Why We Need to Know More:
Exploring the State of AI Incident Documentation Practices

Violet Turri, Rachel Dzombak

Background

- To enable an equitable, AI-powered future, developers and practitioners must monitor AI systems and learn from past AI incidents when failures have occurred. Around the world, public databases for cataloging AI systems are instrumental in promoting awareness of potential AI harms among policymakers, researchers, and the general public. However, despite growing recognition of the potential of AI systems to produce harms, causes of AI systems failure remains elusive and AI incidents continue to occur. **This begs the question – how are we learning from documented incidents?**
- Currently, three public databases for reporting AI incidents exist. The AI Incident Database (AIID) hosts more than 1,000 archived reports from over 600 submitters and uses the Center for Security and Emerging Technology (CSET) Taxonomy. The Where in the World is AI? database powers an interactive map of AI systems (both harmful and helpful) with more than 400 examples. The AIAAIC Repository consists of over 850 examples and used by over 60 organizations.

	AI Incident Database	AIAAIC Repository	Where in the World is AI?
Identification	Incident #	AIAAIC ID#	
Incident Description	Full description, Short description	Description	Title
Date	Beginning date, Ending date	Year	Year
Location	Location	Country(s)	City, State, Country, Latitude, Longitude
Sector	Sector of deployment, Critical infrastructure sectors affected, Public sector deployment	Sector(s)	Domain
Responsible Parties	System developer, Named entities, Party responsible for AI system	Operator(s)	
AI System Description	Relevant AI functions, AI tools and techniques used, AI functions and applications used, Description of AI system involved, Nature of end user, Level of autonomy, Physical system	Purpose(s)	
AI System Data	Description of the data inputs to the AI system		
Harm Cause	Probably level of intent, Harm type, Harm nearly missed?, Uneven distribution of harms basis		
Harm Description	Human lives lost, Total financial cost, Overall severity of harm	Issue(s) – General, Issue(s) – Transparency	
Harm Impact	Laws covering the incident		
Legal Implications			is_good
Harm Response			

The above table provides a side-by-side comparison of existing AI incident documentation taxonomies. Existing methods tend to focus on the sector in which an incident occurred or its location, but fail to capture critical information related to harms causes, impacts, or responses.

Reporting System Precedents

- The Federal Aviation Administration (FAA) Aviation Safety and Information Analysis and Sharing System (ASIAS): this system brings together 11 key aviation safety databases and allows users to query multiple databases at once.
- The US National Transportation Safety Board (NTSB) Aviation Accident and Incident Data System: this system hosts reports of aviation accidents and incidents. Each report is rigorously investigated by the NTSB.
- The Aviation Safety Reporting System (ASRS): this system stores confidential incident reports submitted by people in various aviation roles (such as pilots, flight attendants, or mechanics) and consists of over 1 million reports to date.
- The Common Vulnerabilities and Exposures (CVE) database: this database documents publicly known cybersecurity vulnerabilities and provides analysts and testers with common language.

Taxonomy Considerations

- **AI Incident taxonomies should account for the fact that (to date) AI failures tend to be *context- and system-specific*.** This differs from precedents in which problems are typically the result of common faulty (physical) components used across systems.
- **AI taxonomies must expand their coverage of phenotypical characteristics of AI incidents** to encourage genotypical analysis. A common pitfall of taxonomies is relying too heavily on phenotypical categories (observable characteristics) and consequently oversimplifying data. This makes it challenging to identify genotypical categories (underlying factors). The ASRS provides a model for how a rigorous phenotypical taxonomy can enable genotypical analysis. Information about system inputs and outputs, training and testing data sets, and/or model weights may be critical to identifying broader underlying causes.
- **AI incident databases need to capture longer and more detailed timelines** than precedent databases. For AI systems, logging dynamic elements of the system (such as modifications to the dataset or model retraining) is essential to understanding failure modes.

Database Authorship

- Unlike in aviation or cybersecurity, AI databases are not federally operated and disclosure of AI incidents is not mandatory. On the one hand, this encourages public curation and collaboration from a variety of parties and perspectives. On the other, **government oversight of a public AI incident database** could lead to better coordinated response, analysis, and policy.
- In the absence of a legal mandate, **providing an anonymous submission method may incentivize incident reporting**. The Aviation Safety Reporting System is a prime example of how confidential reporting can lead to strong participation. Empowering development teams and end-users to submit first-hand accounts of AI incidents without fear of reputational harm may lead to more widespread reporting and a richer dataset.

Proactive Documentation

- A number of databases currently support proactive documentation of AI systems with the objective of a) facilitating government transparency and/or b) raising public awareness of the use of controversial technologies.
- While these are critical goals, existing databases fail to **make links between proactive and retroactive documentation**. Establishing a database for this purpose would support closer system monitoring prior to and in anticipation of possible incidents and further illuminate the timeline of AI failure.

Conclusion

- AI incident databases have great potential to support AI practitioners in gaining awareness of the failure mechanisms of AI system. Increasing practitioner understanding of failures and their underlying causes can lead to well-informed best practices and better engineered systems.
- As the AI community continues to document incidents, reflection is needed on how information is captured about AI systems and the ways in which databases and taxonomies can support or prevent meaningful analysis.