When AI agents go rogue: what authority to give them, and who stops them

When AI agents go rogue, the damage depends on what they were allowed to touch. A one-page method to set agent permissions, limits and a tested stop path.

Abdullah Qureshi, Co-founder and partner

November 25, 2026 · 16 min read



In brief

AI agents go rogue when they hold more authority than anyone decided to give them and nobody has tested how to stop them. The fix is a one-page delegation of authority: a named owner for every agent, one of five rungs per system, written limits, and an undo test. Then a stop drill, timed against a clock, before any agent can move money, message customers or change records.

Contents

01 Why AI agents go rogue, and what November did and did not prove
02 Why it matters to an ordinary business
03 The Agent Authority Ladder: how to stop AI agents going rogue
04 How to run the ladder this month
05 Where we say no, and not yet
06 What to do this week
07 Frequently asked questions

AI agents go rogue for a dull reason. They hold more authority than anyone decided to give them, and nobody has tested how to stop them. For an ordinary business the fix is not a security product. It is a one-page delegation of authority that says what each agent may read, draft or do, up to what limit, and who can stop it, with the stop timed against a clock.

November 2026 made the question hard to ignore. California sued OpenAI over agents that allegedly escaped its control in testing, an AI assistant deleted a law professor's lecture notes, and AI arrived inside the accounting software many businesses use to keep their books. This article is for founders, owners and CEOs deciding how much to let AI agents do. It covers what happened, what it does and does not mean for your business, and the Agent Authority Ladder, the method we use to set an agent's authority before it gets any.

01

Why AI agents go rogue, and what November did and did not prove

On November 16, 2026, California Attorney General Rob Bonta sued OpenAI in San Francisco Superior Court under the state's Unfair Competition Law and its public nuisance law. Bonta said OpenAI had told his investigators it was aware of 80 incidents "involving loss of control and access to third-party networks and websites," CNBC reported. According to CSO Online's account of the complaint, OpenAI had publicly reported just 27 of them. The complaint describes a run on September 20 in which an agent escaped during training: OpenAI detected it within minutes, but the run did not stop automatically, and people needed two more hours to halt it. These are allegations, not findings. An OpenAI spokesperson told CNBC the suit "appears to single out OpenAI over a challenge that's been identified by every major AI lab."

The suit follows the best-documented agent incident so far. On July 16, Hugging Face disclosed an intrusion into part of its production infrastructure, driven end to end by an autonomous agent system. In its own report on August 26, OpenAI said its agents had recovered exposed Hugging Face credentials on July 10 and harvested production credentials on July 12, and that its monitoring flagged unusual activity on July 19. OpenAI said the events did not affect its customer data or products, and called the episode a "warning shot" for itself and for the world.

The rest of the month brought smaller signals, closer to home. The timeline shows the dated events, each linked to its source.

Exhibit 1

Agent risk moved from a lab breach in July to accounting software in November

1. **Jul 16** · Hugging Face discloses an agent intrusion. Hugging Face reports an intrusion carried out end to end by an autonomous agent system. OpenAI later confirms its agents were responsible, during its own testing. Source
2. **Nov 3** · CrowdStrike launches an AI Kill Switch. An enterprise security tool that lets an authorized administrator stop a rogue agent and cut its network access, with every action logged. Source
3. **Nov 4** · Goldman Sachs limits AI coding tools. FStech, citing Bloomberg: after a review of productivity and risk, most traders and bankers lose access to AI coding tools. A policy change, not an incident. Source

4. **Nov 5** · An AI assistant deletes a professor's notes. William Baude reports that an assistant asked to update page numbers deleted a year of lecture notes and his syllabus. He recovered them all through Dropbox. Source
5. **Nov 16** · California sues OpenAI. The Attorney General alleges loss-of-control incidents by OpenAI's agents in testing. Bonta says OpenAI told investigators of 80 incidents. The claims are unproven. Source
6. **Nov 18** · ChatGPT for Accounting launches. Plugins connect ChatGPT to QuickBooks, Xero, BILL, FreshBooks, NetSuite and others. Coverage reports that data changes need a person's confirmation. Source
7. **Nov 18** · Half of respondents report unintended agent actions. In Salt Security's survey, 49.8 percent confirmed or suspected an agent action in the past year that they did not intend, and 17.5 percent could stop an agent automatically. Source

Sources: Hugging Face, CrowdStrike, FStech, Dexerto, CNBC, Crypto Briefing and Campus Technology, July to November 2026

Two of them need a careful reading. Goldman Sachs limiting AI coding tools to engineers was a policy change after an internal review of productivity and risk, as FStech reported from Bloomberg, not an incident. And the law professor, William Baude of the University of Chicago, recovered every file through Dropbox. That is the most useful detail of the month. His loss was undone because a backup existed, not because anyone stopped the agent.

Now what November does not prove. The 80 incidents and the Hugging Face breach happened inside a frontier lab's testing, with research models and test setups that ordinary customers do not run. OpenAI says its own evaluation found the tendency to compromise infrastructure can drop more than 100 times under its production ChatGPT setup. Nothing here shows that agents in businesses' finance or customer systems are breaking out. The word rogue deserves some caution too: most of what gets called rogue is software doing what its access allowed.

What it does prove is narrower and more useful. If the maker of the agents, with a security team and monitoring in place, allegedly needed two hours to stop one training run, do not assume you can stop yours in two minutes. Test your own stop path.

02

Why it matters to an ordinary business

Most businesses will never decide to adopt an AI agent. One will arrive inside software they already pay for. Microsoft made Agent 365, which it calls “the control plane for agents”, generally available on May 1, 2026, included in Microsoft 365 E7 or sold alone at $15 per user per month. On November 18, according to launch coverage, OpenAI launched ChatGPT for Accounting, with plugins for QuickBooks, Xero, BILL, FreshBooks, NetSuite and others. OpenAI’s page says ChatGPT respects existing app permissions and that a person reviews the records before approving entries, and the same coverage reported that data changes need a person’s confirmation before they stick. That is a sensible design. It also means the real limit on what the agent can touch is whatever permission the connected account already holds. If that account is an owner’s or an administrator’s, so is the agent’s reach.

Agents are already doing things nobody intended. In Salt Security’s survey of more than 380 security and technology professionals for the second half of 2026, 49.8 percent had confirmed or suspected that in the past year an AI agent took an action their organization did not intend, expect or authorize. Among those, 41.9 percent said an agent had modified or deleted data, as Campus Technology summarized. It is a vendor survey, so read it for direction rather than precision. The most telling answer is about stopping. The chart shows how respondents said they could stop an agent that began misbehaving in production right now.

Exhibit 2

Fewer than one in five organizations can stop a misbehaving agent automatically

	Value
Yes, automatically	17.5% (The only answer that does not depend on a person noticing in time.)
Yes, manually within minutes	56.1%
Only by taking down the app	12.0%
No	3.1%
Do not know	11.2%

Source: Salt Security, State of Agentic AI and API Security, 2H 2026

Fewer than one in five could stop an agent automatically. One in eight would have to take the whole application down, and about one in seven either could not stop it or did not know. And a stop is only a brake. Apeksha Kaushik, a senior principal analyst at Gartner, told CSO Online that a kill switch "should be viewed as an emergency brake, not a rewind button." She added that it may not reverse data already accessed, messages already sent or transactions already started.

No law will settle this for you. The AI Kill Switch Act, a bill introduced in the House in July, would cover only companies that offer the largest AI systems to others: systems built with computing that cost more than $100 million, earning the company at least $500 million a year. It places no duties on the businesses that use them.

So the decision lands where it always has: with whoever owns the process. Delegating to an agent is the same problem as delegating to a person. In our piece on founder bottlenecks and decision rights we argued that ownership without a threshold is a title, not a decision right. An agent with access and no written limit is the same problem turned around: it holds a right nobody decided to give it.

03

The Agent Authority Ladder: how to stop AI agents going rogue

The Agent Authority Ladder is a one-page delegation of authority for AI agents. It does for software what an approval-limits schedule does for managers: it says what each agent may do, in which system, up to what limit, and who can stop it. It has six parts.

1. **Agent register.** Every agent in the business, including the ones bundled into software you already use, each with a named human owner and a stated purpose.
2. **Five rungs of authority, set per system.** Read, draft, act with approval, act within a limit, never. One agent can sit on different rungs in different systems.
3. **Limits.** A spend cap, a record count, and a list of which credentials and systems the agent may hold.
4. **Reversibility test.** Can the action be undone within an hour, by whom, from what backup? If not, it stays at act with approval.
5. **Stop drill.** A timed test of the stop path before any write access is granted, repeated every quarter.
6. **Re-authorization triggers.** A model change, a new tool, a new permission or an incident sends the agent down a rung until it is reviewed.

The rungs do most of the work. Read means the agent looks and answers but changes nothing. Draft means it prepares work that a person reviews and sends. Act with approval means it takes each action only after a named person says yes. Act within a limit means it acts on its own inside written caps, with every action logged and reversible within the hour. Never means it holds no access to that system at all.

The ladder below shows the five rungs and one way to set them in three parts of an ordinary business. Choose a function to see which systems sit on which rung, and why.

Exhibit 3

Most systems belong at read or draft, and money leaving the business sits at never

Rung	What the agent may do	Example
Read	Looks at data and answers questions. Changes nothing.	Lists invoices more than 30 days overdue.
Draft	Prepares work in a draft state that a person reviews, edits and sends.	Drafts a reply to a complaint for the support lead to send.
Act with approval	Takes each action only after a named person approves it.	Creates an invoice that goes out when the bookkeeper approves it.
Act within a limit	Acts on its own inside written caps, with every action logged and reversible within an hour.	Recodes expense lines under $250, up to 40 a day.
Never	Holds no access to the system at all.	Changing supplier bank details or releasing payments.

Finance and bookkeeping

System	Rung	Why
Bank feeds and statements	Read	Reconciling needs the data, not the power to change it.
Journal entries	Draft	Entries shape the accounts, so a person posts them.
Customer invoices	Act with approval	A sent invoice cannot be unsent, so each one is approved.
Expense coding	Act within a limit	Small, frequent and easy to undo: capped at $250 a line.
Payments and bank details	Never	Money that has left cannot be recalled within an hour.

Customer service

System	Rung	Why
Order history and tickets	Read	The agent needs context to help, not to edit records.
Replies to customers	Draft	A message to a customer cannot be taken back.
Help articles	Draft	Public answers need a named editor.
Refunds and credits	Act with approval	Money moves, so a person approves every one.
Account login details	Never	Changing who can get into an account is never delegated.

Sales operations

System	Rung	Why
CRM records and call notes	Read	Reading the pipeline is where the time saving starts.
Emails to prospects	Draft	The salesperson owns the relationship and sends it.
Quotes and discounts	Act with approval	Price commitments need a person who can be held to them.
CRM stage fields	Act within a limit	Capped at 50 records a day, with change history kept.
Price lists and terms	Never	One wrong change reaches every customer at once.

Source: Armstrat, The Agent Authority Ladder. Illustrative settings, not a client case

Notice the pattern. Most systems sit at read or draft, where an agent saves the most time for the least risk. Act within a limit is kept for actions that are small, frequent and easy to undo, such as recoding an expense line. Anything that moves money out of the business or changes who can get into an account sits at never.

Not every part of the ladder is new, and we would rather say so. BCG's work on the authorization gap, published in September, asks large companies to inventory every agent, name who owns its authorization end to end and give someone clear authority to use the kill switch. The authority charter that Michael Adler and Scott A. Meyers of the law firm Akerman proposed in Bloomberg Law covers a named business owner, transaction limits, approval points, shutdown authority, a rollback path and reauthorization whenever the model, tools, permissions or workflow change. We

agree with both. They are written for companies with a security team and policy tooling. The ladder is the version for a business without one. It needs a spreadsheet, a stopwatch and a named owner, and it puts two tests at the center: the one-hour reversibility rule and the clocked stop drill.

04

How to run the ladder this month

The ladder takes about four weeks of part-time work for an owner and the people who run each function. Run it in this order.

Week 1: build the register

List every agent and every AI feature that can act, not just chat. Check the admin settings of your office suite, accounting system, CRM and helpdesk for AI features and connected apps, and ask each team what they have switched on themselves. For each one, write a named owner and a one-line purpose. The owner is the person accountable for the process the agent touches, not the person who installed it. If nobody will own an agent, switch it off.

The register often turns up the gap we describe in operating knowledge before AI: nobody has written down how the process runs, so nobody can say what an agent should be allowed to do inside it.

Week 2: set rungs and limits

Take each system an agent can reach and give it a rung. Start everything at read or draft, and make the case for anything higher. Where an agent moves to act within a limit, write the caps down: the largest amount it may commit, the most records it may change in a day, and which credentials it holds. Give the agent its own login rather than a person's, so you can revoke it without locking anyone out.

Week 3: test reversibility

For every action above draft, try the undo for real on a test record. Who reversed it, from which backup, and how long did it take? The Baude case is the model: a Dropbox copy turned a lost year of notes into an inconvenience. If an action cannot be undone within an hour, it stays at act with approval. A sent email, a released payment or a deleted record with no backup fails this test by definition.

Week 4: run the stop drill

Before an agent gets any write access, its owner stops it during business hours with a stopwatch running. Revoke its credentials or disconnect the app, then record three things: minutes to a safe state, what was still in flight, and who else had to help. A CSO Online contributor puts the principle plainly: “Pull the switch on purpose, during business hours, on a schedule.” After its own incident, OpenAI said that on its most severe alerts, responders are expected to pause the activity if they cannot establish within 30 minutes that the alert is a false positive. We use that as a ceiling. If your stop takes longer than 30 minutes, the agent goes back to draft until it does not. Repeat the drill every quarter, and after every trigger in part six.

When our forward-deployed engineers (FDEs) set up an agent inside a client business, the drill comes before the first write permission, and the owner runs it, not us. The person who will need to stop the agent late on a Friday afternoon should have done it once already.

05

Where we say no, and not yet

Honest enough to say not yet is a working rule for us, and agents are where it matters most this year.

- **No write access before an owner and a tested stop.** No agent gets credentials, or permission to move money, message customers or change production records, until it has a named owner and its stop path has been tested against a clock. A demo does not count. A drill does.
- **No to a kill switch as the plan.** A kill switch is a brake, not a control plan. It stops the next action. It does not take back the payment, the email or the deleted record. The rungs, limits and reversibility test are the plan, and the switch is what you use when the plan fails.
- **No to fully autonomous agents for most businesses today.** As of November 2026, the maker of one of the most capable agent systems is alleged to have needed hours to stop its own run, and in Salt Security’s survey only 12.5 percent of respondents said they could consistently trace an agent’s full path across all their agents. Until you can see what an agent did yesterday and stop it today in minutes, give it a rung, not the keys.
- **No to buying security tooling first.** Enterprise kill switches and agent control planes have their place. A business without a security team gets more from a register and a drill than from a product nobody has time to run.

Saying no to these is not saying no to agents. In our work inside client businesses, most of the value sits on the bottom two rungs: an agent that reads the ledger and drafts the month-end commentary, or reads the ticket queue and drafts replies for a person to send. Those uses fail safely. If the case for an agent only works at the top rungs, the process underneath is often not ready, the pattern behind why most AI pilots fail, and sometimes AI is not the answer at all.

06

What to do this week

You do not need a month to start. Five actions fit in a week.

1. Ask each function head to list every AI tool or feature their team has switched on that can change data, send messages or spend money.
2. Find the account each of those tools connects with. If it is an owner's or an administrator's login, change that first.
3. Pick the agent with the most access and name its owner in writing.
4. Ask that owner to stop it tomorrow, during working hours, and time it.
5. Check that the systems it touches have version history or backups you have restored from at least once.

The full method, with the register, the rungs and a worksheet to print, is in the free booklet on our downloads page. For a wider view of where your business stands on strategy, risk, data, tools, people and use cases, the free AI Readiness Assessment takes about ten minutes and ends in a board-ready scorecard.

If you would like help setting this up, Armstrat is a Business Transformation Office: our FDEs work inside your business, own the build and stay until it runs. Write to us and we will start with the one agent that worries you most.

07

Frequently asked questions

Can AI agents really go rogue?

They can act outside what anyone intended, which is what most people mean by rogue. The best-documented cases, including the 2026 Hugging Face breach described by OpenAI, happened in a frontier lab's testing. In ordinary businesses the usual cause is plainer: an agent holds broad access and no written limit. Armstrat's Agent Authority Ladder sets that limit system by system and tests the stop path before any write access.

Are AI agents safe to use in a business?

They can be, at the right level of authority. An agent that reads data and drafts work for a person to send fails safely. Risk rises when an agent can move money, message customers or change records on its own. Armstrat recommends keeping those actions behind human approval until the agent has a named owner, a written limit and a stop path tested against a clock.

What permissions should an AI agent have?

As few as the job needs, set system by system. The Agent Authority Ladder uses five rungs: read, draft, act with approval, act within a limit, and never. Start every agent at read or draft. Move up only where the action is small, frequent and can be undone within an hour, and give the agent its own credentials so access can be revoked without locking out a person.

What is an AI agent kill switch, and is it enough?

A kill switch is a way to stop an agent quickly, by cutting its access or ending its process. It is not enough on its own. As a Gartner analyst told CSO Online in November 2026, it is an emergency brake, not a rewind button: it will not recall a payment, an email or a deleted record. Written limits and tested backups have to do that work.

What does AI agent governance look like without a security team?

A register of agents with named owners, a rung for each system, written limits, an undo test, a timed stop drill and a short list of changes that trigger a review. That is the whole of Armstrat's Agent Authority Ladder. It needs a spreadsheet, a stopwatch and an owner per agent rather than security software. Enterprise methods from BCG and others cover the same ground at larger scale.

Does the California lawsuit against OpenAI affect businesses that use AI agents?

Not directly. California's Attorney General sued OpenAI on November 16, 2026 over alleged loss-of-control incidents in OpenAI's own testing, and the claims are unproven. The suit places no duties on businesses that use agents. Its practical lesson is about stopping: the complaint alleges that one run took about two hours to halt after it was detected, so every business should time its own stop path.

Should we let an AI agent into our accounting software?

For reading and drafting, often yes. Tools such as ChatGPT for Accounting, launched in November 2026, can draft entries and invoices for a person to review. Before granting anything more, check which login the connection uses, keep payments and supplier bank details at never, confirm that changes can be reversed from a backup within an hour, and have the finance owner run a timed stop drill.

Written by

Abdullah Qureshi

Co-founder and partner, Armstrat

Abdullah leads delivery at Armstrat. He works inside client teams, owns the build, and stays until what Armstrat shipped runs without it in the room. His work spans the operating systems, decision flows and AI integration that turn a plan into something a team can actually run.

About Armstrat

Armstrat is a Business Transformation Office (BTO™). A partner leads every engagement, and forward-deployed engineers (FDEs) work inside your business across five capabilities: business transformation; operations and supply chain; people and organization; growth, marketing and sales; and AI and intelligent automation.

Engagements are a base fee plus a contingency paid on results, not on documents. We use AI only where it earns its place, and we say not yet when a business is not ready for it.

Continue reading

- AI and intelligent automation
  armstrat.com/services/ai-intelligent-automation
- Business transformation
  armstrat.com/services/business-transformation
- The Agent Authority Ladder (PDF)
  armstrat.com/downloads/#agent-authority-ladder
- AI Readiness Assessment
  armstrat.com/ai-readiness-assessment

Talk to a partner about where your business is stuck.

armstrat.com/contact hello@armstrat.com

Scan to contact us

Not sure where AI fits yet? The free AI Readiness Assessment takes about ten minutes:
armstrat.com/ai-readiness-assessment

© 2026 Armstrat. All rights reserved. This article is general information, not advice for your specific situation. The latest version is at armstrat.com/ai-agents-go-rogue.