



BURAK GALİBA
Thoughts on AI Transformation

ESSAY · ENTERPRISE AI

Shipped Is Not the Same as Working — Who Checks Your AI Agent on Day 90?

Launch testing proves an agent worked once. Most enterprises have no way to notice when it stops — no owner, no living eval set, no incident path. What "in production" should actually mean.

AGENTIC AI

GOVERNANCE

LEADERSHIP



Burak Galiba · October 4, 2026 · 6 min read

<https://burakgaliba.com/blog/shipped-is-not-working>

SCAN TO READ ONLINE

Most of what I've written about AI agents is about getting them over the line: why [the demo isn't the finish](#), why [pilots pile up](#), why [agents stall at IT](#). This one is about what happens after the line — because the day an agent goes live is the day most organizations stop looking at it.

The launch goes well. There was a test set, the answers were checked, the sponsor signed off, the announcement went out. Ninety days later someone in the business mentions, in passing, that they've stopped using it — the answers "got a bit off" a while ago. Nobody can say when. Nobody was paged, because nothing broke. There was no outage, no error, no ticket. The agent kept answering the whole time, fluently and confidently, and it was wrong for weeks before anyone said so out loud.

That is the failure mode I'd worry about most in enterprise AI right now, and it's the one almost nobody is instrumented to see.

Software fails loudly. AI fails quietly.

Everything about how enterprises run production systems assumes failure is visible. A service goes down, a dashboard turns red, someone gets woken up. Monitoring is built around availability and errors because, for conventional software, a wrong answer is a bug you find once and fix once.

An agent doesn't work like that. It can be up, fast, error-free and wrong — and the wrongness doesn't arrive as an event. It drifts in. Three things cause most of it:

- **The content underneath moves.** The policy was revised, the price list changed, a product was retired. The agent is still answering from what it was given, and [documents rot faster than anyone maintains them](#). Nothing in the agent changed; the world it describes did.
- **The model underneath moves.** Providers update models, retire versions, and change behavior between releases. This autumn alone saw a wave of new models at sharply lower prices, and every one of those is an invitation to switch. Switching to save money is reasonable. Switching without re-running your evaluation set is how an agent that was right in June becomes subtly different in October, with nobody having decided that it should.
- **The questions move.** Users find uses nobody designed for. The test set covered what the builders imagined; production traffic covers what people actually need. The gap between the two widens every week the agent is popular.

None of these trip an alarm. All of them erode trust — and trust, once users quietly route around a tool, is very expensive to win back.

You can't monitor what you can't list

Before monitoring, there's a more embarrassing question: how many agents are actually running in your organization, and who owns each one?

In most enterprises, nobody can answer that. Agents were built in workshops, inside licensed platforms, by business units and by IT, some officially and many not. This autumn the vendor market started selling agent inventories as a product category — software whose main job is to tell a company what AI it is already running. When a problem becomes a product, it's usually because it became common first. If you need to buy a tool to find out what agents you have, the governance gap came well before the tooling gap.

The inventory doesn't need to be clever. A list with four columns gets you most of the way: what the agent does, what data it reads, who owns it, and when it was last evaluated. The last column is the one that tells you whether anything after this point is happening at all.

Three things every production agent needs

I'd argue an agent isn't in production — whatever the launch email said — until it has all three of these.

A named owner who is not the builder. The person who wrote the prompt or the integration moves on to the next project; that's their job. The owner is someone in the business who is accountable for the agent's answers being right, the way a process owner is accountable for a process. If a regulator, a customer or an executive asks "why did it say that?", this is whose phone rings. An agent whose owner is "the AI team" has no owner.

A living evaluation set. Launch testing is a snapshot. What you need is a set of real questions with known-correct answers that gets re-run on a schedule and every time something underneath changes — new model version, new source documents, new prompt. And it has to keep growing: the best source of new test cases is production itself. Every complaint, every escalation, every answer a user flagged as wrong becomes a test case. The [hundred-question set I suggested building in a workshop](#) is where this starts; the discipline is that it never stops.

An incident path and a kill switch. Decide in advance what happens when the agent gets something wrong that matters: who users report it to, who triages, what the threshold is for pulling it, and who has the authority to pull it without convening a meeting. Agents increasingly *do* things — send, update, approve — rather than just answer, and an agent with write access and no off switch is not a product. It's an exposure.

None of this is exotic. It's how mature organizations already run anything that affects customers or money. The only new part is accepting that "running without errors" is no longer evidence of "working."

Measure the thing that actually degrades

When teams do monitor agents, they usually measure what's easy: volume, latency, cost per query, thumbs-up rate. Those are worth having, but they're mostly measures of activity. Volume can climb while quality falls, because people keep asking until they get something usable. Thumbs-up rates are dominated by the few users who bother to click.

The measures that move first when quality slips are less glamorous:

- **Eval pass rate over time**, re-run on the same set — the closest thing you have to a quality trend line.
- **Escalation and override rate** — how often a human had to step in or correct the output.
- **Abandonment** — conversations that end without the user getting what they came for, and users who simply stop coming back.
- **Source freshness** — the age of the documents the agent is actually citing.

Pick a baseline at launch, the same way you'd [baseline the business case](#), and review it monthly with the owner. A falling eval score you catch in week two is a maintenance task. The same drop discovered in month four is a credibility problem.

The cost nobody budgets for

Here's the uncomfortable part: all of this is ongoing work, and almost no AI business case includes it. The project plan funds the build and the launch. The run cost line has the API bill and maybe some hosting. The owner's time, the eval maintenance, the monthly review — none of it is costed, so none of it is staffed, so none of it happens.

That's not a side detail. If [IT shouldn't own the value number](#), the business can't own the value without owning the upkeep. An agent nobody maintains doesn't stay at launch quality; it decays toward the day someone notices, and that day usually arrives through a complaint rather than a dashboard.

Where to start

Pick the one agent your organization relies on most. Then answer four questions about it, honestly, today: who owns its answers, when was it last evaluated against a fixed test set, what would you see if it started getting worse, and who could switch it off this afternoon.

If any of those answers is "nobody," "at launch," "nothing," or "we'd need a meeting," you've found your next piece of work — and it isn't building another agent.

I built a free [AI Readiness Score](#) that asks a version of these questions alongside 19 others — 20 questions across pilots, data, talent and governance, about ten minutes. The governance section scores exactly this: whether anything is watching your AI after the launch email goes out.

Read the essay online at <https://burakgaliba.com/blog/shipped-is-not-working>

I built a free AI Readiness Score — 20 questions across pilots, data, talent and governance, about ten minutes, no email required: <https://burakgaliba.com/readiness>

© 2026 Burak Galiba · <https://burakgaliba.com> · Views are my own.