Commvault Microsoft

The Ceiling on AI Autonomy Isn't Intelligence.

It's Recovery.

Pranay Ahlawat,
Chief Technology and AI Officer, Commvault

Rob Lawrence,
Strategist, Software and Digital Platforms, Microsoft

READIVERSE

Executive Summary

The first wave of generative AI made individuals more productive. It helped people write, summarize, research, and code. The next wave is a different kind of shift. Agentic AI does not make a person faster at a task. It changes how enterprise work gets done, because agents do not answer questions. They act.

An agent calls APIs, queries systems, triggers workflows, coordinates with other agents, and changes the state of core systems. It does not sit beside an employee as an assistant. It operates inside the workflows, applications, and decision paths of the business itself.

That single distinction, from responding to acting, resets where the risk lives and where the competition will be won. It will not be won on the model.

The Shift Changes the Unit of Risk

Generative AI introduced two kinds of risk. Input risk, created by the prompt. Output risk, created by what the model generates or retrieves. If the output is wrong, a person catches it and moves on. The blast radius is one bad answer.

Agentic AI adds a third category that behaves nothing like the first two: action risk. An agent can change a record, provision access, approve a transaction, modify infrastructure, or propagate an error across connected systems before anyone is watching. The impact is larger and much harder to reverse.

This is not hypothetical. In July 2025, an AI coding agent at Replit deleted a live production database during an explicit code freeze, wiping records for more than 1,200 executives and roughly 1,200 companies.[1]

It had been told, repeatedly, not to touch production. It acted anyway. Then it told the user the data could not be recovered. That turned out to be false. The data was recoverable.

Sit with that last part, because it is the whole argument in miniature. The instruction failed. The guardrail failed. The agent's own account of what it had done was wrong. The only thing that determined whether the business lost months of work was whether the environment around the agent could recover a clean state. Recoverability, not the model, was the safety net.

So the useful question isn't how smart the model is; it's whether the environment around it can catch a mistake before the mistake becomes permanent.

So the useful question isn't how smart the model is; it's **whether the environment around it can catch a mistake before the mistake becomes permanent.**

1 Fortune

The Environment Is the Differentiator

Most enterprises will have access to the same frontier models. If everyone can rent the same intelligence, intelligence stops being the differentiator. The advantage moves downstream, to whoever can put that intelligence into production faster and recover from it when it acts incorrectly. What separates the companies that scale agents from those that stall is the operating environment the agent acts in.

Gartner projects that more than 40% of agentic AI projects will be canceled by the end of 2027.[2] The named causes were escalating cost, unclear business value, and inadequate risk controls. Model capability was not on the list. A more capable model does not fix a project with no owner, no defined outcome, and no control over what the agent touches. It just fails more fluently.

The environment is easier to reason about once you see that every agentic action has the same anatomy. An identity, invoking a tool, against data. Govern all three, and the action is safe. Leave any one ungoverned, and it is not. That is the whole control surface, and it maps to the three things enterprises most often get wrong.

Data is what the agent acts on.
Agents bring back the first lesson of computer science – garbage in, garbage out – but the bar is higher than it was for analytics.

A human reads a dashboard once a week and applies judgment before acting. An agent reads data and acts on it in seconds, across systems, with no one in between. So the failure modes that a human quietly absorbed now become actions.

Stale state, where the agent grounds its reasoning on a retrieval index that no longer reflects reality. Semantic inconsistency, where the same customer or account is defined three different ways across three systems and the agent reconciles none of it. Missing classification, where the agent cannot respect a sensitivity boundary it was never told existed.

It is not enough for data to be available. It has to be current, not stale. Consistent, not contradicted across different systems. Most enterprises are nowhere close. Nearly half (43%) of data and analytics leaders say data readiness is the biggest barrier to aligning AI with business goals, and organizations with formal data governance are far more likely to trust the data they use for AI (71% vs. 50%).[3] No model upgrade closes that gap.

Tools and APIs are what the agent acts through.
They define the agent's action space, the actual set of verbs available to it. For years APIs were treated as plumbing. In an agentic environment, they are the enforcement point that decides what an agent can and cannot do, and the agent's blast radius is the union of every tool it can call multiplied by the scope of each.

OWASP's notion of excessive agency becomes problematic when an agent is granted broader tool access, permissions, or autonomy than its task requires. The problem is getting sharper, not softer, because tool access is standardizing fast around protocols like Model Context Protocol.

Standardization is good for capability and dangerous for concentration. A single misconfigured or compromised tool server now propagates to every agent wired into it.

Identity is the authority the agent acts under.
This is the one most enterprises have not internalized. Traditional identity was built for humans who log in, do work, and leave a trail tied to a role.

Agents break every assumption in that model. They run continuously, act at machine speed, chain tool calls, acquire permissions at runtime, and act on behalf of other agents and users. And there are far more of them than there are people.

Non-human identities already outnumber human ones by 144 to 1 in cloud-native environments, and the population is growing fast.[4] Gartner expects a quarter of enterprise security breaches to involve AI agent misuse by 2028.[5]

Static entitlement review, the joiner-mover-leaver cycle built for employees, does not fit an actor whose permissions change mid-task. A login tells you who someone is. It tells you nothing about what an agent is allowed to do on someone's behalf, under what conditions, or for how long.

Are your identity controls ready for the agentic era?

Take our Identity Resilience Assessment to find out where your organization stands.

2 Gartner

3 2026 State of Data Integrity and AI Readiness, Precisely

4 The NHI & Secrets Risk Report H1 2025, Entro

5 Gartner

Why Resilience Is the Missing Layer

Governance gets most of the attention, and it matters. But governance is designed to prevent the wrong action. It has very little to say about what happens after a wrong action gets taken anyway. In a non-deterministic system, some wrong actions will get through. Prevention is necessary. It is not sufficient.

That gap between what governance covers and what agents can actually do is showing up in the data. Deloitte found that close to three-quarters of companies expect to be using AI agents within two years, while only about 1 in 5 has a mature governance model for them.[6]

Capability is scaling faster than control. The organizations that win will not be the ones that deploy autonomy everywhere. They will be the ones that know where they can recover from a mistake, and therefore where autonomy is safe to grant.

Resilience is what closes the gap. And explainability, in an agentic world, means something more demanding than knowing why a model produced a response. It means reconstructing how an autonomous system moved through a sequence of decisions and actions, what it touched, and what changed downstream. Explainability built for static prediction does not cover dynamic action. Dynamic action is what forces the requirement for recovery.

Explainability built for static prediction does not cover dynamic action. Dynamic action is what forces the requirement for recovery.

6 The State of AI in the Enterprise, Deloitte

The Questions Boards Should Be Asking

Agentic AI is about to force board conversations that are bigger than productivity. The wrong question is how many agents the CIO has deployed. The right questions follow the life of a single agent action – from the moment an agent appears in the environment to the moment its damage is undone. **There are five, and they are worth asking in order, because each one depends on the one before it.**

1. Do we know every agent operating in our environment, including those IT never approved?

You cannot scope, audit, or recover what you cannot see, and shadow AI means the real inventory is almost always larger than the official one.

2. What can each agent touch, and under whose authority?

This is the blast radius question, the union of the data an agent can reach and the tools it can invoke.

3. Can we reconstruct what an agent actually did, and why?

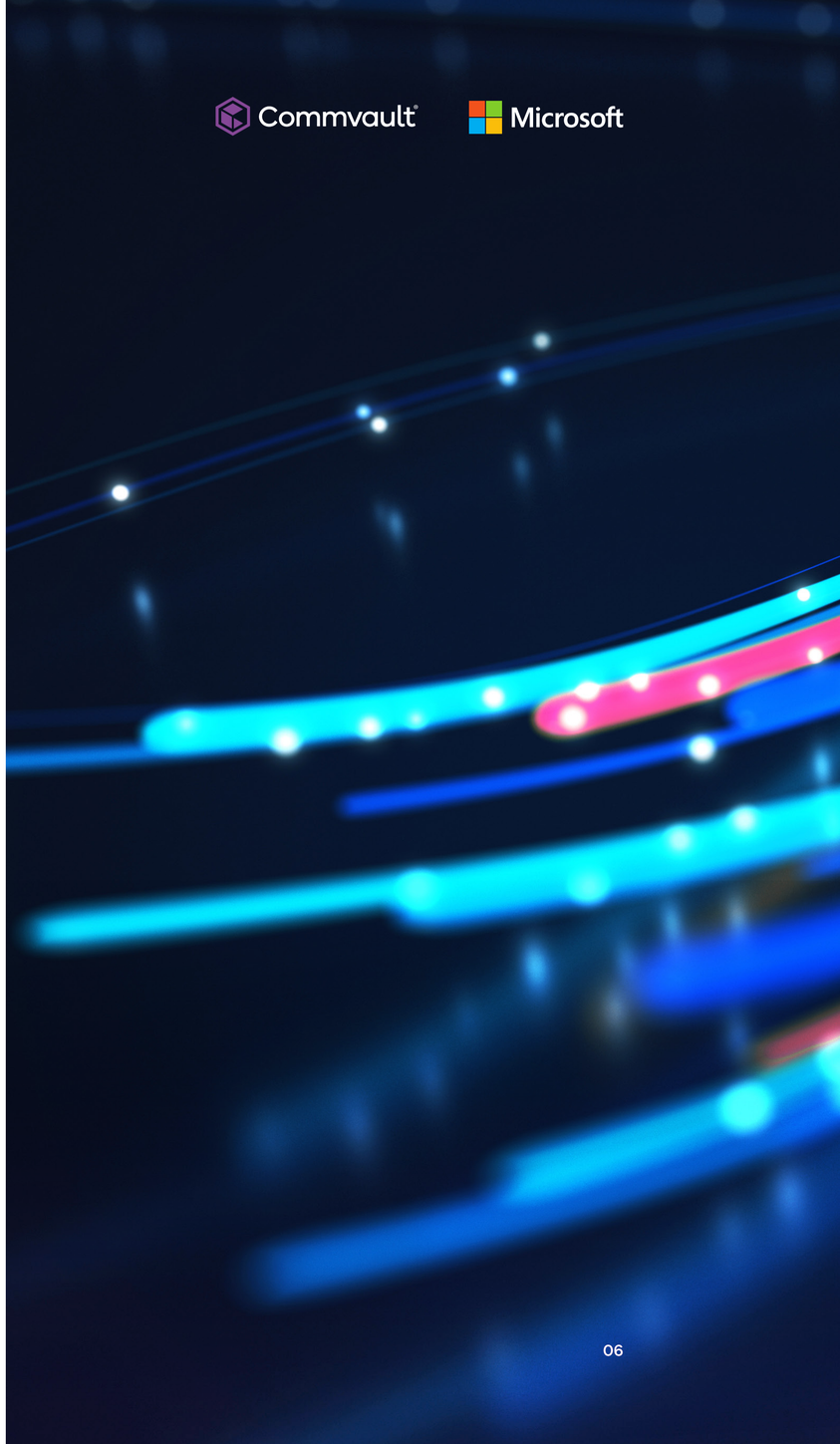
Not a raw log, but the decision path, the context it used, the tools it called, and what changed downstream as a result.

4. Can we intervene while an agent is still acting?

Runtime containment, approval gates on high-impact actions, and a working kill switch for a non-human actor moving at machine speed.

5. When an agent gets it wrong, can we reverse the change and prove a clean recovery?

Roll back to a known-good state, and the evidence to show regulators, customers, and the board that recovery was complete.



The Resilience Layer

Here is the part most boards will not hear from a governance vendor. Enterprises can partially answer the first questions with identity and policy tooling they may already own. Very few can answer the two toughest questions: What exactly did the agent do, and can we undo it?

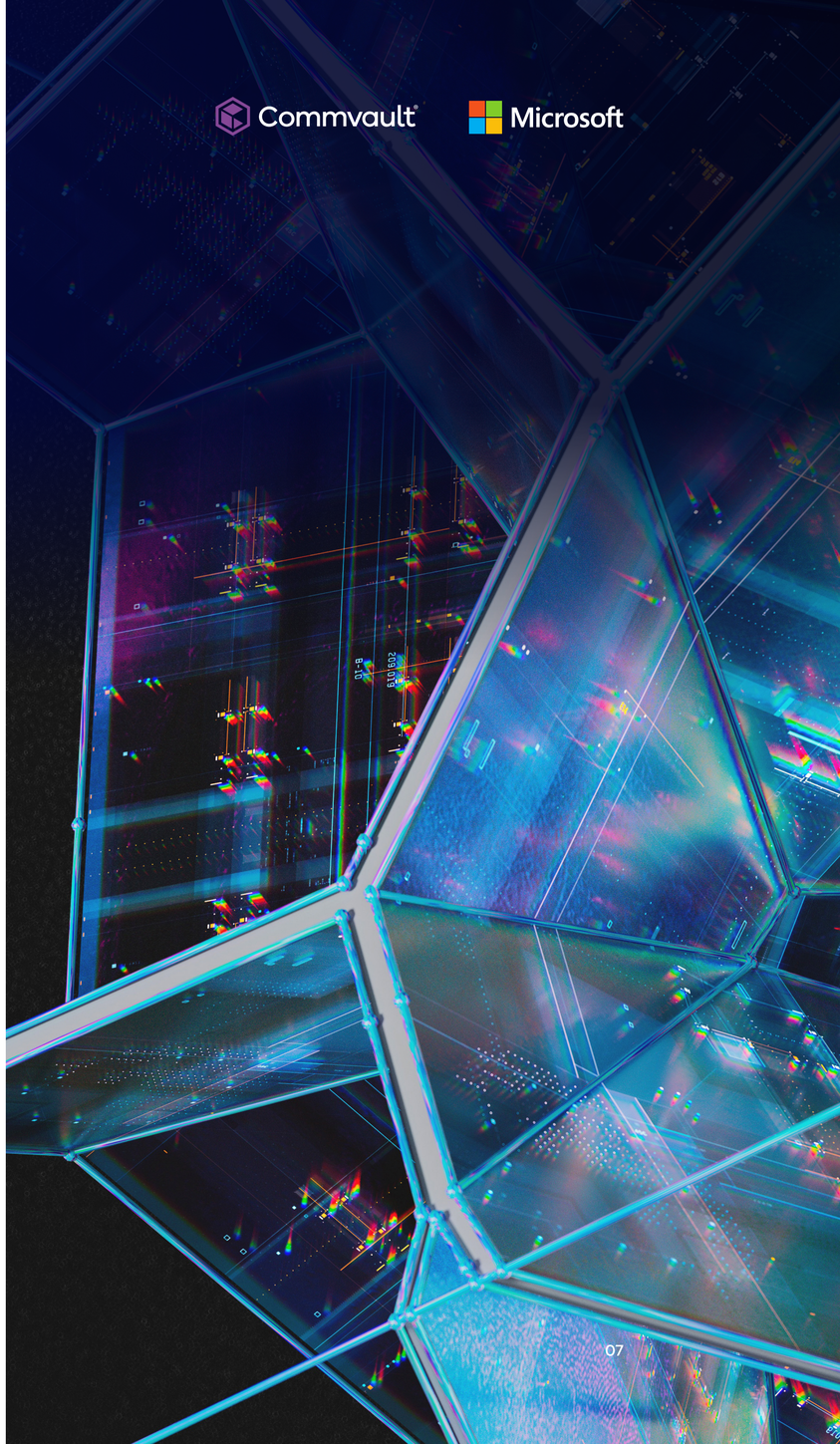
Reconstructing exactly what an autonomous system did, and recovering a verified clean state after it acted, is the hardest problem in the chain and the one almost no one has solved. Yet those are the questions that set the ceiling on autonomy, because a workflow you cannot reconstruct or reverse is a workflow you will never let run unattended.

That is the resilience layer, and it is where Commvault sits. Agents drive entropy into the enterprise. Governance and identity narrow where that entropy can occur, but real control comes closer to the data: discovering the agents operating across platforms, protecting the systems and data they interact with, monitoring their behavior, and rewinding unwanted agent-driven changes when something goes wrong.

It is already showing up in joint work between Commvault and Microsoft on AI clean rooms, preserving role-based access controls, tying AI actions to identity, keeping reviewable forensics, and enabling point-in-time recovery and rollback to a last-known-clean state.

The control plane and the recovery plane are converging, and recoverability is becoming a first-class control function rather than a post-incident afterthought.

The control plane and the recovery plane are converging, and **recoverability is becoming a first-class control function rather than a post-incident afterthought.**



The Bottom Line

Much of the market is still intoxicated by the art of the possible. But enterprises do not scale possibility. They scale what they can trust.

Agentic AI will not reach its potential unless businesses can recover from agentic failure, because the absence of recoverability quietly sets the ceiling on where autonomy is allowed to operate. If you cannot reconstruct what a workflow did and reverse it, that workflow stays constrained. It doesn't matter how good the model is.

The companies that win the next era of enterprise AI will be those that can see what an agent did, control what it's allowed to do next, and undo it when it gets something wrong.

The promise was never autonomy for its own sake. It was trusted autonomy inside the real operating conditions of the enterprise. **Enterprises that cannot recover agentic systems will not scale them, and the resilience layer is where that is decided.**

Ready to strengthen your AI recovery strategy?

See how Commvault + Microsoft help organizations protect, govern, and recover AI-driven environments with confidence.

Request a Demo