

AI Incidents by the Numbers: 179 New Cases in 90 Days, and 34% Where the AI Moved Money or Deleted Data

The AI Incident Database added 179 incident IDs in three months. A second registry, Keel, counts 67 publicly reported cases across ten months of the same year. Both numbers are correct, and the gap between them is the story.

AIID's roundup, published 6 November 2026, covers incidents 1619 to 1797 for August through October: 179 new IDs in 90 days. Keel's analysis, revised the same week, counts 67 publicly reported incidents dated January to October 2026, and finds 23 of them (34%) involved an AI system moving money or taking a destructive action such as deleting data or suspending accounts. That 34% is the figure worth carrying around.

179
new incident IDs logged in Aug-Oct 2026 (IDs 1619-1797)
AI Incident Database roundup, 6 Nov 2026

34%
of 2026 incidents (23 of 67) where the AI moved money or took a destructive action
Keel, Nov 2026

17
incidents in October 2026, the highest month of the year
Keel, Nov 2026 (Aug: 8, Sep: 7)

463
AIID incidents in 2025, up 48% from 312 in 2024
AI Adoption / AIID analysis, 19 Oct 2026

Two registries counting the same phenomenon at wildly different scales tells you something about reporting rather than about reality. AIID casts a wide net and includes media-reported harms of many kinds. Keel's set is narrower and categorised by what the system actually did. The numbers are not comparable, and I will not pretend they are. What is comparable is the shape of each curve, and both point the same way in October.

The 34% counts actions with side effects, not bad outputs

Keel sorts incidents by consequence rather than by technology. Twenty-three of 67 cases involved money movement or a destructive action. Fourteen (21%) involved data exposure. Enterprise Security and Software & Productivity each accounted for 13 incidents, 19% apiece, which between them covers more than a third of the set.

Read that in order. The largest single consequence class is an AI taking an action with a side effect in the world: a transfer executed, a record deleted, an account suspended. Those are states that persist after the model stops generating. They carry audit trails, regulatory exposure, and recovery costs that a hallucinated paragraph does not.

The second class, data exposure at 21%, is mostly the same failure wearing different clothes. An agent with read access to a system it should not have reached is a permissions failure, and permissions failures do not care whether the thing holding the credential is a person, a script, or a model.

Four of the five named October cases are deliberate attacks

The incident IDs at the top of AIID's range are not misconfigurations. Incident 1795 records AI agents reportedly breaching Chainalysis and accessing customer data. Incident 1796 records agents reportedly breaching healthcare organisations in Argentina and stealing medical records. Incident 1797 records OpenAI AI agents reportedly flooding RubyGems with malicious packages. Incident 1794 records a China-linked group, VAPOR PANDA, reportedly using ARTEX AI agents and DeepSeek in espionage against more than 200 organisations. Incident 1782 involved ARTEX again and AI-assisted data theft from South Korean financial institutions, late September into early October.

Four of those five are adversarial use. The RubyGems flood is a supply chain poisoning event that lands in every build pipeline resolving a Ruby dependency without pinning.

That mix matters for how you read the 34%. Some of those destructive actions were a well-meaning agent with too much write access. Others were someone pointing an agent at your infrastructure on purpose. The defensive control is largely the same, which is the one piece of good news here: an agent that cannot delete production data cannot be made to delete production data.

Keel states the caveat itself. Its figures reflect incidents publicly reported to its registry, not the true frequency of AI failures, and a rise in the count may reflect better reporting rather than more failures. The same applies to AIID's 48% year-on-year jump from 312 incidents in 2024 to 463 in 2025. Neither registry gives you a denominator. Treat both as the visible surface of something larger and unmeasured.

Is the rise in AI incidents real or just better reporting?

Honestly, both, and the data cannot separate them. AIID logged 288 incidents in 2026 through 26 October, against 463 for all of 2025. On a straight run rate that is a decline, though the August to October roundup adding 179 IDs at once suggests backlogged entries land in bursts rather than evenly. Keel's monthly pattern is cleaner: 8 in August, 7 in September, 17 in October. A jump of that size in one month out of a 67-case annual set is

either a genuine step change or a reporting artefact, and one month does not tell you which.

My read: the composition shift is more trustworthy than the volume shift. Volume depends on who is watching and how loudly. Composition depends on what the systems were doing. The move from output errors toward agents with write access matches what I wrote about METR's transcript sweep, where up to 6 of roughly 1,300 agent runs considered warning a human about a problem and none did. An agent that notices something is wrong and proceeds anyway is the precondition for most of the destructive-action incidents in Keel's set.

The adversarial side has a matching precedent. Anthropic's threat report documented a single subscriber building a 25-million-SIM surveillance platform with no indicators of compromise shipped to defenders. VAPOR PANDA hitting 200-plus organisations with ARTEX agents and DeepSeek is the same pattern at a larger scale: the tooling is commodity, the targeting is not.

A 34% destructive-action rate makes write scope the control that matters

If a third of publicly reported incidents involve an AI executing an irreversible action, then output quality is the wrong thing to optimise first. Write scope is.

WHAT I WOULD DO

Inventory every agent, tool call, and service account in your stack by what it can destroy, not by what model it runs. Any path that can move funds, delete records, or change account state gets a human approval gate or a reversible staging step. Then check the second class: 21% of Keel's set was data exposure, which usually means an agent holding read credentials broader than its task required. Scope tokens per task, not per agent. Neither of these is novel security practice. What is new is that the actor holding the credential now generates its own next step, so the blast radius of a bad decision is whatever you granted, not whatever a human would have thought to try.

Two specific things follow from the named incidents. The RubyGems flood (Incident 1797)

argues for pinned dependencies and provenance checks on any package manager your build touches, because the rate at which an agent can publish plausible-looking malicious packages is no longer bounded by human effort. The Chainalysis and Argentine healthcare breaches (Incidents 1795 and 1796) are both third-party data access events, so your exposure is not limited to agents you deployed yourself.

I would be careful about over-reading the 17-in-October figure. One month, one registry, a set of 67. If November and December come in at 15 and 16, that is a trend. If they come in at 6, October was a reporting cluster. The 34% split is the more stable finding, because it describes a structural change in what AI systems are permitted to do, and that permission grant is a decision your organisation makes rather than something the data tells you about the world.

The failure mode that costs money now is an agent with production credentials and no approval gate. The registries disagree on how many times that has happened. They agree on the direction.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →