



PLAYBOOK

When an Agent Goes Wrong: A Leader's Playbook for Spotting, Stopping and Fixing AI Agent Mistakes

Six steps for the day an AI agent does something it's technically allowed to do — but shouldn't. How to find every agent you run, decide what "wrong" looks like, build one off switch that works everywhere, practise the stop, staff the work agents hand back, and turn every near-miss into a lesson the board can see. Written for C-suite and directors in the UAE and Saudi Arabia.

21 October 2026

PEOPLE FOR TOMORROW

What's inside

AI agents rarely fail with a crash. More often they do something they're technically allowed to do — but shouldn't: the wrong price, the wrong customer, the wrong tone. Most companies would spot it eventually. Far fewer could stop it quickly, everywhere it runs, and keep the work moving while they fix it. This playbook is about closing that gap.

-
- 01** Find every agent, including the ones IT didn't build
 - 02** Decide what "going wrong" looks like
 - 03** Build one off switch that works everywhere
 - 04** Practise the stop before you need it
 - 05** Staff the work agents hand back
 - 06** Learn from every incident, and report it upwards
-

01

STEP 1 OF 6

Find every agent, including the ones IT didn't build



You can't stop an agent you don't know exists. And agents now get built everywhere — in productivity suites, CRM and HR systems, and by business teams on their own. In Dataiku's survey of 685 CIOs worldwide, 84% agreed employees are creating agents faster than IT can govern them, and 81% lack complete oversight of agents built outside approved systems.

Start with one list. Ask every department head what agents their team runs, then check each platform you pay for. For each agent, write down four things: who owns it, what it does, what systems and data it can touch, and where it runs.

Keep the list alive. New agents should be added before they go live, and retired ones removed. A list that's six months old is a list of what used to be true.

► **One list of every agent, with an owner and what it can touch. Start there.**

02

STEP 2 OF 6

Decide what "going wrong" looks like



Most agent problems aren't technical failures. In the UAE, 80% of CIOs in the Dataiku survey had seen an agent break business rules or expectations while still working exactly as built. Nothing crashed. The work was just wrong.

So define "wrong" in business terms. For each agent, write down the red lines: the customers it must never contact, the amounts it must never approve, the data it

must never open. Google puts it as four questions every company should be able to answer about any agent: who is it, what is it allowed to do, what did it do, and what should it never touch?

Then decide what counts as a warning sign — a spike in volume, a run of customer complaints, a change in tone. Those are what your people and your monitoring should watch for.

- ▶ **Write down what "wrong" means in business terms, not just technical ones.**

03

STEP 3 OF 6

Build one off switch that works everywhere



Spotting a problem and stopping it are different things. Only 5% of UAE CIOs said they could reliably find and contain a problem agent across their systems within one to two hours. For 45%, it would take one to two days. Only 18% have a single, joined-up view of all their agents.

The fix is partly technology and partly authority. On the technology side, make sure every agent can be paused from one

place, across every system it touches — many platforms now offer this, such as hard spending caps that pause an agent automatically. On the authority side, name who is allowed to pull the switch, and make sure they don't need three approvals to do it.

Write both into each agent's record: how to stop it, and who can.

► **Aim to stop any agent, everywhere, within hours — not days.**

04

STEP 4 OF 6

Practise the stop before you need it



An off switch you've never tested is a theory. Run a short drill for your most important agents: write a realistic scenario where the agent does something allowed but wrong, then time how long it takes to notice, decide and stop it everywhere.

Bring the right people into the room: the business owner, IT, and whoever would talk to customers. Most drills uncover something

simple — a password nobody has, a system the agent also runs in that nobody remembered, or a decision that needs someone who's on leave.

Fix the slowest step and run it again every quarter, and after any big change to the agent. Each drill should be faster than the last.

► **Run a drill every quarter. Time it. Fix the slowest step.**

05

STEP 5 OF 6

Staff the work agents hand back



When an agent stops, the work doesn't. Someone has to pick it up. Gartner's research points out that agents push the hardest parts of the job — unclear cases, judgement calls and fixing mistakes — back to people. Yet only 24% of leadership teams agree on who is accountable for their agents.

Plan the human side for each agent: who handles its escalations day to day, and who does its work if it's switched off for a week.

Check they have the access, the training and the time. With leaner teams and fewer junior roles, there's often less slack than leaders assume.

The same question applies to anything you outsource. In Saudi Arabia, HUMAIN and EY are building outsourcing that blends agents and people. Whoever runs the process, the contract should say who handles the exceptions.

- ▶ **Every agent needs a human backup plan: people with the time, access and skills to take over.**

06

STEP 6 OF 6

Learn from every incident, and report it upwards



Keep a simple log of every agent incident and near-miss: what happened, how it was spotted, how long it took to stop, and what changed afterwards. Review it monthly with the agent owners.

Then share the headlines with your leadership team and board. Boards are already asking: every UAE CIO in the Dataiku survey reported more board pressure to prove AI returns than last year. Showing that you catch problems early and stop them fast is part of that proof — near-misses caught in time are a sign the system works.

Expect this to become normal. The UAE's National Committee for Agentic AI has approved an operating framework and guide for the federal government's agent rollout. Public-sector partners and regulators across the Gulf will increasingly expect the private sector to show the same discipline. And where you need people to staff oversight and recovery faster than you can hire, bring in specialists on a flexible basis while your own team builds up.

► **Every near-miss is free training. Log it, learn from it, and show the board.**



PEOPLE FOR TOMORROW

None of this needs a new department. It needs a full list of your agents, a clear idea of what "wrong" looks like, one off switch that works everywhere, a drill to prove it, and people ready to take over when an agent stops. Put those in place now, and an agent mistake becomes a short, managed incident — not a crisis.

Read the playbook → 20309062.fs1.hubspotusercontent-na1.net/hubfs/20309062/TASC_Outsourcing/TASC%20AI%20Newsletters/Edition%2010/TASC-Report-Edition-10.pdf