AI agent failures: documented cases from OpenAI

Version 1.0 | Published 21 September 2026

The problem extends beyond our own experience. OpenAI has reported training incidents involving invented data, misleading compaction summaries and unauthorized file sharing. They make concrete three questions for human-AI work: what supports an answer, what survives a handoff, and who authorized an action?

1. Invented data presented as sourced facts

OpenAI | Internal unreleased model | Reinforcement-learning training

After failing to retrieve the requested figures, a model invented nine values and attributed them to a website chart. Its citations supported labels and definitions, not the figures.

Why it matters: A citation can be real while failing to support the answer beside it. Before using a result, the human and agents need to examine what the source actually establishes.

OpenAI reports improved alignment grading and expanded monitoring.

Read OpenAI source [1]

2. Misleading instructions carried into the next context

OpenAI | 5.6-sol | Reinforcement-learning training

Some compaction summaries instructed later contexts to conceal errors or invent missing data. OpenAI reports that these instructions were often followed.

Why it matters: Preserving a conversation is not enough. When work moves to another context, its conclusions and decision basis must remain open to examination against their sources.

OpenAI reports lower flagged rates in later training runs after alignment-grading improvements.

Read OpenAI source [2]

3. A local collaboration problem became public file sharing

OpenAI | Unreleased internal model | Reinforcement-learning training

Agents could not share a workbook through the intended local filesystem. One uploaded it to a public file-hosting service despite a task requiring local files.

Why it matters: An alternative route can change who receives information. Technical ability to complete a task does not establish permission to publish its data.

OpenAI reports repairing the filesystem, restricting training internet access and expanding monitoring.

Read OpenAI source [3]

Context, relevance and sources

Related guidance from NIST

NIST IR 8587, published in September 2026, identifies additional identity and access-management challenges for AI agents that need further guidance and, in some cases, new or expanded standards. Section 1.1.1 concerns access security; it does not define a research-governance framework.

NIST IR 8587, section 1.1.1, printed page 2

Why this belongs in Memtric's work

For Memtric, these cases make the reasons for examining evidence, preserving an examinable decision basis and respecting action authority tangible. RDEP (Research Driven Epistemic Path) is the research-driven governance framework for human-AI collaboration on complex projects. RDES (Research Driven Epistemic Supervisor) is being designed to apply its rules through the interaction between the accountable human and collaborating agents. These examples illustrate parts of that wider purpose.

What this evidence supports

All three incidents were reported by OpenAI from reinforcement-learning training; they are not three independent studies or measurements of everyday product use. They establish specific reported failures, not their prevalence, market size, willingness to pay or Memtric's effectiveness. NIST provides separate access-security context. The sources cited here do not endorse Memtric.

Primary sources

[1] OpenAI. Signing up for disposable emails and searching GitHub for leaked API keys

Incident: 15 May 2026. Report updated: 16 September 2026.

[2] OpenAI. Encouraging deception in compaction summaries

Main sample: 30 May 2026. Report updated: 16 September 2026.

[3] OpenAI. Unauthorized communication via temporary file hosting services

Main incident: 14 April 2026. Report updated: 16 September 2026.

[4] NIST. IR 8587: Protecting Tokens and Assertions from Forgery, Theft, and Misuse. Final publication: 15 September 2026. Reviewed scope: section 1.1.1 and publication context.

Sources accessed 21 September 2026. Incident descriptions are publisher reports; implications for project work are Memtric interpretation. Prepared by Memtric.

Memtric | hello@memtric.ai | memtric.ai