TAIWHA / ENTERPRISE AI PORTFOLIO / PUBLIC REPORT

Why Enterprise AI Pilots Are Hard to Stop

2026 Enterprise AI Portfolios: Kill, Repair, and Bounded-Scale Outcomes and Management Implications

REPORT TYPE	LANGUAGE	SOURCE CUTOFF	VERSION
Public Flagship Report	English Public Edition	2026-07-14	v2.2 / 2026 Bilingual Public Edition

RESEARCH AND EDITORIAL INSTITUTION	EDITORIAL REVIEW
Taiwha Intelligence Institute	Public editorial and evidence-boundary review

USE BOUNDARY

This report uses public information available through 2026-07-14 and wholly fictional conditions for 6 pilots to illustrate 1 bounded Scale, 3 time-boxed Repairs, and 2 Kill decisions. The examples are not outcomes from a real company and do not replace local data or the necessary financial, legal, security, and compliance review.

OUTCOMES, EVIDENCE, AND ACTION

Contents

Start with management's conclusion and the 1 / 3 / 2 illustrative decisions for 6 fictional pilots, then check the evidence, three gates, dashboard, and 90-day actions.

EVIDENCE BOUNDARY

Survey measures are not collapsed into one adoption or success rate; task-level effects are not extrapolated directly to enterprise profit; the P1–P6 classifications use wholly fictional illustrative conditions and are not outcomes from a real company.

A RESULTS-FIRST READING PATH

- Begin with 1 Scale, 3 Repairs, and 2 Kill decisions, then read the project-level reasons for P1–P6.
- Next, check the survey denominators and task studies so adoption, ROI, EBIT impact, and local productivity are not treated as equivalent measures.
- Finish with the three gates, portfolio dashboard, and 90-day cadence to turn the illustration into testable local action.

Decision outcomes and management implications for the 2026 enterprise AI portfolio: Kill / Repair / Scale

If you oversee enterprise AI budgets, business workflows, technology delivery, or risk, and pilots are multiplying faster than integration, review, and change capacity, this report is for you. It addresses one portfolio decision: which projects can scale, which need time-boxed repair, and which should lose their current claim on scarce resources.

What management needs to decide

The report's central judgment is that many companies should examine one portfolio risk in 2026: pilots may be growing faster than comparable evidence and shared delivery capacity. When that mismatch appears, weak-evidence projects may continue to absorb integration, review, and change resources while stronger-evidence projects wait.

In the wholly fictional 6-project portfolio below, 1 project qualifies for bounded Scale, 3 can enter only time-boxed Repair, and 2 should lose their current priority claim on scarce resources. The distribution illustrates the decision criteria only; it is not a result from a real company, an industry benchmark, or a project recommendation.

6 fictional pilots: illustrative decisions

P1–P6 are illustrative identifiers. All evidence states, constraints, and decisions are fictional.

Pilot	Public outcome	Principal basis	Next boundary
P1 Customer-response drafting	Time-boxed Repair	Only a response-time change is observed. There is no credible comparator, and reopens, senior review, and incorrect advice are not fully recorded	Do not expand the current scope; add comparator and quality-burden evidence, and remove current priority if the named gap remains open at the deadline
P2 Pre-meeting sales research	Time-boxed Repair	Only self-reported time savings; CRM write remains unapproved	Keep read-only, confirm whether the business accepts the outcome, and do not expand system permissions
P3 Invoice-anomaly pre-check	Time-boxed Repair	Compared evidence exists, but false-negative severity remains unresolved	Keep as a read-only pre-check; grant no payment authority until false-negative risk is tested and controlled
P4 Internal knowledge retrieval	Bounded Scale	Comparable results repeat across at least 2 review windows; the named owner, read-only permission filter, and operating burden remain stable	Scale only over the filtered SOP corpus; continue monitoring source freshness, permission filtering, and exceptions
P5 Contract issue spotting	Kill — remove current priority claim	Evidence is only observed; missed issues have asymmetric impact and legal review is the current bottleneck	Release the current claim on scale capacity; retain testing and issue-taxonomy assets for reassessment if risk or review capacity changes
P6 Code-change agent	Kill — remove current priority claim	Change proposals increased, but production writes, rollback, and permission boundaries remain hard blockers	Do not expand into production; resubmit only after the control gaps close and new evidence exists

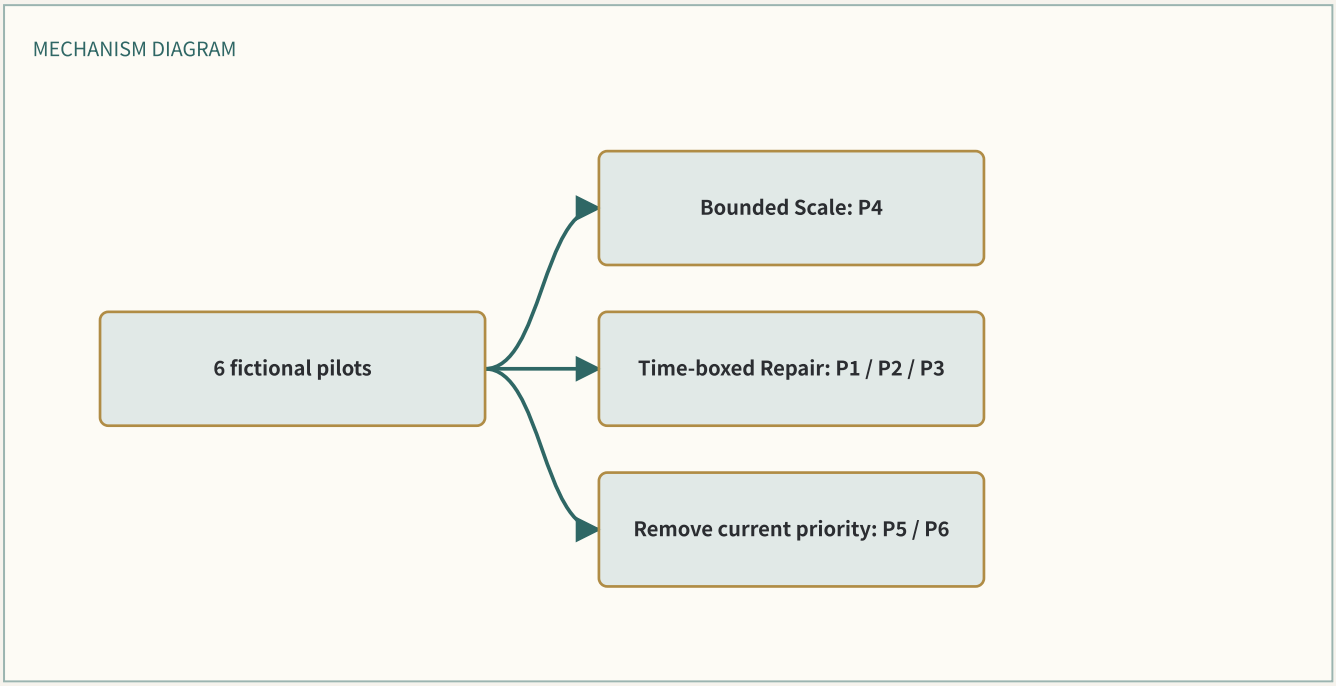

Figure 1 | Illustrative decisions for the fictional portfolio. Kill removes priority within the current portfolio cycle. It is not a permanent prohibition or a judgment against AI or the project team as a whole.

Management implications and principal risks

- Scale requires stable outcome evidence, operating burden, and controls; use or a single demonstration is not sufficient.
- Repair needs a named gap, owner, deadline, and default exit. Without them, it becomes indefinite continuation.
- System writes, permissions, rollback, and high-impact false negatives are hard constraints that potential benefit cannot offset.
- Portfolio decisions need to include integration, review, rework, exception, and maintenance costs carried by shared teams.
- When current priority is removed, retain comparator designs, issue taxonomies, and control records for reassessment if conditions change.

Why portfolio decisions matter more in 2026

The **BCG 2026 executive survey** covered 2,360 executives and found that respondents expected AI spending to rise from about 0.8% of revenue in 2025 to about 1.7% in 2026. Of those surveyed, 94% said their organisation would continue investing even if it generated no return in 2026. The survey measures plans, confidence, and pressure, not audited ROI. It cannot prove that stopping has become harder, but it suggests that stop, reduce, or defer decisions may face greater organisational and reputational resistance among the surveyed organisations.

At the same time, a **McKinsey 2025 survey** of 1,993 respondents found that 88% said their organisation regularly used AI in at least one function, while nearly two-thirds had not reached enterprise-wide scale; 39% reported any enterprise-level EBIT impact. An **IBM 2025 survey** of 2,000 CEOs used another denominator: respondents believed only 25% of AI initiatives had delivered expected ROI, 16% had scaled enterprise-wide, and 50% acknowledged that rapid investment had created disconnected technology.

These figures cannot be averaged because their samples, questions, and denominators differ. They support a limited but important management conclusion: willingness to keep investing may be stronger than the maturity of attributable results. Management therefore needs consistent continuation criteria across projects instead of allowing each pilot to define success on its own.

Apparently conflicting figures measure different things

Source	Sample or method	What it measures	What it can support	What it cannot support
U.S. Census 2026	Nationally representative BTOS AI supplement	Whether companies use AI in business functions	Use remains shallow within an official business frame	Project ROI or management quality
NBER *Firm Data on AI*	Nearly 6,000 senior executives across four countries	Company use, past impact, and future expectations	Broad use can coexist with limited realised impact	A global corporate failure rate
McKinsey 2025	1,993 online survey respondents	Organisational use, scale, and EBIT impact	Scaling and impact are uneven among surveyed organisations	A nationally representative adoption rate
IBM CEO Study 2025	2,000 CEOs	Whether initiatives met expected ROI	A CEO-perceived value-conversion gap	Audited, standardised ROI
Microsoft / NBER / METR	Field experiments or task studies	Productivity in specific tasks	Some settings record measurable local effects, with heterogeneous results	Whole-enterprise profit

A **U.S. Census 2026 working paper** estimated that about 18% of companies used AI in at least one function between 2025/11 and 2026/1. Among functional adopters, 57% covered only three or fewer functions; among companies reporting worker-task use, about 65% covered only three or fewer worker tasks. The survey question changed in 2025/11, so the results should not be spliced directly into the earlier series.

NBER's 2026 *Firm Data on AI* is based on nearly 6,000 senior executives in the United States, United Kingdom, Germany, and Australia. Of the surveyed companies, 69% said they were actively using AI; more than 90% reported no observed employment impact over the past three years, 89% reported no observed productivity impact, and more than 80% reported neither. These findings come from executive self-reports and a four-country sample. They cannot prove that companies achieved no local benefits, but they remind management to distinguish "in use" from "changed enterprise results".

For that reason, this report does not create a single adoption or success chart. Plotting different denominators on one trend line would imply a stronger conclusion than the underlying studies support.

Local productivity changes can be measured, but they are not enterprise value

Task research shows that local benefits can be measured in some settings. An **NBER staggered-rollout study** of 5,179 customer-support agents found that generative AI increased issues resolved per hour by an average of 14%, with gains concentrated among less experienced and lower-skilled workers. **Microsoft Research** combined three field experiments involving 4,867 developers and reported an increase of about 26% in completed tasks, with a standard error of 10.3 percentage points. These studies support the claim that AI can produce measurable improvements in some structured work.

But effects do not point in the same direction across all tasks. A **METR 2025 study** of 16 experienced developers familiar with large open-source codebases and 246 real tasks found that early-2025 AI tools increased completion time by an average of 19%, although participants later believed they had worked 20% faster. This result can only serve as a historical counterexample: METR has explicitly marked the 2025 result as outdated and not representative of current tool effects, while its **2026 follow-up** did not provide a reliable replacement estimate for current effects.

Management should therefore avoid extracting a single average productivity figure from these studies. A more useful approach separates three layers of effect:

- **Task effect:** whether a specific task becomes faster or more accurate.
- **Workflow conversion:** whether the local improvement reduces cycle time, error, or waiting across the workflow.
- **Portfolio value:** whether the improvement deserves shared integration, review, risk, and change capacity.

A pilot can succeed at the first layer, fail at the second, and still not deserve priority at the third. Portfolio decisions therefore need to retain benefit evidence, shared costs, and opportunity costs at the same time.

Why weak-evidence pilots may survive

The gains and losses facing different participants are asymmetric. An AI product team may benefit from use, feature launches, and sponsor approval. The workflow owner handles edge cases and manual fallback. The CIO carries long-term integration and maintenance. The CISO and legal team may carry low-probability, high-impact incidents. The CFO is then asked to show whether these distributed changes reach the income statement.

When each department selects its own success metric, project comparison can deteriorate into a contest of narratives. Figure 2 summarises the resulting mechanism chain.

The principal-agent problem offers one possible explanation for this cycle. Project sponsors are closer to the pilot and see demonstrations and local data sooner. Shared teams understand the long-term operating burden better, but often enter only after budget approval. Information asymmetry may make evidence gaps difficult to see at the decision interface and may also lead organisations to misread “stop” as “failure”.

MECHANISM DIAGRAM

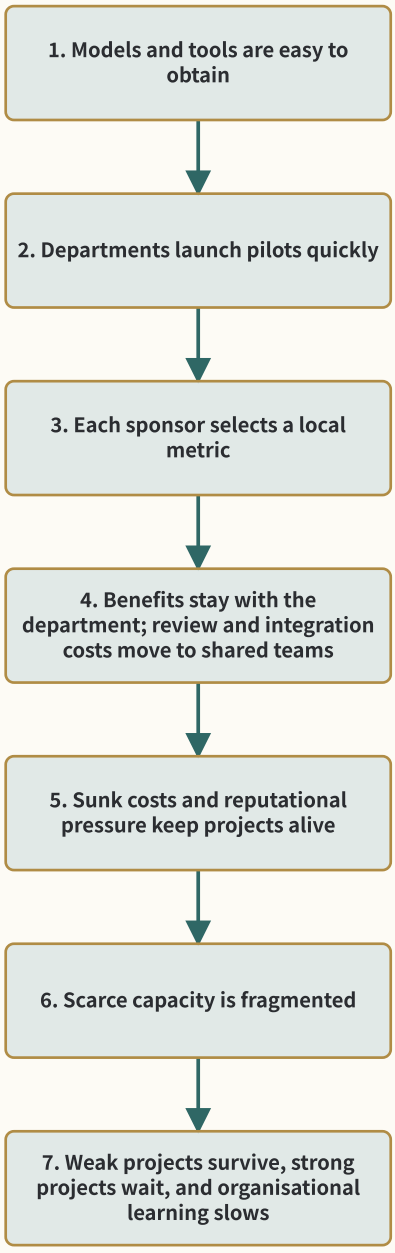

Figure 2 | Pilot-survival mechanism chain. This is a mechanism hypothesis, not an allegation about any company or team. Local metrics, shared costs, and sunk costs may jointly change project survival conditions. Its actual strength needs to be tested against a company's own portfolio history.

Change the unit of comparison: from a tool to a named workflow

Before a pilot can claim scarce delivery capacity, it should submit a one-page contract. The unit of analysis is an end-to-end business workflow with an accepted outcome; the model, seat, and features are only inputs.

In this report, capacity means shared delivery capacity. An accepted outcome is a result accepted by the business that remains valid downstream. A hard gate cannot be offset by a higher score elsewhere. A system write changes the state of a business system, while rollback returns the workflow to a known safe state.

Field	Minimum requirement	Treatment when insufficient
Workflow	Start event, end result, and owner	If it cannot be named, Kill or return it to the idea pool
Accepted outcome	Accepted by the business and still valid downstream	Raw output alone leads to Repair
Baseline / comparator	Concurrent, phased, matched cohort, or stable historical line	Subjective time savings alone cannot support an ROI claim
Measurement window	Start and end dates, task volume, and exception definition	An open-ended pilot cannot consume integration capacity
Operating burden	Manual review, rework, integration, monitoring, change, and fallback	Complete missing fields or retain them as uncertain costs
Owners	Business, technical, and risk owner	Any missing owner blocks Scale
Stop rule	Error, permission, review, or maintenance boundary	If not written before testing, do not start

This contract makes differences visible without pretending that every project needs the same threshold. High-frequency, low-risk knowledge retrieval and a low-frequency, high-impact invoice-control agent require different limits, but both must state what outcome counts, what evidence supports it, and who carries the burden.

Three gates: Evidence, Conversion, and Control

Gate A | Comparable evidence

Management first checks whether the project names its workflow, accepted outcome, baseline or comparator, measurement window, exception definition, and three owners. Use, message count, output volume, or demo speed may still have learning value, but they do not justify the same scale budget as outcome evidence.

Gate B | Conversion evidence and burden

This report does not develop an enterprise ROI model, a complete counterfactual design, or weighted prioritisation. At portfolio level, it asks three questions: how strong is the accepted-outcome evidence against a comparator; how much operating burden do the teams carrying the work confirm; and what other work is displaced when the project uses scarce integration and review capacity?

Field	Portfolio-level record	Treatment when unknown
Conversion evidence	Unknown / Observed / Compared / Repeated	Cannot enter Scale
Operating burden	Unknown / Low / Medium / High; define local bands using hours/month, integration days, or a cost range, then ask the teams carrying the burden to confirm them	An undefined or unconfirmed band leads to time-boxed Repair
Opportunity cost	Integration capacity, security-review slots, validation hours, and delayed projects	Explicit confirmation by the budget owner

Detailed value, cost, and causal attribution require local company data and a separate workflow-measurement package. Without a baseline, an industry average should not be used to fabricate a precise ROI.

Gate C | Control and constraint fit

NIST's GenAI Profile identifies Governance, Content Provenance, Pre-deployment Testing, and Incident Disclosure as four primary considerations, and its suggested actions discuss continuous monitoring and incident response. The **NIST 2026 AI Agent Standards Initiative** foregrounds agent reliability, interoperability, security, and identity. Its launch material also points to **identity and authorisation work**, as well as the access context agents have for external systems and internal data.

This framework therefore includes permissions, traceability, human escalation, rollback, and continuity in the pre-launch check. For pilots involving system writes, customer communication, movement of funds, regulated judgments, or production changes, an unresolved control gap is defined as a hard blocker. This framework judgment does not replace a company's own legal, security, or compliance review.

MECHANISM DIAGRAM

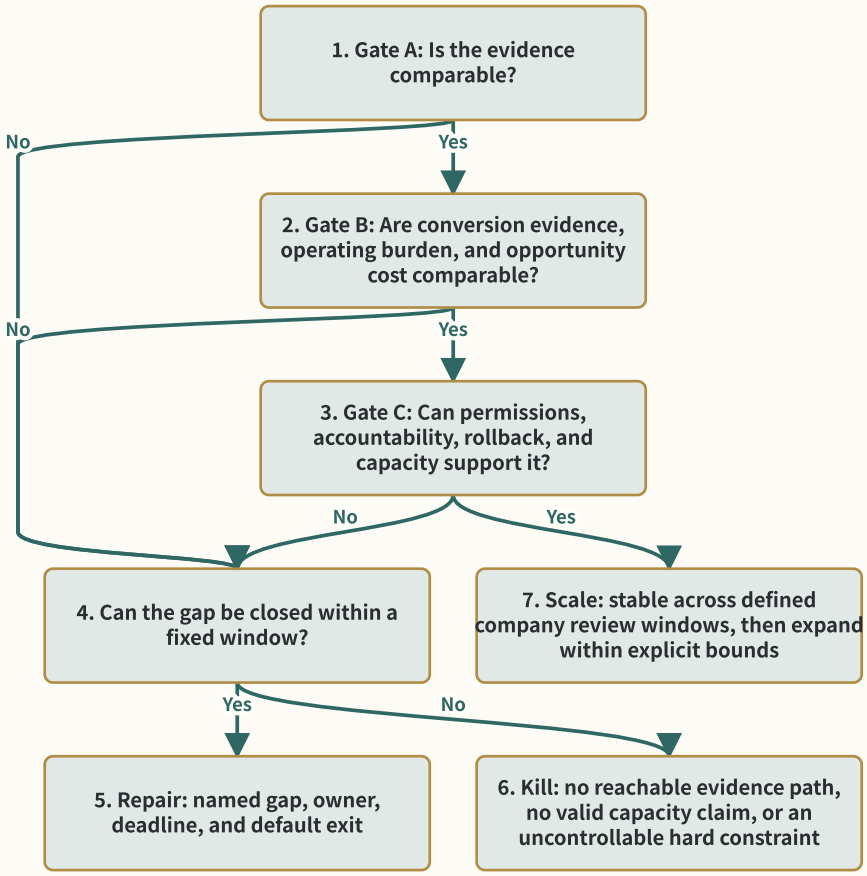

Figure 3 | Three gates and three survival states. Missing evidence or a hard blocker at any gate may lead to Repair, a return to the idea pool, or a stop. Only projects that pass all three gates should be considered for bounded Scale. The 2 consecutive review windows are a cautious illustration used in this report, not an industry benchmark; each company needs to set its own threshold according to workflow risk and local capacity.

Three survival states: Kill / Repair / Scale

Keep can easily become an open-ended intermediate state. Repair instead requires five items to be written down together: the gap, who will close it, when it will be reviewed, what evidence counts as closure, and whether a missed deadline automatically triggers Kill.

State	Applicable conditions	Required record
Kill	No reachable evidence path; an uncontrollable hard constraint; clearly higher opportunity cost	Reason for termination, learning captured, reusable asset, and conditions for resubmission
Repair	Value may still exist and a named evidence or control gap can be closed within a fixed window	Owner, deadline, required evidence, and default exit condition
Scale	Conversion evidence, operating burden, and control conditions remain stable across at least two review windows, and the budget owner confirms the opportunity cost	Bounded scope, capacity budget, monitoring, rollback, and next review

In this framework, Kill removes a project's current priority claim on scarce capacity; it is not a judgment against AI or the team as a whole. If new evidence, lower costs, or improved controls change the conditions, the project may re-enter the idea pool with a new contract.

Portfolio dashboard: separate value evidence from shared capacity

Value and evidence

Metric	Recommended definition	Common misreading
Pilot conversion	Number of projects and time spent moving from idea to test to compared to scale	More projects mean stronger innovation
Accepted-outcome rate	Results accepted by the business and still valid downstream / total results	Output volume equals value
Review burden	Expert-review, exception, and rework hours for every 100 results	Human-in-the-loop automatically means safe
Evidence depth	Portfolio distribution across Unknown, Observed, Compared, and Repeated	A single before-and-after result is causal evidence

Capacity and control

Metric	Recommended definition	Common misreading
Integration congestion	Days waiting for integration, shared dependencies, and maintenance queue	An API connection completes integration
Control integrity	Permission failures, unsafe actions, rollback success, and incidents by severity / 1,000 runs	Low reporting or no incident means mature controls
Overdue unresolved pilot rate	Share of pilots that have passed the repair deadline without producing the required evidence	A continuing pilot is a long-term option
Reallocation speed	Time from kill decision to capacity reallocation	Termination itself has released value

These metrics still require local definitions. Their purpose is to show management portfolio learning and bottlenecks, not to create another composite maturity score.

Metric integrity rules

1. Freeze the eligible cohort, numerator, denominator, and exclusion rule before testing. Any change needs a retained version and a restarted comparison window.
2. The workflow owner certifies the accepted outcome; the sponsor cannot self-certify the denominator, cost, or control.
3. Record excluded cases, shadow work, manual workarounds, and near misses. Missing items are Unknown, not zero.
4. Give each project a stable pilot ID so renaming cannot erase its repair deadline or incident history.
5. Report incident severity, reporting completeness, and rollback success together under control integrity.
6. Pair reallocation speed with the realised outcome after reallocation to avoid rewarding hasty termination.

To prevent the portfolio mechanism from rewarding only easy-to-measure, low-risk projects, reserve a capped validation window for high-uncertainty work. The window expires, cannot bypass production hard gates, and cannot renew automatically.

A reversible 90-day operating cadence

Day 1-10 | Inventory

List every AI pilot, owner, workflow, current permissions, shared dependencies, cumulative spending, available evidence, and next review. If the integration queue is already visibly congested, pause new commitments; low-risk idea discovery can continue.

Day 11-30 | Standardise

Put every project on the same one-page evidence contract. Record use metrics separately from business outcomes, and list missing baselines, review costs, permissions, and rollback as gaps to close.

Day 31-60 | Bounded tests

In companies where shared capacity is genuinely scarce, give priority to evidence-ready or high-learning-value pilots. Use shadow mode, read-only operation, decision support, phased rollout, and recoverable manual processes where appropriate, while recording negative and null results.

Day 61-75 | Stress test

Before finalising a decision, test the provisional conclusion against higher costs, incidents, key-person departure, audit demands, or demand shifts. If a project cannot degrade safely, roll back, or retain clear accountability, move it back to Repair or Kill. The check should cover denominators, incentives, missing owners, and shared-cost transfer, not only model accuracy.

Day 76-90 | Portfolio decision

Make a Kill / Repair / Scale decision for every project. Repair needs a deadline and default exit. Scale needs a bounded scope, capacity budget, monitoring, rollback, and next review. Explicitly reallocate released capacity so it is not consumed invisibly.

What to watch in 2026-2027

The following is a directional watchlist, not a probability forecast.

Watch 1 | From adoption stories to portfolio accountability

- **Directional expectation:** If budget and board pressure continue, CEOs, CFOs, COOs, and technology and risk owners are more likely to decide together whether a project should keep receiving investment, enter repair, or stop.
- **Leading indicators:** Public cases begin to disclose the workflow owner, validation burden, full operating cost, kill decision, and capacity reallocation.
- **Falsifier:** Large enterprises continue to expand only by seats, messages, or pilot count without visible budget, incident, or maintenance consequences.
- **Review dates:** 2026-10-15; 2027-01-15.

Watch 2 | Agent identity and authorisation enter the scale gate

- **Directional expectation:** As system access expands, identity, least privilege, reconstructible logs, human escalation, and rollback are more likely to become prerequisites in customer and architecture reviews.
- **Leading indicators:** The NIST initiative produces implementable standards; procurement, audit, or cyber-insurance requires agent identity and action traces.
- **Falsifier:** Most agent deployments remain low-risk and without system writes for an extended period, while related standards do not affect enterprise procurement or launch.
- **Review dates:** 2026-10-15; 2027-02-15.

Watch 3 | High-quality Kill decisions become a learning metric

- **Directional expectation:** If portfolio congestion rises, mature teams may begin recording kill quality, reusable learning, and reallocation speed.
- **Leading indicators:** Transformation dashboards disclose cancelled pilots, repair deadlines, and reused controls, rather than adoption alone.
- **Falsifier:** Organisational performance and financing narratives continue to reward project count and expansion in only one direction, with no visible stopping mechanism.
- **Review date:** 2027-01-15.

Boundaries between evidence, inference, hypothesis, and recommendation

Fact. The surveys, official firm data, and field experiments cited in this report support the following facts: AI use and spending intentions are rising; enterprise-wide scale and attributable impact are uneven across surveys; and task effects are heterogeneous. The NIST 2026 initiative foregrounds agent reliability, interoperability, security, and identity. Its launch material also points to identity and authorisation work, as well as the access context agents have for external systems and internal data.

Inference. This report infers that portfolio selection and cross-functional accountability may become constraints when multiple pilots compete for shared capacity. The sources support integration, review, data, change, and governance frictions, but cannot prove that these factors are the primary bottleneck in every company.

Hypothesis. Local metrics, cost shifting, sunk costs, and sponsor incentives may increase the survival rate of weak pilots. This mechanism needs to be tested against a company's own portfolio history, meeting records, and cost ledger.

Recommendation. Use a common evidence contract, three gates, and time-boxed Repair; present management with the decision, evidence level, unresolved risks, bounded scope, next review, and default exit. This is a reversible management experiment, not accounting, legal, compliance, investment, or vendor-specific advice.

Uncertainty. This report has no project-level baseline, cross-pilot comparator, or complete recurring-cost ledger from a specific company. It therefore provides portfolio decision criteria rather than deciding which project any company should scale.

Update and correction policy

- Source cutoff: 2026-7-14.
- If a primary source revises its sample, denominator, or conclusion, the affected pages should be checked again.
- Stronger enterprise-level causal evidence, public negative cases, or agent standards should trigger an evidence update.
- If subsequent real-world portfolio outcomes conflict with the report's mechanism, retain the original judgment, record the deviation, and issue a dated review or correction rather than silently rewriting it.
- This is a free public report, version v2.2. If substantive evidence changes the conclusion, the original judgment will be retained and the change explained in a dated correction note. Readers should apply the framework with local data and complete their own financial, legal, security, and compliance review.

Primary sources

- NBER, Firm Data on AI (2026)
- U.S. Census Bureau, The Microstructure of AI Diffusion (2026)
- BCG, As AI Investments Surge, CEOs Take the Lead (2026)
- McKinsey, The State of AI 2025
- IBM, CEO Study 2025
- Microsoft Research, The Effects of Generative AI on High-Skilled Work (2025)
- NBER, Generative AI at Work (2023)
- METR, Early-2025 AI Study of Experienced OSS Developers
- METR, 2026 Uplift Update
- NIST AI 600-1, Generative Artificial Intelligence Profile
- NIST, AI Agent Standards Initiative (2026)
- NCCoE, Software and AI Agent Identity and Authorization

Further reading (not used for central factual judgments)

- OECD, Empowering SMEs in the Age of AI (2026)
- BCG / MIT SMR, Responsible AI Needs More Than Good Intentions (2026)
- Anthropic, The 2026 State of AI Agents
- BCG, CEOs and Boards Are Aligned on AI in Theory but Divided in Practice (2026)