

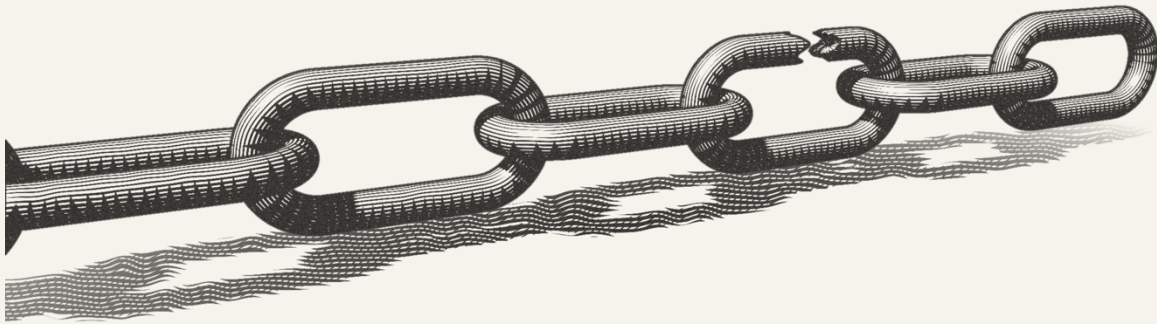
MESA

Capital Partners

WHAT BREAKS NEXT · RESEARCH SERIES · ISSUE 01

Moving AI from Pilots into Enterprise Workflows

The model usually works. The workflow around it is what breaks. A field guide to ownership, data, security, review, and cost.



September 2026 · For enterprise IT practitioners · By Megan Dahlgren

00 • EXECUTIVE SUMMARY

At a glance

Moving AI from pilot to production means managing everything between a request and a verified business outcome. A model can read a ticket perfectly and the service can still fail because of stale data, wrong permissions, or an update nobody can confirm. Those dependencies get harder to absorb as deployments grow.

88% ➞ 7%

use AI in at least one function — vs. have fully scaled it

McKinsey global survey, 2025 · adoption, not failure [1]

40%

faster feature delivery at Doctolib

Anthropic, 2026 · vendor-reported [2]

70–75%

of variable running cost is human oversight — not the model

McKinsey banking example, 2026 [3]

FIVE THINGS THAT BREAK — AND WHAT TO DECIDE

WHERE IT BREAKS

THE DECISION TO MAKE FIRST

1 The workflow boundary	Where the AI's job starts and ends and who owns the whole path?
2 Data Access	Which source is authoritative, what AI may write, and who checks?
3 Security	Which tool / what permissions?
4 Human Review	How much human review and oversight can this agentic workflow justify?
5 Economics	What is the total cost to regularly deliver this agentic outcome?

THE BOTTOM LINE

Give each workflow a business owner and a service owner, define what AI agents may read and change, and value in dollar terms the business outcomes before release. Expand only when a workflow meets its quality bar at an acceptable cost for the desired outcome.

SCENARIO	ILLUSTRATIVE · THE SERVICE DESK
	The pilot that broke scaling to production An AI assistant reads a ticket, finds the right knowledge-base article, and drafts a good answer. The demo works. Then it meets production. The article is out of date. The employee lacks access to the recommended app. A provisioning request times out after the downstream system already accepted it. Retry, and you create a duplicate. Stop, and the employee waits. The model understood the ticket perfectly. The workflow still failed.

Pilots survive because engineers quietly patch what surrounds the model: data access, transaction state, human decisions, and ownership. Production has to make each of those explicit. This report follows where they break, in the order most teams hit them.

An impressive demo proves a capability. Daily operations expose the decisions, dependencies, and costs that determine whether it delivers value.

01 · WORKFLOW

The workflow boundary breaks first

Drafting, recommending, and executing are three different jobs. Calling them all “ticket automation” hides decisions that need both engineering and business sign-off. Before the build starts, builders, the business owner, and security must agree exactly where the AI’s job begins and ends.

Mapping the path from request to verified outcome across systems of record, human decisions, and exceptions is a cultural shift as much as a process one. The path crosses silos, each with its own executive. Someone must own it end to end, and in most organizations that question has no answer yet.

MORGAN STANLEY



STORY · COMPANY-REPORTED, 2024

Debrief: a bounded design

Morgan Stanley Wealth Management offers a rare counterexample to the enterprise AI adoption curve. After launching AI @ Morgan Stanley Assistant, a GPT-4-powered knowledge tool built on roughly 100,000 curated internal documents, the firm reported that more than 98% of its advisor teams were actively using it, far above the 30–40% plateau typical of enterprise software rollouts. The firm estimates that advisors' access to relevant internal knowledge rose from about 20% to 80%. A follow-on tool, AI @ Morgan Stanley Debrief, extends the same approach to client meetings by automating notes, CRM logging, and follow-up drafts, reportedly saving around 30 minutes per meeting.

The AI drafts. The advisor decides. The boundary is the design. [4]

RESEARCH CALL · WE WANT TO HEAR FROM YOU

Operators: tell us about a handoff that failed after rollout the impact, and the change that fixed it.

Share your experience → mesapartners.io

02 · DATA

Data access becomes an operating job

Once an agent can read and write business data, three questions need owners:

- **Authority.** Which source is the record of truth?
- **Verification.** Did the update reach the CRM and is it right?
- **Attribution.** Who is accountable for what the AI changed, especially customer and financial data?

Retrieval should carry provenance, freshness, and permissions with it, and the systems behind the AI must enforce them. When someone's access is removed, test that they can't reach the same information through search indexes, caches, or saved outputs. Telling the model to respect permissions is not a control.

ROLE

OWNS

Business owner	What the data means
CRM administrator	Where the AI is allowed to write
Designated reviewer	Resolving questionable updates

Every change should be traceable to its source, the automation that made it, and any human approval.

GONG



STORY · PRODUCT DOCUMENTATION

One tool, three lessons

Gong's AI Data Extractor does something every sales leader wants: it listens to calls, pulls out the answers that matter, and drops them straight into the CRM. No more reps forgetting to update the deal stage. But look closely at how it handles data, and three quieter lessons emerge.

It writes on its own. Once an admin maps extracted answers to CRM fields, the agent keeps overwriting them as new calls come in, with no human in the loop. That's the point, and also the risk, which is why Gong itself recommends testing before you publish. [5]

It reads more than you can. An extractor can draw on calls the admin who configured it isn't allowed to see. Access to a recording and access to what's derived from it turn out to be two different permissions. [5]

It doesn't truly forget. Delete data in Gong, and it can quietly re-sync from a connected source. Real deletion must happen upstream, where the data lives. [6]

None of this is a flaw in Gong. It's what happens when an agent is trusted to write, read, and retain on a company's behalf, and it's exactly where governance tends to break next.

*Something said on a call doesn't automatically become a business record.
E.g. "We might buy in Q3" is not a commitment.*

Authenticated and still wrong

A connected system can be available, authenticated, and working as designed and still write the wrong thing to the CRM. Traditional integration controls still matter. AI adds a new question: did the system interpret the input correctly before it acted?

Start with one workflow and agree on what success looks like:

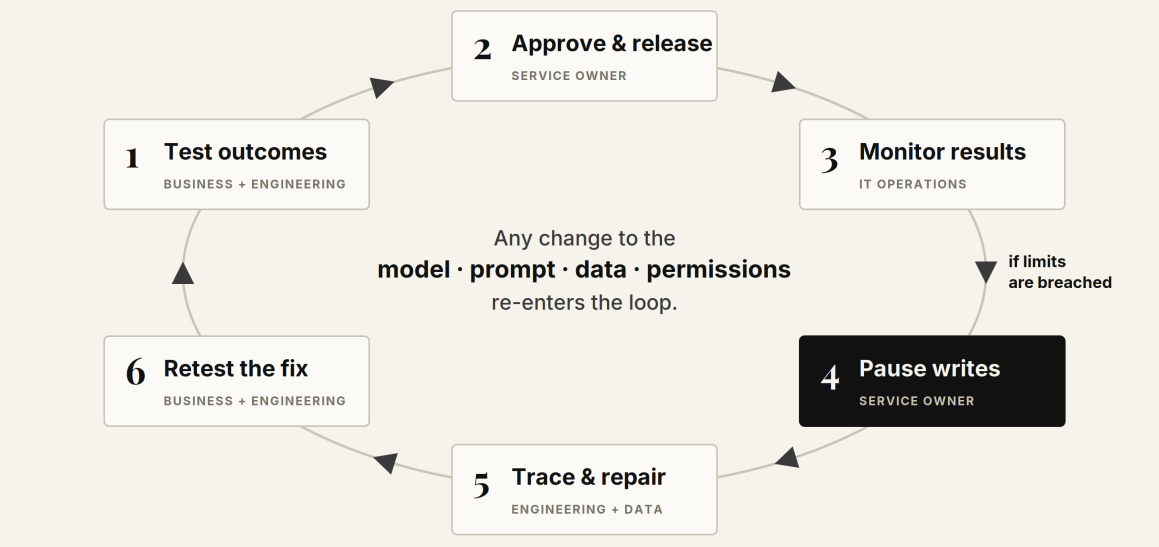
TEAM	RESPONSIBLE FOR
Sales operations	What becomes an official record, and what needs review
Engineering & security	Enforcing those boundaries
Service owner	Detecting failures and restoring the business process
Compliance	Evidence retained to explain a consequential update

Make it part of release. A high success rate means little if the failures include the wrong customer record. Test the tasks people actually perform: a customer changing their mind, two calls that conflict, an update outside the agent's permissions, and agree in advance which failures block release. [7]

Plan the operational response. If an update times out, the service desk must know whether it failed or succeeded with a lost confirmation. Check the destination, or use duplicate-safe writes, before retrying. Keep enough history to show which source informed a change, which workflow version ran, and who approved it.

WHAT BREAKS NEXT · FIGURE 1

Every change re-enters the loop



Example: customer conversations → CRM updates. Controls scale with business impact. Analyst synthesis.

Figure 1 · The operating loop for an AI workflow, from test to retest.

FOR THE VP OF IT

Manage AI like any other production service. Changes to the model, prompts, data sources, or permissions go through release management, with rigor scaled to the business impact of failure.

RESEARCH CALL · WE WANT TO HEAR FROM YOU

Evaluation practitioners: what failure did your aggregate metrics miss, and which test finally caught it?

Share your experience → mesapartners.io

More autonomy expands the security boundary

Every tool you add changes what an AI agent can reach and what damage it can cause. Each one warrants a fresh security review.

The lethal trifecta

An AI agent becomes dangerous when it has three things at once: access to private data, exposure to untrusted content (emails, documents, web pages), and a way to send information out. Combine them, and a single hidden instruction buried in a document can tell the agent to quietly send your data to an attacker, and it may comply. Remove any one of the three, and the attack fails. *(Term coined by security researcher Simon Willison, 2025.)* [8]

ANTHROPIC



STORY · THREAT INTELLIGENCE, SEPT 2026

Stolen keys from a sandbox

The lethal trifecta isn't theoretical. In its September 2026 threat report, Anthropic describes a financially motivated group that had spent years breaking into hotel booking and fintech platforms, then turned its attention to the AI industry itself.

The target was an AI vendor's automated evaluation sandbox, the kind of test environment most companies think of as low-stakes. But this one held production API keys for several model providers, and it read content from outside. The attackers planted instructions in that content, and the sandbox handed over its keys.

They didn't stop there. With working credentials in hand, the group launched attacks on roughly thirty AI companies in about four days. Their goal, access to an unreleased model, was never reached. But no model was hacked along the way. All it took was a test system holding real keys next to text an attacker could write.

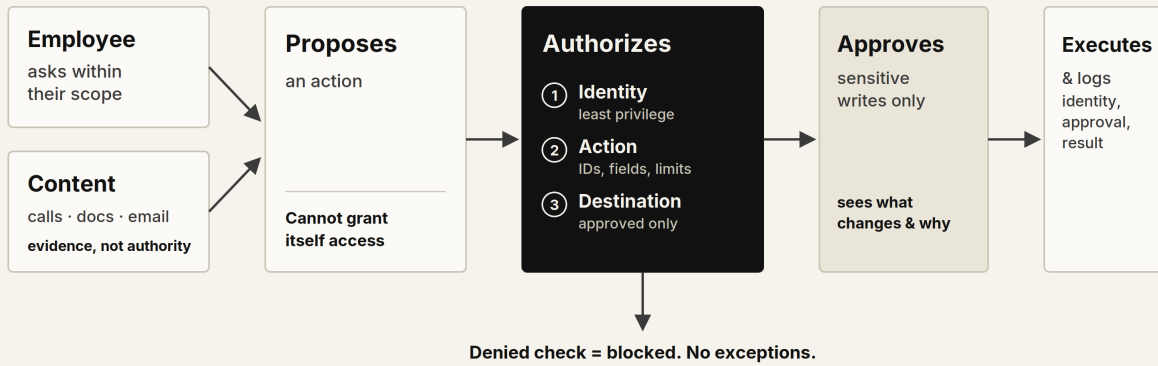
Anthropic's advice is blunt: treat AI keys and agent integrations with the same seriousness as production credentials, because attackers already do. [10]

Anthropic's engineering guidance recommends the simplest architecture that works; a fixed workflow with one AI classification step is often enough. More autonomy should earn its operating cost. [9]

For security and ITSM: restrict each agent's identity, data, and destinations. Tie those permissions to a named service owner and a tested way to suspend access. Agree which events trigger investigation unexpected exports, new destinations, unauthorized writes and who responds. Recovery covers both the stolen credential and the workflow that exposed it.

WHAT BREAKS NEXT · FIGURE 2

The AI proposes. Your systems authorize.



A sentence inside a recording or document can never expand what the agent is allowed to do.

Permissions are enforced by apps, APIs, and identity systems — outside the model. Analyst synthesis.

Figure 2 · Keep permission to act outside the model.

WHAT BREAKS NEXT · FIGURE 3

Approve the workflow, not just the app



OWNED JOINTLY BY

👤 **Security** 👤 **ITSM** 👤 **Compliance**

TRACK approved adoption · shadow AI · data exposure · business value

Adapted from Palo Alto Networks (2024), pp. 7–9 [11]. Ownership and recovery steps: analyst synthesis.

Figure 3 · Security, ITSM, and compliance share one inventory and one set of controls.

Where the controls would have stepped in

Run the Anthropic sandbox attack back through these controls and it breaks at almost every stage. The evaluation sandbox held production keys for several providers, when a test system needed, at most, a scoped test key. Restricting each agent's identity and data would have left the attackers with nothing worth stealing. Even with a key in hand, restricting destinations would have blocked the sandbox from sending credentials anywhere outside approved endpoints, removing the third leg of the lethal trifecta.

Keeping permission to act outside the model (Figure 2) matters most here. The planted instructions worked because the sandbox could do whatever the text told it to. When permissions live in the surrounding system instead of the model's judgment, a hijacked prompt can ask for the keys, but the system won't hand them over.

If something still slips through, the next layer is detection. An unexpected export and a new destination are exactly the events your triggers watch for, and a named service owner with a tested kill switch can suspend access in minutes, not after thirty companies have been hit. Recovery then covers both halves: rotating the stolen keys and fixing the workflow that exposed them, so the same injection can't work twice.

And a shared inventory (Figure 3) answers the question most organizations can't: which test systems are quietly holding production credentials?

RESEARCH CALL · WE WANT TO HEAR FROM YOU

Security specialists: what permission or tool-design change reduced exposure without making the workflow unusable?

Share your experience → mesapartners.io

05 · HUMAN REVIEW

Human review can become the bottleneck

“Human in the loop” is not a design. Who reviews? What evidence do they see? How long can a request wait? Can they reject the recommendation without restarting the task?

A review queue needs capacity planning and an escalation path. If AI drafts faster than people can check, you've accelerated one step while growing unfinished work and cost.

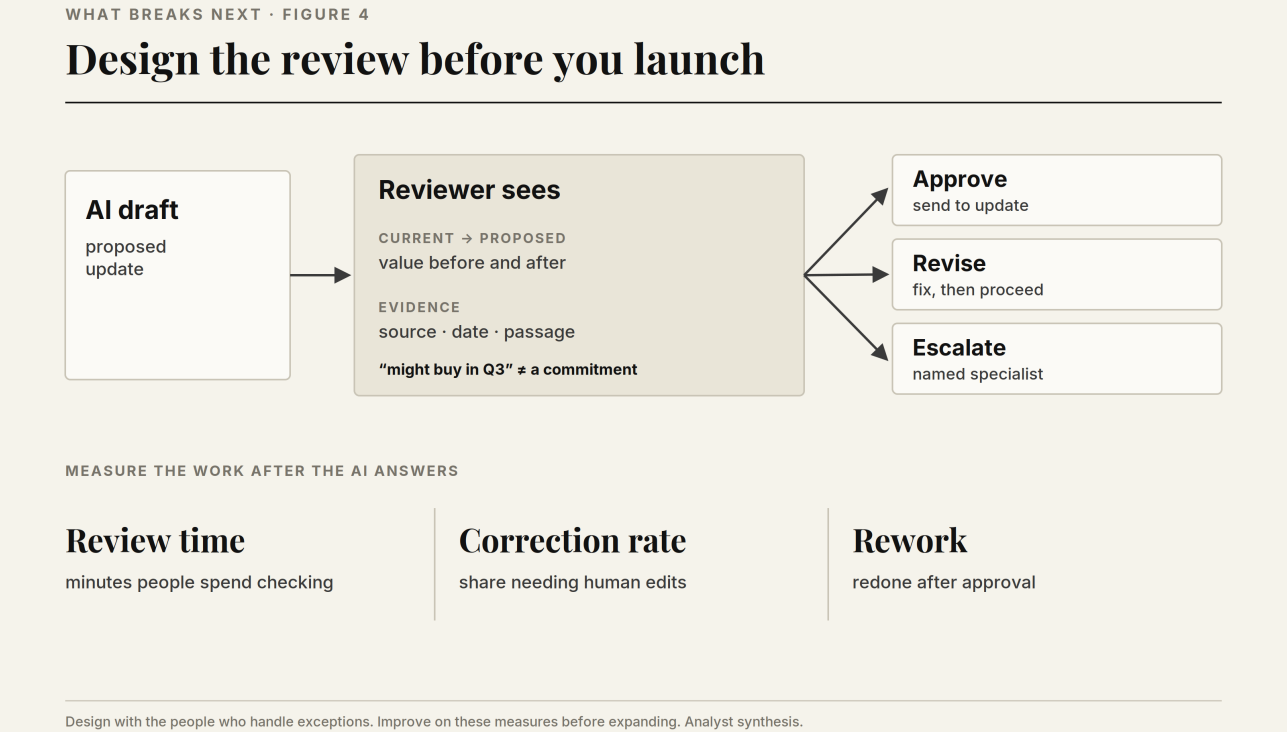


Figure 4 · A decision-ready review, and the three measures that show whether it's working.

06 · GOVERNANCE

Governance frameworks move into the workflow

When a workflow combines an external model, a vendor’s agent, an integrator’s configuration, and internal data, no single supplier’s assurance covers the whole. Someone must investigate failures that cross those boundaries. Established bodies are responding each for a different job:

FRAMEWORK	USE IT FOR	REF
ITIL AI Governance v5	Service-management oversight; AI capability and governance-improvement models	[12]
COBIT for AI (ISACA)	Linking AI to business objectives, resources, and risk across the lifecycle	[13]
NIST AI RMF + GenAI Profile	Structuring risk work (voluntary)	[14]
ISO/IEC 42001	Requirements for a certifiable AI management system	[15]
CSA AI Controls Matrix	Detailed control catalog	[16]
OWASP GenAI Security	Application-level threats and mitigations	[17]

Adopting one does not satisfy the others.

THE PRACTICAL TEST

Before approving any workflow, can you name its owner, permitted actions, evaluation evidence, escalation route, and conditions for suspension? If not, the policy leaves production to improvise.

07 · CAPABILITY

Governance is becoming a profession

Security manages access, ITSM manages change, and legal interprets obligations but someone must connect them across workflows and data. AI governance professionals turn policy into decisions: which uses are acceptable, what evidence is required before launch, and who acts when systems cause harm.

5

functions leading AI governance today: privacy, legal & compliance, IT, data governance, security

IAPP & Credo AI, 2025 [18]

AIGP

IAPP's governance credential proof of foundations, not of working controls

IAPP [19]

That makes governance a career extension for experienced staff as much as a new hire. A practical build model:

SOURCE	BEST FOR
Internal lead	Authority across business, technology, security, and compliance. Risk acceptance, service ownership, and oversight stay inside.
Specialist advisers	Defined gaps: an assessment, a technical evaluation, program design.
Larger firms	Rollouts spanning business units, jurisdictions, or platforms.

WHERE PRACTITIONERS CONNECT

COMMUNITY	GOOD FOR	REF
IAPP	Professional development, AIGP credential, AI Governance Center	[19]
Responsible AI Institute	Organizational practice and independent assessment pathways	[20]
IAGA	Standards and assurance across the AI value chain; free individual tier	[21]

AIGN Global Network	Peer network with its own frameworks and trust labels (not a standards body)	[22]
Government AI Association	Public-sector perspective — confirm current programs	[23]
The Future Society	Independent policy research (not a membership body)	[24]

Participation supports a program; it doesn't replace enterprise accountability.

RESEARCH CALL · WE WANT TO HEAR FROM YOU

Governance practitioners: when a workflow crosses suppliers and internal teams, what evidence must travel with it and who investigates a failure?

Share your experience → mesapartners.io

08 · ECONOMICS

The economics when work is measured



Different workflows and methods not comparable ROI or proof of cash savings.

Speed isn't savings. In McKinsey's banking customer-service example, human oversight is 70–75% of variable running cost; model tokens are 20–25%. Cutting exceptions and making review easier may be worth more than a cheaper model. [3]

WHAT BREAKS NEXT · FIGURE 5

The model bill is the small part

Human oversight
risk & functional experts



Model tokens
inference usage



Share of variable running cost · McKinsey banking customer-service example [3]

THE NUMBER THAT MATTERS

**Cost per
accepted outcome**

=

build + run + human review + repair

outcomes that meet the quality bar

Example-specific ranges, not a universal TCO split; fixed costs excluded. Formula: analyst synthesis.

Figure 5 · Where the running cost actually goes — and the measure that captures it.

Build the business case around a completed outcome: onboarding, resolved incident, approved record update. Count infrastructure, integration, monitoring, review, and repair against outcomes that meet your quality bar. EY recommends spend ceilings, caps on calls and retries, and a named owner for agent economics. [26]

What breaks next? A successful pilot becomes an expensive service when its exception queue grows faster than its useful output.

RESEARCH CALL · WE WANT TO HEAR FROM YOU

Organizations: will you share before-and-after costs, quality measures, and whether the gains held up?

Share your experience → mesapartners.io

09 · RECOMMENDATIONS

Five decisions before you scale

1

Name two owners.

A business owner for the outcome; a service owner for the running system.

2

Write the permission map.

What the AI may read, what it may change, and what needs a human.

3

Test outcomes, not answers.

Agree which failures block release before you start testing.

4

Rehearse the stop.

Security, ITSM, engineering, and compliance agree how to detect, suspend, and restore — and practice it.

5

Price the whole job.

Staff the review queue. Measure cost per accepted outcome, including repair.

CONCLUSION

From demonstration to dependable service

Moving from pilot to production makes AI a shared operating responsibility. The gains at Doctolib and Novo Nordisk show why enterprises are investing; the cost of human oversight shows why speed alone can't make the business case. What breaks next will usually surface at a handoff between systems, suppliers, or people. Making those handoffs visible, and measuring their effect on quality, effort, and cost, is how a successful demo becomes a dependable service.

NEXT IN THE SERIES

Our next phase of research examines that transition with practitioners: what failed after rollout, what teams changed, and whether the gains held up. Get in touch at mesapartners.io.

SOURCES

Sources

-
- [1] McKinsey, “AI at work but not at scale,” Dec 10, 2025. Survey of 1,993 participants. [Link →](#)
 - [2] Anthropic, The 2026 State of AI Agents Report, pp. 25–26. Vendor-published customer results. [Link →](#)
 - [3] McKinsey, “Where AI agents pay off: A practical guide to the economics of agentic workflows,” Aug 2026. [Link →](#)
 - [4] Morgan Stanley, Debrief launch announcement, June 26, 2024. Company-reported. [Link →](#)
 - [5] Gong, AI Data Extractor documentation. [Link →](#)
 - [6] Gong, “Security best practices for the Gong Revenue AI OS.” [Link →](#)
 - [7] Hamel Husain, “Your AI Product Needs Evals,” Mar 29, 2024. [Link →](#)
 - [8] Simon Willison, “The lethal trifecta for AI agents,” June 16, 2025. [Link →](#)
 - [9] Anthropic, “Building effective agents,” Dec 19, 2024. Vendor engineering guidance. [Link →](#)
 - [10] Anthropic, Threat Intelligence Report, Sept 2026. [Link →](#)
 - [11] Palo Alto Networks, “The C-Suite Guide to GenAI Risk Management,” Oct 2024, pp. 7–9. Basis for Figure 3. [Link →](#)
 - [12] PeopleCert, ITIL AI Governance Version 5, syllabus and certification description. [Link →](#)
 - [13] ISACA, “Leveraging COBIT for Effective AI System Governance,” 2025. [Link →](#)
 - [14] NIST, AI Risk Management Framework and Generative AI Profile. [Link →](#)
 - [15] ISO, ISO/IEC 42001:2023, AI management systems. [Link →](#)
 - [16] Cloud Security Alliance, AI Controls Matrix v1.1, June 22, 2026. [Link →](#)
 - [17] OWASP GenAI Security Project. [Link →](#)
 - [18] IAPP & Credo AI, AI Governance Profession Report, 2025. [Link →](#)
 - [19] IAPP, membership, AI Governance Center, and AIGP credential. [Link →](#)
 - [20] Responsible AI Institute, programs and membership. [Link →](#)
 - [21] International AI Governance Association (formerly AI Assurance Club), history, framework, and membership. [Link →](#)
 - [22] AIGN Global Network, community and membership. [Link →](#)
 - [23] Government AI Association, mission and activities. [Link →](#)
 - [24] The Future Society, mission and areas of work. [Link →](#)
 - [25] Brynjolfsson, Li & Raymond, “Generative AI at Work,” Quarterly Journal of Economics, 2025 (earlier as NBER WP 31161). [Link →](#)
 - [26] EY, “Five shifts reshaping the agentic AI economy,” Aug 2026, p. 3. [Link →](#)
-

About this research. What Breaks Next is a Mesa Capital Partners research series on what fails when AI and Physical AI move into production. Research and drafting were assisted by AI tools (ChatGPT, Claude); all claims are linked to sources. Vendor-published results are labeled as such.