

Title

Why Most AI Incidents Are Evidence Failures, Not Model Failures

<https://www.aivojournal.org/why-most-ai-incidents-are-evidence-failures-not-model-failures/>

Authors

Tim de Rosen
AIVO Journal

Abstract

Public discourse on AI risk continues to frame incidents primarily as technical failures: model bias, hallucination, or misconfiguration. This article advances a different interpretation grounded in governance practice. Drawing on patterns observable in the OECD AI Incidents Monitor, it argues that many AI incidents escalate not because models fail, but because institutions cannot reconstruct what AI systems said, when they said it, and how those representations were framed at the moment of reliance.

The article does not assess model accuracy, internal system design, or causality. Instead, it examines AI incidents as post-event accountability failures driven by missing or non-inspectable evidence. Through sector-agnostic walkthroughs spanning finance, healthcare, and public administration, it demonstrates a recurring governance failure mode: once scrutiny occurs, the absence of contemporaneous, interaction-specific records converts uncertainty into institutional exposure regardless of technical intent or system quality.

The paper reframes AI incident management as an evidentiary control problem rather than a model optimization problem. It concludes that, in non-deterministic systems deployed as external representation channels, accountability depends less on improving prediction accuracy than on preserving inspectable records of AI-mediated representations at the point of human reliance.

Keywords

AI governance
AI incidents
AI accountability
Evidence and auditability
Post-market oversight
AI risk management

Description

This deposit contains the full article “**Why Most AI Incidents Are Evidence Failures, Not Model Failures**”, including two appendices that provide empirical and operational context:

- **Appendix A** summarizes observable patterns in OECD-catalogued AI incidents relevant to post-incident evidentiary disputes.
- **Appendix B** maps common OECD incident attributes to evidentiary control requirements that arise under scrutiny.

The article is intentionally non-prescriptive and avoids claims about causality, liability, or regulatory mandates. Its purpose is to clarify how accountability failures emerge once AI-mediated representations are challenged, independent of whether model behavior is ultimately judged correct or incorrect.

This work is suitable for readers in AI governance, risk management, audit, compliance, public policy, and legal oversight.

Notes on Methodology and Scope

- The OECD AI Incidents Monitor is used as a **descriptive corpus**, not as proof of causation.
- All claims are **interpretive and governance focused**.
- The article does **not** evaluate model internals, training data, or system design choices.
- The analysis is limited to **publicly reported incidents** and therefore represents a lower bound on institutional exposure.