# AI Incident Collection: An Observational Study of the Great AI Experiment

By Heather Frase, PhD (hf276@georgetown.edu) and Ren Bin Lee Dixon

We are in the midst of a global AI adoption experiment. Discussion about how to limit harm and regulate these systems is widespread. As investigators, we should gather information during this Great AI Experiment by executing an observational study that collects critical, structured and unstructured data on AI incidents as they occur in the wild. While AI systems are often tested before deployment, those tests are not always robust to real-world situations, because AI systems may act differently when exposed to untested conditions. As a result, we learn much about AI behavior post-deployment, when it is in use and has the opportunity to harm.

Informed by CSET's work on the AI Harm Framework and other ongoing research, this explainer recommends a structured, observational study that builds robust, updateable data sets on AI incidents to inform regulation and harm mitigation efforts.

## What Good AI Incident Collection Requires

Gathering information on AI incidents and risks can help to improve AI safety, shape policy, support risk mitigation, and build trust. However, simply collecting data about AI incidents does not automatically lead to these benefits. To be effective, the collection should have:

- **Clear goals that support action**—why we are collecting particular data?
- **Collaboration**—who is involved in planning collection efforts?
- **Analyzable and meaningful data**—what we are collecting?
- **Clear and specific requirements**—when should we collect?
- **Infrastructure that is easy to use**—where should we collect data, and how should we process it?
- **Updateable processes**—how can we improve the collection with time and experience?
- **Adequate resourcing**—what do we need to maintain the collection?
- **Identified roles, responsibilities, and authorities**—who does what?

## Clear Goals that Support Action

Incident data collection should provide actionable information and serve clear goals. The goals address 'why' we are collecting data, and they guide the who, what, when, where, and how of collection—with periodic checks that we are meeting these goals. Information collection is typically more impactful when research is outcome-focused with specific, measurable, trackable goals.

## Collaboration

The importance of collaboration for AI-related projects and processes is well-established. We make systems better by collaborating across industry, racial, ethnic, gender, and other identities and affiliations, as both NIST's AI Risk Management Framework and White House's Blueprint for an AI Bill of Rights recommend. By reflecting multiple perspectives and use-cases, a group of collaborators from diverse backgrounds also improves data collection by reducing blind spots, improving data quality and representativeness. The tradeoff for greater inclusivity can sometimes be delays/slower results due to the need to build consensus.

**Box 1. Tradeoffs Between Highly Structured and Less Structured Data in the Financial Crimes Mitigation**

The Bank Security Act (BSA) of 1970 requires financial institutions to collect information on monetary transactions with the goal of detecting criminal activity. Under the Act, financial institutions must file a highly structured Currency Transaction Report (CTR) for any transaction involving more than $10,000. Unfortunately, criminals found ways to move money that did not trigger a CTR filing.

A later amendment to the BSA required reporting called the Suspicious Activity Report (SAR), which collects a wider range and variety of data via free-text fields, many categorical fields that are updated as new trends emerge by Treasury's FinCEN, and an option to add pages of background information. This combination of structured and open-ended text and frequent updates permits observations of financial crime patterns and trends that are not restricted to the transaction amounts. The SAR is messier and harder to use and analyze, but it is much better at catching new problems and criminal activity. Thus, the move to the SAR from the CTR poses a tradeoff between easy to analyze data (CTR) and more meaningful data (SAR).

## Analyzable and Meaningful Data Collection

Data is most useful when it is analyzable and meaningful. Analyzable data needs to be consistent and understandable with definitions, described data relationships to enable data annotator and analyzer clarity. We should develop special tools, taxonomies, and

ontologies to support the analyzability of the data we collect. These tools can aid communication between communities and improve the reproducibility of analytics. To be meaningful, these tools must capture all the essential data and information we need to achieve our goals.

Sometimes, we must make tradeoffs between analyzable and meaningful data (see Box 1). When we analyze data, it is better to have clear, simple, organized, and structured information. But, having more context and details helps to understand things deeply or gain new insights. Often, we cannot organize this additional information systematically or neatly.

## Clear and Specific Requirements

AI incident collection needs detailed reporting criteria that can address the wide variety of AI incidents to ensure coverage and timeliness of data collection. We need to say precisely when to report AI incidents, what information is given, and how quickly it is submitted. We should collect data on both smaller problems that happen often and severe issues that may happen rarely. Reporting less serious incidents encourages developers and deployers to implement preventative measures while reporting serious events helps us understand how they happen and can inform policymakers.

However, capturing the range of incidents does not mean that data about all incidents should be reported in the same way. We do not want relevant incidents to slip through the cracks. Thus, there should be multiple, clearly defined, triggers for reporting, with different reporting criteria, timelines, and information for different types of events.

Triggers or rules for reporting different kinds of incidents can depend upon multiple factors: severity, number of affected people, sector of use, amount of autonomy, the type of AI, etc. For example, incidents with severe consequences, like death, may require fast initial reporting with basic details. Then, later on, initial reporting can be supplemented with additional root cause or other analysis. Additionally, there could also be sector-based reporting requirements. For example, reporting all AI incidents involving law enforcement regardless of the severity of impact[1] while small issues, like an autonomous personal vacuum robot getting stuck on stairs, do not.

## Infrastructure that is Easy to Use

A successful collection effort needs infrastructure that is both user-friendly and able to scale, adapt, and evolve. Infrastructure will likely include the tools, workforce, processes, and environments to upload, maintain, and disseminate reports. Since

[1] The AI Act proposed by the European Union is slated to be finalized at the end of 2023. It is expected to adopt a risk-based approach and identities certain use-sectors as high-risk. Law enforcement is one of the listed sectors.

collaboration is vital to effective AI incident collection, infrastructure must support different stakeholder processes depending on function or filing frequency. Additionally, as the goals and type of data collection evolve, collection infrastructure should seamlessly accommodate both historical data maintenance and future needs.

Effectively disseminating data can increase its impact on safety, but transparency must be balanced with privacy and security. Some AI incident reports may contain information about vulnerabilities that could be exploited, while others may contain personally identifiable information. Thus, effective infrastructure should support compartmented access, granting different entities access to different aspects of the data. The general public may only need to see summary information or trends of incidents, while regulators and auditors may require access to root-cause analysis of individual incidents.

### Updateable Processes

Since initial data collection efforts may be imperfect and the technology will continue to adapt, incident collection and analysis procedures must be able to evolve. Factors improving adaptability include, but are not limited to, modularity, a faceted (instead of a hierarchical) data taxonomy, clear documentation with version control, regular reviews, and feedback mechanisms.

### Adequate Resourcing

Before collecting AI incident information, we need to determine and secure the resources (money, people, infrastructure) required to stand up and sustain the effort. Sometimes, we may need more resources for the planned incident collection. If this happens, we might need to change the goals to something that requires less resources.

### Identified Roles, Responsibilities, and Authorities

To successfully build and maintain an AI incident database, we need to decide who is doing different jobs, like data collection, infrastructure support, data sharing, oversight, and updating. We need to ensure responsibilities and authorities are delineated and deconflicted. Maintaining an independent oversight group that ensures processes are followed and goals are met will also be important to ensure the system is meeting its goals and complying with applicable laws and policies.

## The Must Versus Can of Incident Reporting

Conceptually, we can divide incident reporting into three types: mandatory, voluntary, and citizen.

- Mandatory reporting involves legal or regulatory requirements to report specific incidents with specific information. Mandatory reports are usually submitted to a government agency database.
- Voluntary reporting allows people and organizations the choice to report, often with recommendations on when to report and guidance on what information to provide. Voluntary reporting is often submitted to databases run by government agencies or professional organizations.
- Citizen reporting has a lot in common with voluntary reporting. However, citizen reporting is done by people and organizations serving as watchdogs. Such reporters may be academics, journalists, social media posters, or non-profit organizations. There are cases, like the Pandemic Response Accountability Committee (PRAC) hotline, where government agencies collect citizen reporting.

Table 1 summarizes the pros and cons of these three groups of reporting. Hybrid reporting approaches, blending all three, are also possible. For example, SARs are hybrid with mandatory reporting conditions, but flexibility for financial institutions to submit SARs for any activity that they consider suspicious.

Currently, AI incident reporting is dominated by voluntary 'citizen' reporting. Different organizations are collecting these citizen reports into databases.[2] These collections are valuable, and CSET is actively studying their content. However, data is often messy, incomplete, and time-consuming to analyze. We will likely need hybrid reporting of AI incidents to detect new harms and mitigate risk adequately. In a future publication, CSET will further analyze the pros and cons of mandatory and voluntary AI incident reporting.

[2] In the US, the primary AI incident databases are the AI Incident Database (AIID), the Algorithmic, and Automation Incidents and Controversies (AIAAIC) repository, the AI Vulnerability Database (AVID), and the Emerging Technology Institute's AI Litigation Database.

Table 1. Pros and Cons of Different AI Incident Reporting Regimes

| Type | Pros | Cons |
|---|---|---|
| Mandatory | • Consistency in when and what is reported<br>• Reduced reporting gaps<br>• Detailed data that could contain root-cause analysis, which is essential for mitigation<br>• Improved safety awareness and culture<br>• Option for investigative safety board<br>• Early detection | • Imposes administrative burden on reporting entities, which can be significant for small organizations<br>• Overly broad reporting criteria may yield low-value reports that do not support action or change<br>• Inflexible<br>• Compliance needs to be enforced<br>• Slow startup because they require legislation or policy<br>• Industry fear of repercussion from what reporting exposes |
| Voluntary | • Less burden on reporting<br>• Fewer, low-value reports<br>• More flexible than mandatory reporting | • Underreporting and selection bias.<br>• Organizations may not report or reduce content out of self-interest.<br>• Inconsistent data collection due to lack of standardization<br>• Increased reporting gaps<br>• Data may not accurately represent trends and issues<br>• Collection may not be maintained |
| Citizen | • Extremely flexible<br>• Likely to catch novel or unexpected harms<br>• Can be established quickly<br>• Reports are fully accessible to everyone<br>• Can foster AI literacy<br>• Means for self-agency when people experience AI harm<br>• Transparency of reports may increase AI trust | • The most inconsistent and messy of the three<br>• Unlikely to have root-cause or information about incident mechanisms<br>• Burden of reporting on private citizens and harmed entities<br>• Data gaps and inconsistencies.<br>• Reporting tends to peek for new or high-interest items and then drop. Reporting may not reflect reality.<br>• Poor databases maintenance |