InfoSecAI

A BRIEF FOR CISOS · BISOS · GRC LEADERSHIP

WHITE PAPER · PAPER 4 · 2026

Why AI Transformation Fails After the Pilot

A practitioner's brief for CIOs, CTOs, Chief AI Officers, transformation leaders, executive sponsors and CISOs who need AI pilots to convert into measurable enterprise value.

AUTHORED BY

Paul Jolliffe

By Paul Jolliffe, Founder and Director, InfoSecAI Limited · MBA · CISSP · ISO 27001 Lead Auditor

01 · EXECUTIVE SUMMARY

Executive summary

Most large organisations now have artificial intelligence (AI) pilots. Many have copilots deployed. Several have run more than fifty internal experiments. Very few can point to a process that has been redesigned around AI, a control set updated to reflect the new workflow, and a measurable outcome that closes the business case the pilot was authorised against.

That is the AI execution gap. Usage is rising. Enterprise value is uneven. The pattern is consistent across sectors and consistent enough that it is no longer best explained by model capability, pilot quality or vendor selection.

This brief is the fourth in InfoSecAI's five-part executive series, From AI Ambition to AI Assurance. The first three covered the operating model, the discovery of the AI estate, and the control plane for agentic AI. This paper extends the argument into the place where most executive AI investment is currently stalling: the transition from pilot to scale.

The thesis is direct. AI pilots rarely fail because the demo was bad. They fail because nobody redesigned the workflow, the data, the controls, the ownership or the measurement model around them. Security and governance are not what slow AI transformation down. Weak operating models are what prevent it from scaling.

02 · WHY THIS MATTERS NOW

Why this matters now

Three pressures are converging on executive teams.

The first is investor and board scrutiny. The capital allocated to AI in 2024 and 2025 was largely permissive: every meaningful pilot got funded. The 2026 conversation is shifting to return on the money already spent. Boards are asking what measurable outcomes the pilots produced, and most cannot answer with operational evidence.

The second is the published academic and industry signal on the pilot-to-value gap. Recent research by the Massachusetts Institute of Technology (MIT) Center for Information Systems Research and parallel reporting from large consultancies have repeatedly observed that the majority of enterprise generative AI (GenAI) pilots have not produced measurable business value, even where individual user productivity gains are real. The numbers reported vary by methodology and definition; the directional finding is consistent.

The third is operational. AI pilots run by individual teams have not changed how work is organised, who owns the outcome, or how value is measured. A copilot that saves an analyst one hour a day produces 220 hours of saved time across a team in a year. Whether the organisation captures that as cost reduction, throughput increase, product velocity or quality improvement depends on the operating model, not the tool.

The combined picture is uncomfortable. Capability is real. Investment has been substantial. The execution gap is now the dominant story.

03 · WHY PILOTS FAIL TO SCALE

Why pilots fail to scale

Five patterns recur in the pilots InfoSecAI has reviewed.

The first is the tool-shaped pilot. The pilot is designed around what the tool can do, not around what a business process needs. A summarisation tool gets piloted in three or four teams. Each team uses it differently. None of the teams change the surrounding workflow. The pilot reports show usage and user satisfaction. The business case asks for measurable productivity, cost or quality improvement, and the report cannot answer it.

The second is the unowned pilot. The pilot has a project sponsor and a vendor lead, but no operational owner who will run the process after the pilot ends. When the project closes, the tool floats. Adoption decays. The value, if any, is not captured because no one is accountable for the redesigned process.

The third is the measurement-free pilot. The pilot does not start with a value hypothesis and a metric. It starts with "let us see what this can do". Six months later there is no baseline against which to measure improvement, and the success case becomes a story.

The fourth is the governance-blind pilot. The pilot runs without an explicit risk classification, a control set, or evidence of operation. When the time comes to scale, internal audit or risk functions ask questions the pilot cannot answer, and the scale decision pauses.

The fifth is the parallel-process pilot. The pilot operates alongside the existing process rather than replacing it. The team uses both. Throughput does not improve because the new path is additive. Cost does not reduce because the old path still exists. The pilot is a thing the team does on top of the day job.

The common factor is structural. Each pattern has been rational at the pilot stage. None survives contact with scale.

04 · THE AI VALUE CHAIN

The AI value chain

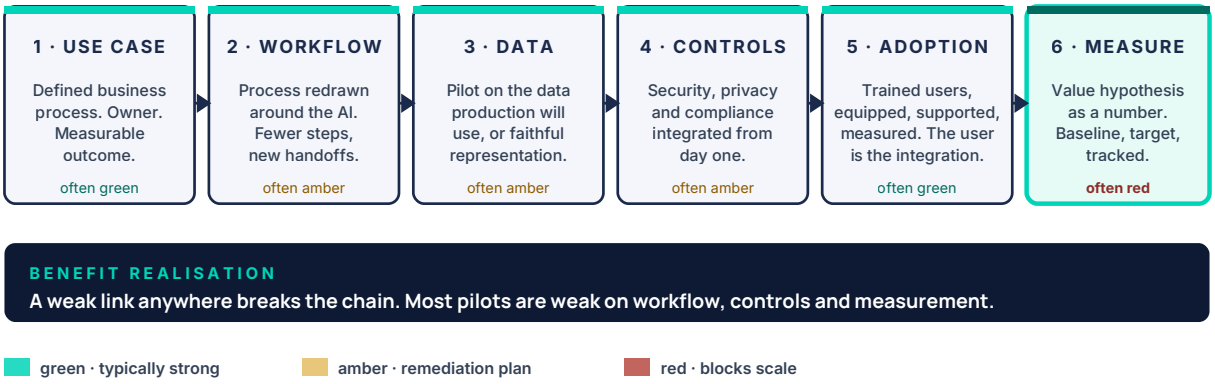
Value from AI flows through a chain. The chain has six links. A weak link anywhere breaks the chain.

The links are sequential and dependent. A pilot can be strong on use-case selection and on user adoption and still produce no measurable enterprise value, because the middle links carry the value. Workflow that has not been redesigned around the AI keeps the existing process cost. Data that does not match production conditions hides the production failure mode. Controls that are retrofitted after the pilot delay scale by six months. Measurement that is bolted on at the end leaves the value undefined.

The pattern across the pilots InfoSecAI reviews is consistent. Pilots score well on the first and fifth links, and weakly on the middle three. The implication is that AI value programmes have to be designed around the weakest links, not the strongest. Investment in adoption training while measurement is missing is wasted. Investment in measurement infrastructure while controls are not integrated is premature.

FIGURE 1 · The AI value chain.

Six links to benefit. The chain is only as strong as its weakest link.



Use-case selection. The use cases that scale share three characteristics: a defined business process, a measurable outcome, and a sponsor who owns the process after the AI is introduced. Use cases that lack any of the three should not enter pilot.

Workflow redesign. Inserting AI into a workflow without redesigning the workflow rarely produces value. The redesign question is "what does the work look like when AI is doing this part". The answer is a redrawn process with fewer steps, different handoffs, and different measurement points.

Data readiness. AI inherits the quality of the data it is given. Pilots that succeed on demo data and fail on production data are usually data-readiness failures, not model failures.

Controls. Security, privacy and compliance controls must be integrated at the pilot stage, not retrofitted at scale. Retrofitting is what produces the six-month delay between pilot success and production deployment.

Adoption. The user is the integration layer. The user has to be trained, equipped, supported and measured. Adoption is not a single training event; it is a workflow change.

Measurement. The pilot starts with a value hypothesis stated as a number. Cost saved per month, throughput per analyst, defects detected, customer time to resolution. The measurement design is built into the pilot, not bolted on after.

05 · THE PILOT-TO-PRODUCTION READINESS SCORECARD

The pilot-to-production readiness scorecard

A pilot is ready to scale when six dimensions are green. The scorecard makes this explicit. Each dimension has a pass criterion that the sponsor, the operational owner and a control-

function representative can score jointly. The pass criteria are deliberately narrow so the scoring conversation is short.

The discipline is the test, not the score. If the three reviewers cannot agree on the rating for any one dimension, the pilot is not ready to scale; the disagreement is the signal. In practice, most disagreements concentrate on workflow redesign and measurement, where the gap between the pilot team's view and the operational owner's view is widest.

FIGURE 2 · Pilot-to-production readiness scorecard.

Six dimensions. Any red blocks scale. Three or more amber means redesign.

DIMENSION	PASS CRITERION
Value hypothesis	Stated as a number, with baseline and target
Workflow redesign	Process redrawn around the AI, not added beside it
Data readiness	Pilot uses production-equivalent data
Controls	Risk classification, control set, evidence operating
Adoption	Trained users, equipped, supported, measured
Measurement	Metric tracked, refreshed monthly, owner named

ALL GREEN · SCALE | 1-2 AMBER · SCALE W/ CONTROLS | 3+ AMBER · REDESIGN | ANY RED · STOP

The discipline is the test, not the score. The scorecard is run jointly by the sponsor, the operational owner and a control function representative. If the three cannot agree on the score, the pilot is not ready.

Most pilots InfoSecAI sees are amber on workflow redesign, amber or red on measurement, and amber on controls. They are green on adoption and on value hypothesis. The pattern is recognisable: the pilot got far enough to demonstrate that users will use the tool, and not far enough to demonstrate that the business will get the value.

06 · THE STOP, SCALE, REDESIGN MATRIX

The stop, scale, redesign matrix

Not every pilot should be scaled. The matrix below provides a decision framing built around two axes that any executive can score.

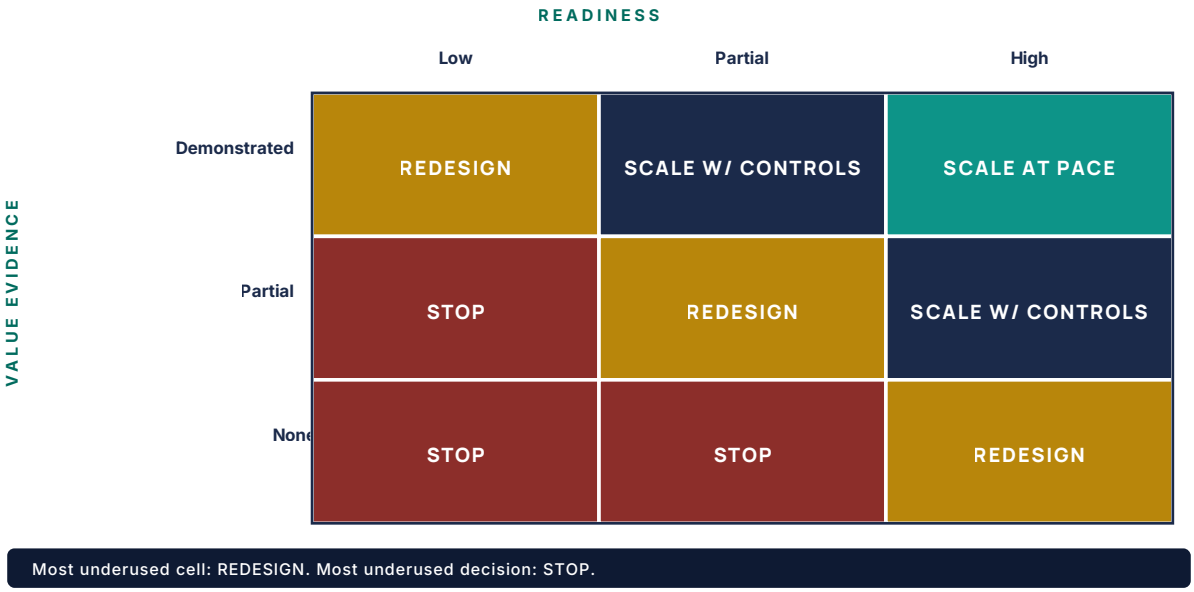
Value evidence is the first axis. Demonstrated value means a measurable change against baseline in the pilot conditions. Partial value means directional improvement without statistical confidence. None means user satisfaction or activity counts without a business outcome.

Readiness is the second axis. High readiness means the pilot would pass the scorecard above. Partial means at least one amber dimension. Low means two or more red dimensions, typically including measurement or workflow.

The cells map to four decisions and one rule: nothing scales from the red column. The matrix is the place where political pilots are stopped.

FIGURE 3 · Stop, scale, redesign decision matrix.

Pilots map to one of four decisions. Most 2026 pilots land in REDESIGN, not SCALE.



Most underused cell: REDESIGN. Most underused decision: STOP.

The redesign cell is the cell most organisations underuse. Redesign means the use case is sound and the value is plausible, but the workflow, data or measurement design is not. The pilot does not progress; it goes back into design with a defined remediation list and a re-pilot.

The stop cell is the cell most organisations underuse for political reasons. A pilot with weak value evidence and weak readiness is not a future scale candidate. Stopping is a decision, not a failure.

The scale-with-controls cell is the most common destination. The use case is sound, the value evidence is partial but credible, and the readiness is high enough to deploy with the existing control framework. The pilot scales with explicit acceptance of the remaining risk and an owner for the residual remediation.

The scale-at-pace cell is rare. It requires demonstrated value and high readiness. Pilots that reach this cell typically scale within ninety days.

07 · THE EXECUTION GAP

The execution gap

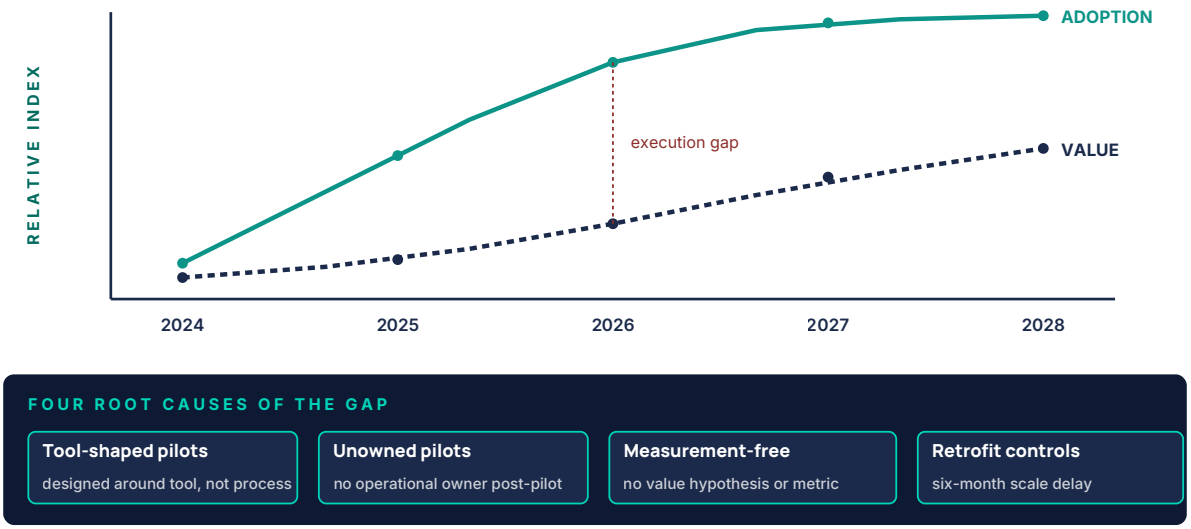
The pattern across the five failure modes and the scorecard results is summarised in the gap diagram. The chart shows two trend lines from 2024 to a projected 2028. One traces adoption, defined as the percentage of knowledge-work hours touched by AI tools. The other traces measurable enterprise value, defined as documented cost, throughput, quality or revenue change attributable to AI deployments.

Adoption is rising steeply. Measurable value is rising slowly. The widening gap between them is the execution gap, and it is the dominant story in 2026 enterprise AI. It is not a model problem. It is not a tool problem. It is an operating-model problem.

The same chart attributes the gap to four root causes from the failure modes earlier in this paper. The size of each root cause indicates how much of the gap it explains in the portfolios InfoSecAI has reviewed. Tool-shaped pilots and measurement-free pilots account for more than half between them.

FIGURE 4 · The AI execution gap.

Adoption rising. Measurable value rising slowly. The gap is operating-model failure, not model failure.



The frame is consequential. An organisation that treats the gap as a model problem will keep buying tools. An organisation that treats it as an operating-model problem will redesign work, name owners, build measurement and integrate controls early. Only the second route closes the gap.

08 · INFORMATION SECURITY IMPLICATIONS

Information security implications

Transformation pilots that succeed and transformation pilots that stall both end up on the CISO's desk. The first arrive needing control evidence that should have been built in; the second arrive when an internal audit asks what happened.

The implication is that AI transformation and the security control plane have to be wired together at pilot day one, not after the scale decision is made. Six integration points carry the load.

Use-case intake. Every pilot enters through a single intake that classifies risk and data exposure, assigns an owner, and triggers the relevant control work. Pilots that bypass intake are unauthorised by definition.

Data access. The pilot runs on the data the production deployment will run on, or a faithful representation. Pilots that use sanitised demo data systematically understate the data readiness gap.

Supplier and AI feature review. Pilots that depend on a supplier's AI feature are subject to the supplier review process, including the no-training, data-residency and audit clauses described in earlier papers in this series.

Logging and reconstructibility. From pilot day one, the standard is whether the security team can reconstruct what the AI did. Pilots that do not meet the standard cannot scale without rework.

Incident response. The incident classification standard operating procedure includes an AI branch covering hallucination harms, data exposure and agentic action errors. The pilot tests the branch on a tabletop exercise before scale.

Business continuity. AI-enabled processes need a defined manual fallback. Pilots that erode the manual path without proving the AI path can carry the load are creating operational risk.

09 · EXECUTIVE DECISIONS AND THE 30-DAY CLARITY SPRINT

Executive decisions and the 30-day clarity sprint

The six executive decisions

Six decisions reliably move stalled portfolios. Each is an executive call, not an analyst recommendation.

Who is accountable for AI value realisation? One named individual, typically the chief operating officer or transformation leader. The chief information security officer (CISO) is accountable for risk; the chief financial officer is accountable for value capture.

What is the maximum number of concurrent pilots? Fewer than the current number, in almost every organisation. Reducing pilot count reliably increases pilot quality.

What is the standing value hypothesis template? One page, three sections: the metric, the baseline, the target. Every pilot uses it. No template, no pilot.

What is the standard control set every pilot inherits on day one? The set drawn from the AI governance operating model in Paper 1 of this series. Pilots that need additional controls request them; pilots that need fewer do not exist.

What is the decision cadence for the stop, scale, redesign matrix? Monthly for high-velocity portfolios; quarterly for the rest. Skipped reviews are decisions to do nothing.

What evidence proves the portfolio is producing value? Three artefacts: the value scorecard, the latest scorecard refresh, and a narrative of two pilots that progressed and one that stopped.

The 30-day clarity sprint

The first thirty days do not require new platforms or new vendors. The sprint runs in four ordered weeks.

Week 1, inventory. Every pilot, with sponsor, owner, value hypothesis, current state, and last scorecard. The output is a single-page register that everyone can read.

Week 2, score. Run every pilot through the readiness scorecard. Categorise as stop, redesign, scale with controls, or scale at pace. Disagreements on the score are escalated to the sponsor.

Week 3, act. Stop the stops. Begin the redesigns with a remediation list and a date. Authorise the scales with documented controls. Visible action in week three is what makes week four matter.

Week 4, report. Build the quarterly board page: portfolio composition, value progress, four decisions, three numbers. The board page is the deliverable; the sprint is the means.

Worked example: Northgate Logistics plc

Northgate Logistics plc, a mid-sized UK logistics business, ran the clarity sprint in April 2026 after a board challenge on AI return on investment.

The portfolio review surfaced 31 pilots. Sixteen were stopped within the month, including five that had no operational owner and four that had no measurable hypothesis. Eleven moved to redesign with named owners and remediation plans. Three scaled with controls, including a route-planning copilot that delivered a documented 4 per cent fuel saving across the controlled trial. One scaled at pace: customer-service email triage with a fully redesigned workflow and a measured 22 per cent reduction in time to first response.

The board pack changed from a list of 31 pilots to a value chart showing four decisions and three numbers. The board met once on the new format; the executive cadence then moved to monthly.

10 · QUESTIONS EVERY LEADER SHOULD ASK NOW

Questions every leader should ask now

For the three highest-priority AI use cases, can the executive name the operational owner who will run the redesigned process after the pilot ends? If not, those pilots are not ready to scale.

What value did the portfolio produce in the last quarter, expressed in cost saved, throughput gained, defect reduction or revenue captured? If the answer is not a number, the measurement design is missing.

How many of the current pilots have a working scorecard, refreshed in the last sixty days? If fewer than half, the operating model is not yet running the portfolio.

For the most recent pilot to scale into production, how long did it take to retrofit the controls? If the answer is more than thirty days, controls are not being integrated early enough.

What is the single pilot decision that has been pending the longest, and what evidence would unblock it? That decision is the priority.

11 · CLOSING THOUGHT

Closing thought

AI transformation is not failing because the models cannot do the work. It is failing because the surrounding operating model has not been changed to capture the value the models produce. Closing the execution gap is the executive job of the next twenty-four months.

The thread of this series continues: AI assurance evidence, not AI reassurance narrative. An organisation that can point to a measured outcome from a controlled AI deployment is one that can scale. An organisation that can only point to pilot activity is not yet a transformation case.

The final paper in this series, The Board Pack for AI Assurance, sets out what the board needs to see, by quarter, to oversee the work.

12 · SOURCE REGISTER

Source register

All sources verified to primary publisher on 4 June 2026.

#	SOURCE	USE	LINK
1	MIT CISR enterprise GenAI value research	Pilot-to-value gap	https://cisr.mit.edu
2	EU AI Act (Regulation 2024/1689)	Articles 14, 26	https://eur-lex.europa.eu/eli/reg/2024/1689
3	NIST AI RMF 1.0	Govern / Map / Measure / Manage	https://www.nist.gov/itl/ai-risk-management-framework
4	NIST AI 600-1 Generative AI Profile (2024)	GenAI risks and actions	https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
5	NCSC Secure AI System Development Guidelines	Production deployment framing	https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development
6	ISO/IEC 42001:2023 AI Management System	Portfolio governance	https://www.iso.org/standard/42001

13 · ABOUT THIS SERIES

About this series

From AI Ambition to AI Assurance is a five-paper executive briefing series, 1 to 5 June 2026.

1. AI Governance Is No Longer a Policy Problem
2. The Shadow AI Exposure Map
3. Securing Agentic AI Before It Acts
4. Why AI Transformation Fails After the Pilot (this paper)
5. The Board Pack for AI Assurance

Each paper is published as a 12-page executive briefing under the InfoSecAI Blog Template. The full series is available at infosecai.net/insights for subscribers to the InfoSecAI insights list.

14 · PRACTITIONER NOTE

Practitioner note

This briefing is practitioner interpretation, not legal advice. For regulated deployments, validate final claims against current legal obligations, sector-specific requirements and the original primary sources before relying on them.

InfoSecAI

About InfoSecAI

InfoSecAI is an independent UK consultancy helping organisations turn security, regulatory, resilience and AI governance requirements into practical operating models, stronger controls and robust delivery.

We work across strategy, governance, risk, compliance, AI security, assurance, operations and engineering. Our services help leadership teams assess their current position, align to standards and regulation, define the target operating model, and deliver the governance, controls, artefacts and ways of working needed to move from intent to implementation.

Our toolkit capability accelerates structured work across ISO 27001, ISO 22301, ISO 42001, NIST CSF, NIST AI RMF, CIS Controls, Cyber Essentials, DORA, NIS 2, the EU AI Act, GDPR, UK GDPR, SOC 1 and SOC 2. The approach combines AI-enabled workflow support with senior practitioner judgement, so outputs remain proportionate, usable and connected to the way the organisation actually operates.

InfoSecAI was founded in **2025** by **Paul Jolliffe**. The company is built for organisations that need clarity, senior leadership and hands-on delivery across information security and AI governance, without adding unnecessary complexity or treating compliance as a paperwork exercise.

infosecai.net · paul.jolliffe@infosecai.net