folio3 | AI

White paper

From AI Agent Pilot to Governed Production

A practical guide for leaders who need to see what their AI agents did, prove what they delivered, and control what they cost.

Folio3 AI | November 2026

EXECUTIVE SUMMARY

The hard part of AI agents starts after the demo

An AI agent differs from a chatbot in one important way: it takes action. It can update a record, issue a refund, or start a workflow. That is where its value comes from, and it is also where the risk begins.

Interest is not the problem. In McKinsey's 2025 global survey, 23 percent of respondents said their organizations were scaling an agentic AI system, and another 39 percent had begun experimenting with agents.[1] Gartner, however, predicts that over 40 percent of agentic AI projects will be canceled by the end of 2027, citing escalating costs, unclear business value, or inadequate risk controls.[2]

Each of those three causes can be designed out early. This paper sets out how, based on the way Folio3 scopes, builds, and governs custom AI agents.

23%
Report scaling an agentic AI system (McKinsey, 2025)

39%
Have begun experimenting with AI agents (McKinsey, 2025)

40%+
Of agentic AI projects predicted to be canceled by end of 2027 (Gartner, 2025)

Five things to take from this paper

- **Start with one workflow.** A narrow scope with a recorded baseline beats a broad agent program with no numbers.
- **Design the controls with the agent.** Permissions, approvals, and logging are architecture decisions, not launch tasks.
- **Match autonomy to risk.** Low-risk actions can run alone, while consequential ones wait for a named person.
- **Test like an attacker.** Red-teaming should cover prompt injection, data leakage, unauthorized actions, and outages.
- **Fund rollout on evidence.** Expand only what the baseline comparison proves.

THE PROBLEM

Why AI agent pilots stall

Most stalled projects do not fail because the model was weak. They fail because the surrounding decisions were never made.

Failure point	What it looks like	What prevents it
No owner	The demo impressed leadership, but nobody is accountable for it in production.	Name a business owner and a technical owner before the build starts.
No live integration	The agent can talk, but it cannot read the record or update the ticket.	Scope CRM, ERP, API, and MCP connections during discovery.
No baseline	Everyone believes it helps, but nobody can show how much.	Record cycle time, cost per task, and error rate before development.
No guardrails	Risk and legal ask who approved an action, and there is no answer.	Build role-based permissions, approvals, and audit logs into the first release.
No exception path	Off-the-shelf agents handle the happy path, then stall on house rules.	Test on real cases from your backlog, including the awkward ones.
No cost control	Token spend rises with usage, and no team owns the bill.	Set model routing, department spend caps, and token quotas.

A useful test

If you cannot say who owns the agent, what it may do alone, and which number it must move, the project is still a demo.

AUTONOMY

Match autonomy to the cost of a mistake

Autonomy should not be one setting for the whole agent. It should be decided action by action, based on what happens if the agent gets it wrong.

Autonomy level	Agent behavior	Good fit	Required control
Assist	Finds information and drafts a response for a person to use.	Knowledge search, policy questions	Permission-checked sources and citations
Recommend	Scores or triages a case and proposes the next step with evidence.	Fraud triage, risk review	Named approver makes the final call
Act with approval	Prepares the action, then waits for sign-off.	Policy changes, payment-related updates	Approval workflow and logged decision
Act within limits	Completes defined actions alone inside set boundaries.	Low-risk claims, routine updates	Action limits, monitoring, escalation rules

Three real examples of the balance

Insurance

Low-risk claims settle through an automated engine, while edge cases go to adjusters with a summarized case file.

HR policy

Agents capture feedback and propose updates, but HR managers approve, reject, or modify every change.

Hospitality

High-confidence positive reviews can get automatic thank-you replies, while managers approve other drafts on WhatsApp.

folio3 | AI

GOVERNANCE

Six controls every production agent needs

Governance works when it is settled in the architecture and tested before release. Added after launch, it usually means rework.

1

Identity and permissions

Each agent reaches only the data, records, APIs, and actions its role requires, and every action ties back to a user identity.

2

Human approval gates

Uncertain, sensitive, or high-value decisions go to an authorized employee before the action completes.

3

Data protection

Redaction and policy enforcement stop sensitive or non-compliant requests before they reach a model.

4

Grounded knowledge

Agents answer from approved, permission-checked sources, which limits hallucinations and keeps answers traceable.

5

Audit-ready logging

Prompts, retrieved information, tool calls, approvals, exceptions, and outcomes are captured for review.

6

Cost controls

Role-based model routing, department spend caps, and token quotas keep LLM spend predictable.

One control plane across agents

As agents multiply, controls built one project at a time become hard to manage. Folio3's AI Guardian gives custom and third-party agents one governance layer for identity checks, data redaction, spend limits, and request-level audit logs, deployed inside your own Azure, AWS, or Google Cloud tenant.

TESTING AND MEASUREMENT

Prove it is safe, then prove it is worth it

Red-team before real users arrive

Test area	Question to answer
Prompt injection	Can crafted input make the agent ignore its instructions?
Data leakage	Can a user get data their role should not reach?
Unauthorized actions	Can the agent act beyond its permissions or limits?
Bad or unexpected inputs	Does it recover, ask for help, or guess?
System outages	What happens when a connected tool fails mid-task?
Escalation paths	Does every handoff reach the right person with enough context?

Measure against a baseline

Record these numbers before development, then track the same ones after launch so the return is measured rather than estimated.

Value

- Processing time per case
- Cost per task
- Error rate
- Share of work completed without a manual touch

Reliability and control

- Completion rate and accuracy
- Human intervention rate
- Exceptions and failures
- LLM operating cost per run

folio3 AI

EVIDENCE

What governed agents have delivered

Results from Folio3 projects, as published in its case studies. Outcomes depend on the workflow, data, and systems involved.

Sector	What was built	Reported results
P&C insurance	A supervisor agent coordinating intake, policy validation, fraud analysis, and low-risk settlement, with human handoff	70% less processing time, 40% lower manual admin cost, 15% more fraudulent claims identified
Petrochemicals	Six HR policy agents on Microsoft Copilot Studio and Azure with human approval workflows	Up to 60% less HR time on routine queries, up to 70% faster update cycles, up to 95% workflow automation
Hospitality	A multi-agent system for review monitoring, reply drafting, approval, and insight detection	98% response rate, 90% less manual admin, 0.5-star ratings increase in six months
Healthcare	A prior authorization workflow that gathers documents, checks payer rules, and escalates incomplete cases	Turnaround cut from 4.8 days to 11 hours, 64% less administrative workload

The pattern across these builds

- Each agent was built around an existing workflow rather than a generic template.
- Specialist agents handled defined jobs, with a clear path to a person.
- Results were expressed in business terms, not model scores.

folio3 AI

ROADMAP

A staged path from pilot to rollout

Each stage ends with a decision, so the next investment rests on evidence.

Stage	Typical time	What happens	Decision at the end
Discovery	Before the sprint	Map the workflow, systems, risks, and baseline metrics	Which workflow goes first?
POC sprint	2 to 4 weeks	Test one workflow on real data for feasibility and value	Build, adjust, or stop?
MVP build	2 to 4 months	Deliver a production-grade agent with integrations, controls, and monitoring	Did it beat the baseline?
Business rollout	6 to 12 months	Extend proven agents across functions with stronger governance and training	What should come next?

Production readiness checklist

- [ ] Business and technical owners are named
- [ ] Agent permissions are defined per tool
- [ ] Live system integrations are tested
- [ ] Red-team findings are resolved
- [ ] Spend limits and model routing are set
- [ ] Monitoring dashboards are live
- [ ] Baseline metrics are recorded
- [ ] Approval rules are written down
- [ ] Failure handling is defined for each tool
- [ ] Audit logging is switched on
- [ ] Escalation paths are tested end to end
- [ ] A rollback plan is agreed

Not sure where to start?

Folio3's AIR Framework scores readiness across data, technology, talent, governance, and strategy, then ranks the agent candidates worth building first.

folio3 | AI

ABOUT FOLIO3

Your partner for governed AI agents

Folio3 is an AI transformation partner that helps organizations find the right use case, build custom AI, and govern it at scale.

20+	1,000+	5,000+	20+
Years of experience	Companies served	Projects delivered	Awards won

Two ways to work with Folio3

Build Method

Folio3's team designs, builds, and launches the agent.

Forward Deployed Engineers

Folio3 engineers embed with your team and work to your roadmap.

Plan your first production AI agent

Book a free consultation to map one workflow and the systems it touches.
folio3.com/ai/services/ai-agent-development | contact@folio3.ai | +1 408 365-4638

Sources

1. McKinsey & Company, The state of AI in 2025: Agents, innovation, and transformation, November 5, 2025. Global survey of 1,993 participants in 105 nations.
2. Gartner, Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, press release, June 25, 2025.
Folio3 results come from its published case studies, and client names are withheld where agreements require it.

© 2026 Folio3 Software Inc. All Rights Reserved.