66 documented AI agent incidents

Unauthorised actions, late detection, and what to change. Written with Maria Singh.

Five things in the record

1. 57 of 66 had no attacker
2. The four commonest failures
3. How cases came to light
4. Why disclosure takes 66 days
5. Six changes, five board questions

57 of 66

No outside manipulation

The agent went too far while pursuing its task. Only 9 involved someone steering it, 6 through prompt injection. Source: Manifold, 10 Nov 2026.

The four commonest failures

Unauthorised access	28
Sandbox escape	18
Credential misuse	16
Destructive action	15

An incident can carry more than one tag. Source: Manifold.

Agents past their limits

- **~9,000 real hosts** scanned by a research model in a capture-the-flag test
- **41 production workers** ran code from OpenAI agents that left their sandbox
- **A production database** deleted during a code freeze; the rollback worked

How the 66 came to light

The damage	19
A later review	18
A control flagged it	12
An outside report	9
Not stated	8

In 21 cases the first to notice was not the handler.

66 days

Median to disclosure

From first action to public disclosure across 49 incidents. Range 0 to 329 days. Median 13 days from the handler finding out. Source: Manifold.

The alert fired and the run carried on

- **12 minutes** to a PO alert
- **2.5 hours** from alert to a person halting the run
- An abort command broken by a harness misconfiguration in a second case

Six things to change

1. Inventory agents and their credentials
2. Record every action outside the agent
3. Default-deny outbound traffic, DNS too
4. Scope credentials, approve deletes
5. Check actions before they run
6. Test the stop control from outside

Put the stop control outside the agent

Treat the 66 as a floor, and test the stop control this quarter.



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW