months
About 6.2% (the 24-month payback line)	$185,000	$125,000	24 months
5%	$150,000	$90,000	33 months

Source: author's illustrative estimate. Capacity value = team cost × measured gain; it becomes cash only if the freed capacity avoids hiring, absorbs growth or reduces outsourcing. Payback = $250,000 ÷ net annual value × 12.

With these assumptions, a gain under about 6% does not pay back within two years, so the kill criterion sits there. Finance should run its own numbers and fix the threshold before results exist, when nobody has a stake in the answer.

Costs, Risks, and Tradeoffs

Exhibit 3. What the one-bet approach costs, and when it is wrong

Area	Cost or risk	Mitigation
Opportunity cost	Other departments wait while one use case runs	Keep a ranked backlog; the second charter starts the day the first decides
Measurement effort	Baselines and control groups take analyst time and delay launch by two weeks	Use metrics the business already reports; borrow an analyst from finance
Narrow results	A 14% gain in one team is small next to board expectations of transformation	Report the gain, the payback and the next three candidates with their own baselines
False negatives	A tool that needs months of user learning can fail a 90-day test [10]	State the learning period in the charter and extend once, by written decision
Local optimum	One process at a time can miss changes that cross departments	After two successes, charter a cross-process redesign with the same discipline
When this is wrong	Some bets cannot be measured in 90 days: a product feature for customers, a regulated model needing validation, or a competitor move that requires a platform decision now	Use a longer charter with interim milestones and the same kill criterion; make the platform decision on its own terms and say so to the board

Source: author's analysis.

How to Act

Exhibit 4. A one-year roadmap

Window	Actions	Owner	Milestone
First 30 days	List ten candidate processes with volume, metric and owner. Pick one. Write the charter. Start measuring the baseline. Price the buy option.	CEO with CFO; named business owner	Board approves a one-page charter with every field filled
First 90 days	Build or buy the narrow tool. Run it against a control group for at least four weeks. Compare to target, guardrail and kill criterion.	Business owner with technology lead	Scale or stop decision on day 90, reported to the board with the numbers
First 90 days	Keep a ranked backlog of the next candidates, each with a measured baseline.	Head of strategy or transformation	Second charter ready for approval on day 90
First 365 days	Scale the passing use case to the whole process. Run two or three more charters in sequence. Report each result against its baseline every quarter.	CFO with business owners	Each scaled use case shows its gain in the reported P&L line; every stopped project has a written reason

Source: author's recommendation.

Questions to Ask Your Team

1. Which single process would we choose first, and who owns its P&L line?
2. What is that process's metric today, and how long have we measured it?
3. What result on day 90 would make us stop, and have we written it down?
4. Which measure of quality or customer experience must not get worse?
5. Where is the control group, and how do we keep it clean?
6. Have we priced buying a tool before deciding to build one?
7. How many AI pilots are running now, and which of them has a baseline and an end date?
8. If this one succeeds, how does the gain show up in the numbers the board already sees?

Closing

The board's question deserves an answer it can check. Fund one narrow use case with a measured baseline, a target, a guardrail, a 90-day deadline and a kill criterion written before launch. Scale it if it passes and stop it if it fails. Report either result in money. Each charter after that starts from evidence the company owns, and the AI plan becomes a record of measured gains the board can audit.

Sources

[1] Fortune. "MIT report: 95% of generative AI pilots at companies are failing." Coverage of MIT Project NANDA, "The GenAI Divide: State of AI in Business 2025." 2025. https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/.

[2] FM Magazine. "Companies' financial value from AI holds firm in 2026." Coverage of McKinsey & Company, "The State of AI in 2026: On the Road to ROI." 2026. https://www.fm-magazine.com/news/2026/sep/companies-financial-value-from-ai-holds-firm-in-2026/.

[3] Boston Consulting Group. Press release on "The Widening AI Value Gap: Build for the Future 2025." 2025. https://www.bcg.com/press/30september2025-ai-leaders-outpace-laggards-revenue-growth-cost-savings.

[4] Boston Consulting Group. Press release on the Applied AI Index 2026: almost 50% of companies now generate value with AI. 2026. https://www.bcg.com/press/30september2026-ai-starting-to-pay-off-companies-generate-value.

[5] Boston Consulting Group. "AI Adoption in 2024: 74% of Companies Struggle to Achieve and Scale Value." Press release on "Where's the Value in AI?" 2024. https://bcg.com/ja-jp/press/24october2024-ai-adoption-in-2024-74-of-companies-struggle-to-achieve-and-scale-value.

[6] Gartner. "Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025." Press release. 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.

[7] The Star (Reuters). "Over 40% of agentic AI projects will be scrapped by 2027, Gartner says." Coverage of Gartner press release. 2025. https://www.thestar.com.my/tech/tech-news/2025/06/25/over-40-of-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says.

[8] Ryseff, J., De Bruhl, B. F., and Newberry, S. J. "The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed." RAND Corporation, RR-A2680-1. 2024. https://www.rand.org/pubs/research_reports/RRA2680-1.html.

[9] Brynjolfsson, E., Li, D., and Raymond, L. "Generative AI at Work." Quarterly Journal of Economics 140(2). NBER Working Paper 31161. 2025. https://www.nber.org/index.php/papers/w31161.

[10] METR. "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." 2025. https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/.

[11] CX Today. "Klarna Claims Its New AI Assistant Does the Work of 700 Full-Time Agents." 2024. https://www.cxtoday.com/contact-center/klarna-claims-its-new-ai-assistant-does-the-work-of-700-full-time-agents/.

[12] Tech.co. "Klarna Reverses AI Customer Service Overhaul." Coverage of Bloomberg interview. 2025. https://tech.co/news/klarna-reverses-ai-overhaul.

[13] CX Dive. Report on Klarna chief executive Sebastian Siemiatkowski's remarks on the 20VC podcast about an Uber-style customer service model. 2026. https://www.customerexperiencedive.com/news/klarna-pursues-uber-style-customer-service-model/812763/.

[14] Information Age (Australian Computer Society). Report on Commonwealth Bank reversing AI-linked customer service job cuts. 2025. https://ia.acs.org.au/article/2025/cba-reverses-ai-driven-job-cuts--admits--error-.html.

[15] Nation's Restaurant News. "McDonald's is ending its AI drive-thru test with IBM." 2024. https://www.nrn.com/quick-service/mcdonald-s-is-ending-its-ai-drive-thru-test-with-ibm.

[16] The Next Web. "McKinsey's new AI report argues the productivity payoff is real but conditional." Coverage of McKinsey & Company, "AI productivity gains and the performance paradox." 2026. https://thenextweb.com/news/mckinsey-ai-productivity-paradox-enterprise-roi-capex.

The views in this paper are the author's own and do not represent any employer. Scenarios are hypothetical. This paper is not legal, financial, or compliance advice.