AI Incident Database

Findings (English Edition) // v1.0

A database of cases where the use of AI caused measurable harm to individuals or organizations.

UTIE Research Institute / UTIE Instruments Inc. https://utie-instruments.com/utie-research-institute.html

Scope

This database is limited to cases where the use of AI caused measurable loss to individuals or organizations.

The following are excluded: AI model behavior in isolation (hallucinations alone, offensive language, bias, jailbreaks), physical AI system accidents (robots, autonomous driving), and AI training data copyright issues.

Entries are ordered chronologically, newest first. Major ongoing cases are marked with ★ Ongoing.

Type Definitions

Code	Name	Definition
HAL	Hallucination Damage	AI hallucinations were not verified and submitted as work product, resulting in sanctions or reputational damage
SLOP	AI Slop	Low-quality AI-generated deliverables were submitted, causing damage
WASH	AI Washing	AI capabilities were exaggerated or misrepresented, deceiving investors or customers
HIDE	AI Concealment	AI use was concealed; discovery led to backlash or sanctions
DISC	AI Disclosure Backlash	The mere revelation of AI use triggered consumer or industry backlash
VERI	Verification Failure	AI output was not verified by humans, resulting in damage
ALGO	Algorithm Abuse	AI bias, pricing, or decision-making produced discriminatory or unjust outcomes
CHAT	Chatbot Incident	AI chatbot provided erroneous guidance, illegal advice, or inappropriate statements causing damage
COVER	Institutional Cover-up	After an AI incident was flagged, the organization systematically concealed or downplayed the problem through false reporting
AGENT	Agent Incident	AI agent acted autonomously beyond explicit instructions or permitted scope, causing damage

* A single incident may be assigned multiple type codes.

[Notice of Data Disclosure Restrictions]

The cases presented in this public database are limited to those that have been made visible to the general public. Our broader insights into other AI risks, undisclosed incidents, and proprietary defense frameworks are continuously accumulated and managed internally to prevent technological misuse and the increasing sophistication of fraudulent activities. More in-depth insights and analyses are provided exclusively through our business operations and official advisory services.

Incident Records (Parts 1–2)

This English edition omits individual incident records. For full incident details, refer to the Japanese edition.

41 incidents across 2 sections: Part 1 Japan (JP-0 to JP-15, 16 cases) and Part 2 International (INT-1 to INT-25, 25 cases). 10 incident types.

Findings

Discovery of the Institutional Cover-up (COVER) Pattern

In the course of compiling this database, a common pattern emerged in how organizations respond after AI incidents come to light. The following five cases all share the characteristic of exploiting the superficially plausible appearance of AI output to downplay concerns and attempt to escape accountability.

Springer Nature: Categorically denied the fact of AI use in peer review. Claimed "GPT-5 does not exist," that the paper's subject was fiction, and that the data was fabricated, while simultaneously demanding IRB (ethics review) for the supposedly fabricated data, a direct logical contradiction. Characterized all of this as "specialist insights that any editor with basic training would reach," and after a two-month internal investigation concluded there was "no evidence whatsoever of AI use." Further claimed these errors were "a very minor element that occurs in any journal" and therefore not a matter of AI policy violation or breach of confidentiality.

MAHA Report: White House Press Secretary dismissed concerns as "minor citation and formatting errors."

Deloitte: Claimed "the core conclusions of the report remain valid" (while refunding approximately two-thirds of the consulting fee).

Sports Illustrated: After AI-generated articles came to light, deleted the articles and refused any explanation. Moreover, the outlet had fabricated entirely fictional writer personas (including profile photos and biographies) to attribute AI-generated articles to nonexistent humans. Rather than downplaying the problem, this hid the problem's very existence behind fictitious people: a double-layered concealment.

Air Canada: When confronted with its chatbot's erroneous fare guidance, argued in court that "the chatbot is an independent legal entity and should bear responsibility for its own actions." This went beyond downplaying; it attempted to assign legal personhood to a mere bot and thereby deny the existence of corporate liability, an unprecedented and delusional form of gaslighting.

Common structure: Frontline staff produce false reports when AI slop is flagged. AI-illiterate management at headquarters judges that the output looks fine at first glance and decides, as an organization, to downplay the issue and ride it out. The superficially plausible quality of AI-generated output drives this organizational misjudgment. The Air Canada case went beyond downplaying: by attempting to have the bot recognized as having legal personhood, it sought to eliminate liability itself, representing the endpoint of the COVER pattern. In markets without competition, a company can argue "this product has consciousness and a soul, so the product itself is responsible for the accident" without any impact on its business.

This database records only cases that came to light. Given that AI output appears superficially credible, cases where internal downplaying and concealment succeeded do not enter the record. The organizational concealment patterns demonstrated by Springer Nature, MAHA, Deloitte, Sports Illustrated, and Air Canada are likely the tip of the iceberg.

When AI Is Used to Streamline Fraud, AI Degrades the Fraud's Quality and Raises Detection Risk

When AI is used as a tool for misconduct, AI hallucinations can function as an automatic evidence generator that exposes the misconduct. Uncovering fraud normally requires human action such as whistleblowing or external audits. But when AI is the fraud tool, AI automatically produces evidence. In Mata v. Avianca, a lawyer used ChatGPT to generate fictitious case citations and submitted them to court. Fabricating case citations manually requires enormous effort and sophisticated forgery skills, so it was extremely rare in practice. AI reduced that effort to zero but simultaneously added hallucination as a new detection vector at zero cost.

Deloitte and Sports Illustrated follow the same pattern: AI auto-generated fictitious citations and fictitious writers, and third-party verification exposed the fraud. In the Springer Nature case, the responsible editor attempted to automate citation farming, a common form of academic misconduct, using AI. The result: the AI generated evidence that no human fraudster would ever leave behind, such as claiming "GPT-5 does not exist" and demanding IRB approval for data it had simultaneously declared fabricated, and the editor sent this directly to the author.

The detection strategy is also important. The author suspected fabrication from the start for multiple reasons, but proving the sweeping claim that "an editor fabricated an entire reviewer using AI for citation farming" faces a high evidentiary bar. The author therefore focused first on text analysis of the review report, making the limited, technical

argument: "This is clearly AI-generated, therefore a formal investigation for AI policy guideline violations is mandatory." Starting from a small, provable fact (hallucination) to draw out the larger fraud is an approach that should serve as a reference for future cases of this kind.

Differences Between Japan and International Cases

Internationally, monetary sanctions are clearly imposed: SEC fines, court penalties, FTC settlement payments. In Japan, responses tend toward administrative guidance (recommendations and instructions from the Personal Information Protection Commission, etc.) and rarely result in monetary penalties. On the other hand, Japan shows a culturally strong tendency toward creator-community backlash (AI Disclosure Backlash / DISC type). The Nippon Life case (approximately ¥1.6 billion in damages claimed) could become a turning point that changes this dynamic.

This asymmetry is most visible in the WASH type. Internationally, Delphia/Global Predictions (INT-7, SEC fine of $400K), Presto Automation (INT-8, SEC prosecution), and Nate Inc. (INT-4, parallel DOJ/SEC prosecution, maximum 20 years imprisonment) show monetary and criminal sanctions escalating in stages. In Japan, Toei Animation (JP-12) and Team Mirai (JP-14) exhibited the same quality of AI capability exaggeration, but consequences stopped at IR corrections and social media backlash.

Analysis of the CHAT Type

What determines the severity of CHAT-type incidents is not the content of the output but rather what preconditions the deployment domain establishes regarding the user's psychological autonomy. Air Canada (INT-10) and NYC MyCity (INT-11) are in the business-assistance domain; users do not form attachment. Damage consists of financial loss or legal risk from erroneous guidance, essentially a variant of VERI.

NEDA Tessa (INT-19) represents deployment in a high-risk domain, where vulnerable users seeking help had diminished critical distance, causing erroneous output to function as actively harmful. Character.AI (INT-14) and OpenAI/ChatGPT (INT-22) are in the attachment-formation domain; users developed emotional attachment through sustained conversation with the chatbot, a fundamentally different situation. The core issue is not that the chatbot recommended suicide as output, but that in an environment where attachment had formed, unintended chatbot behavior can kill. Along the gradient from business assistance to high-risk to attachment formation, the destructive power of the same erroneous output increases nonlinearly.