CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Explainer for “An Argument for Hybrid AI Incident Reporting”

Artificial intelligence incidents have been occurring with growing frequency since AI capabilities began advancing rapidly in the last decade. Many incidents have a significant impact (e.g., fake robocall imitating President Biden that attempted to mislead voters, facial recognition technology disproportionately misidentifying women, Black, Latino, and Asian shoppers as shoplifters, and pornographic deepfake images of a celebrity circulated on social media), leading to an urgent need to systematically and comprehensively collect and analyze AI incident data. However, current AI incident reporting databases are still in their early stages, relying primarily on citizen reporting, volunteers for data entry, and inconsistent funding from donations.

What is sorely needed is a comprehensive policy that regulates the collection, classification, compilation, and accessibility of data on AI incidents. The knowledge from this data can enhance the understanding of AI harm and help in the efforts to craft effective AI policies to foster the development of safe, secure, and trustworthy AI systems and products. To address this critical gap, in a March 2024 CSET report, authors Ren Bin Lee Dixon and Heather Frase:

- Identified five AI incident reporting databases (AI Incident Database; AI, Algorithmic, and Automation Incidents and Controversies Repository; AI Vulnerability Database; AI Litigation Database; and OECD AI Incident Monitor). These databases provide useful information but are limited in coverage and consistency, hindering comparative analysis that can enhance the understanding of AI harms and creation of effective AI safety policies.

- Compared AI governance initiatives from China, the European Union, Brazil, Canada, and the United States. These initiatives either provide obligations or recommendations for reporting AI incidents. That said, these initiatives do not always include provisions that cover reporting from all relevant stakeholders, or provide a cohesive framework for comprehensive data collection.

- Drew lessons from incident reporting practices in the healthcare, transportation, and cybersecurity sectors. Their finding suggests that a federated and comprehensive AI incident reporting framework can help ensure consistent and thorough data collection. Additionally, establishing an investigative safety board can support effective safety policies.

Based on their analysis, the authors developed recommendations calling U.S. policymakers to:

- Establish clear policies for a federated hybrid reporting framework supported by mandatory, voluntary, and citizen reporting.
- Develop a classification system to standardize AI incident reporting.
- Create an AI incident investigation agency that can provide safety recommendations.
- Explore an automated data collection mechanism that can help extract and process information from AI incident reports more efficiently.

This is the time to act on establishing a comprehensive AI incident reporting framework, since AI is rapidly becoming ubiquitously available and integrated across all sectors of society and the economy. Such a framework will help prevent data gaps, allowing for better analyses of AI incidents, and develop more effective AI safety policies.

What Is Incident Reporting?

Incident reporting has been an integral component of safety practices across different sectors—from healthcare to aviation, manufacturing to occupational safety, and utilities to food safety. When adverse events or harm occur, vital data is collected to help gain deeper insights into the root causes, uncover trends, and prevent past failures from reoccurring. These insights form the basis for developing more accurate and effective policies that foster a robust safety ecosystem.

In this explainer, the authors focus on incident reporting to an independent external organization (e.g., government agency, professional association, oversight body, etc.), which can fall into three main categories:

- **Mandatory reporting:** Organizations must report certain incidents as directed by regulations.
- **Voluntary reporting:** Individuals and groups are permitted and encouraged to report incidents, often with clear guidelines and policies.
- **Citizen reporting:** Similar to voluntary reporting, but incidents are reported by the public, journalists, and organizations acting as watchdogs.

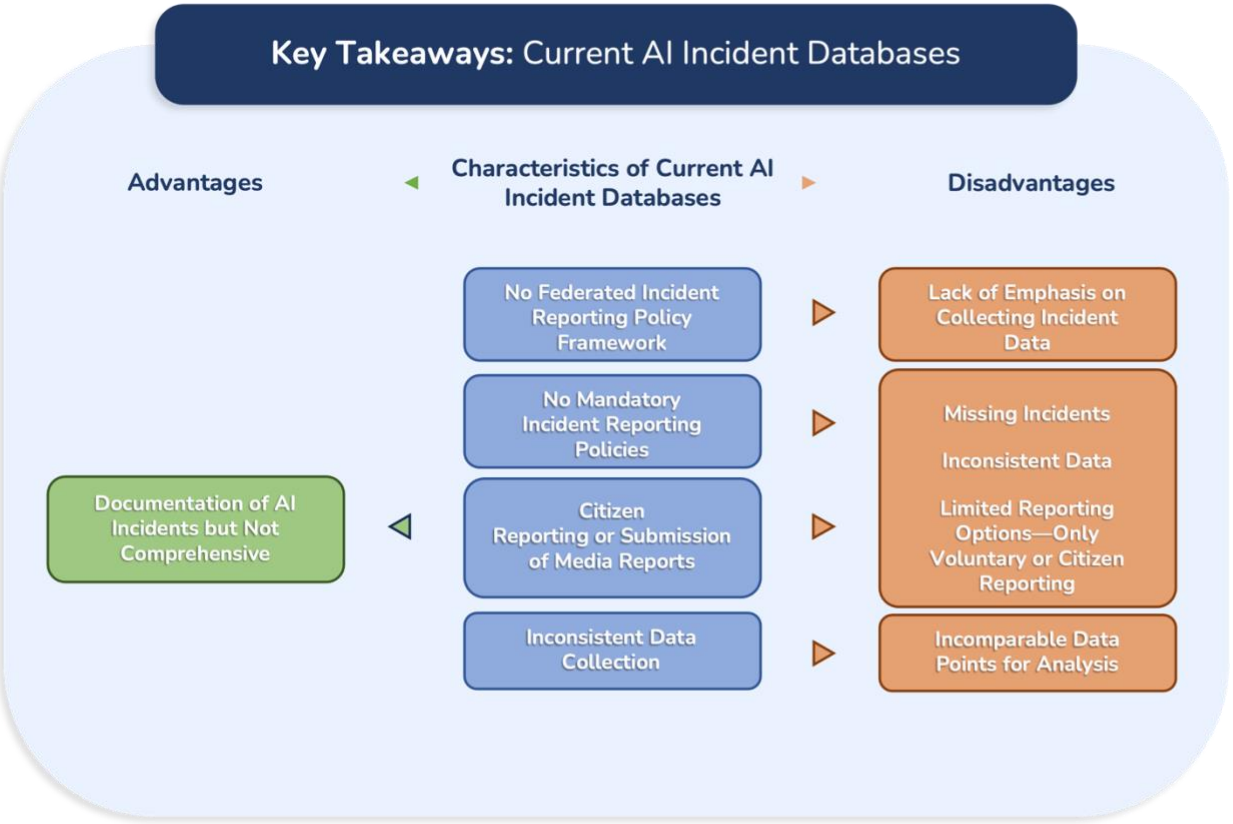
Incident Reporting: AI Incident Databases

A survey of existing AI incident reporting databases yielded a handful of key players:

- AI Incident Database (AIID)
- AI, Algorithmic, and Automation Incidents and Controversies Repository (AIAAIC)
- AI Vulnerability Database
- AI Litigation Database
- OECD AI Incident Monitor (AIM)

AIID, AIAAIC, and AIM are the only AI incident databases that attempt to capture publicly available data related to AI harms and issues. In contrast, the AI Vulnerability Database emphasizes identifying vulnerabilities in AI systems, while the AI Litigation Database focuses on documenting AI-related legal cases.

Figure 2. Key Takeaways: Current AI Incident Databases


Key Takeaways: Current AI Incident Databases
Advantages
Characteristics of Current AI Incident Databases
Disadvantages
No Federated Incident Reporting Policy Framework
Lack of Emphasis on Collecting Incident Data
No Mandatory Incident Reporting Policies
Missing Incidents
Inconsistent Data
Limited Reporting Options—Only Voluntary or Citizen Reporting
Documentation of AI Incidents but Not Comprehensive
Citizen Reporting or Submission of Media Reports
Inconsistent Data Collection
Incomparable Data Points for Analysis

While AIID's, AIAAIC's, and AIM's work is valuable in setting the foundation for documenting AI incidents, these initiatives have developed separate taxonomy and classification frameworks for defining AI incidents and harms. The conflicting definitions of AI incidents and harms make it difficult to conduct comparable research in AI safety.

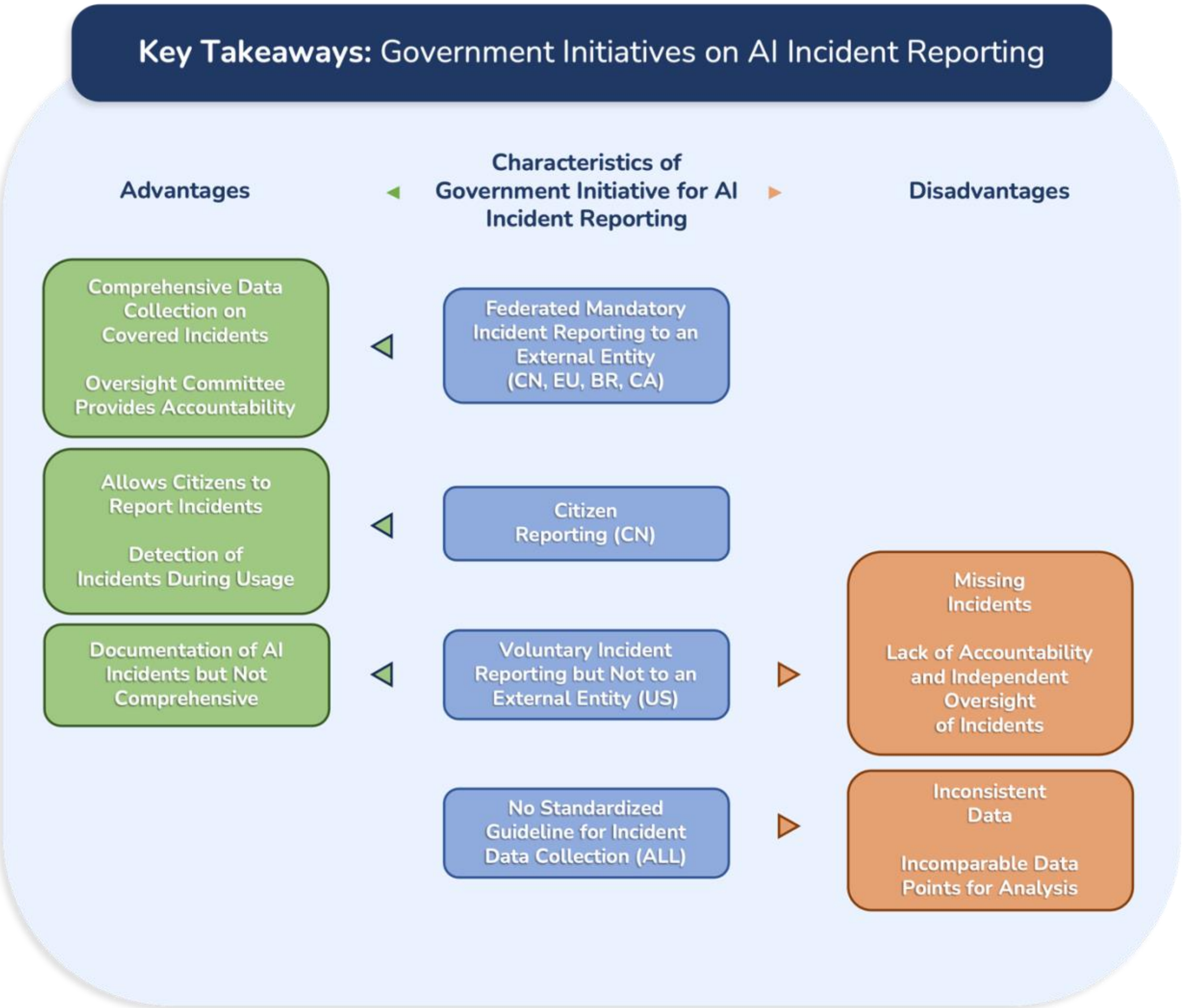
Incident Reporting: Government AI Initiatives

A review of AI government initiatives globally revealed that China, the European Union, Brazil, and Canada have enacted or proposed guidelines for mandatory AI incident reporting.

The legislative initiatives from China, the European Union, Brazil, and Canada outline obligations for AI developers and providers to report incidents. However, the initiatives do not include clear recommendations for implementing consistent federated incident reporting frameworks and data collection. China's legislative initiative is the only one that addresses citizen reporting on AI incidents. The proposals from the European Union, Brazil, and Canada do not extend incident reporting provisions to include other stakeholders, such as the public, that could potentially experience AI harm.

Meanwhile, in the United States, the Executive Order 14110 for the development of safe, secure, and trustworthy AI demonstrates the U.S. government's intent to capture AI incidents data, yet it does not include details on how this should be done. Other U.S. governmental documents that discuss reporting AI incidents are mainly recommendations and guidelines but not necessarily toward an external entity.

Figure 3. Key Takeaways: Government Initiatives on AI Incident Reporting

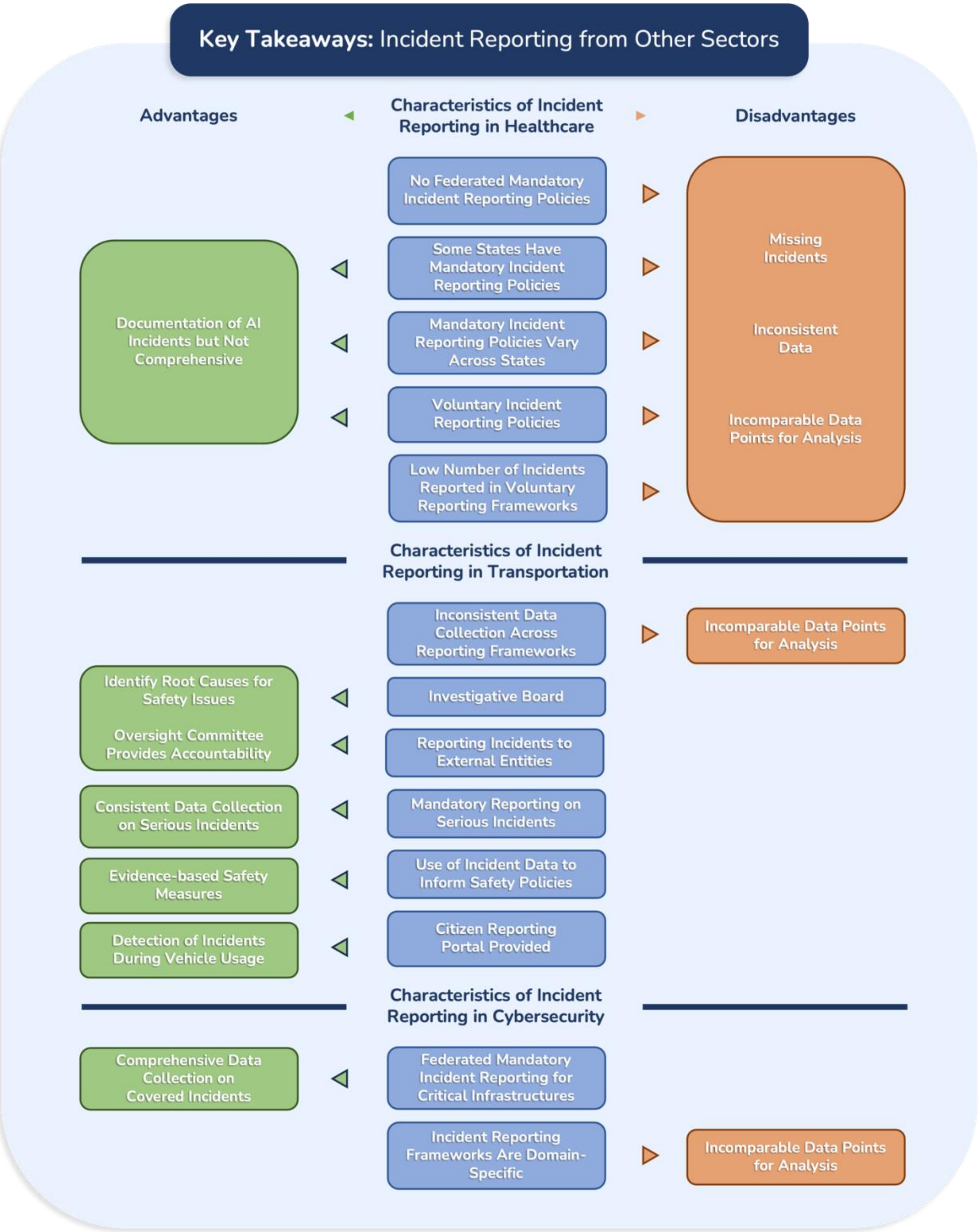

Key Takeaways: Government Initiatives on AI Incident Reporting
Advantages
Characteristics of Government Initiative for AI Incident Reporting
Disadvantages
Comprehensive Data Collection on Covered Incidents
Oversight Committee Provides Accountability
Federated Mandatory Incident Reporting to an External Entity (CN, EU, BR, CA)
Allows Citizens to Report Incidents
Detection of Incidents During Usage
Citizen Reporting (CN)
Documentation of AI Incidents but Not Comprehensive
Voluntary Incident Reporting but Not to an External Entity (US)
Missing Incidents
Lack of Accountability and Independent Oversight of Incidents
No Standardized Guideline for Incident Data Collection (ALL)
Inconsistent Data
Incomparable Data Points for Analysis

Note: CN=China, EU=European Union, BR=Brazil, CA=Canada, and US=United States.

Incident Reporting: Other Sectors

Looking at incident reporting frameworks from the healthcare, transportation, and cybersecurity sectors yielded valuable lessons. The healthcare sector's use of voluntary reporting resulted in missing incidents and incomparable data points for analysis. The transportation sector's established incident reporting frameworks include investigative boards for identifying root causes, which are used to inform evidence-based safety measures. In cybersecurity, the U.S. government has issued a series of mandates requiring mandatory reporting in selected domains, shifting away from relying on standards and soft laws.

Figure 4. Key Takeaways: Incident Reporting from Other Sectors


Key Takeaways: Incident Reporting from Other Sectors
Advantages
Characteristics of Incident Reporting in Healthcare
Disadvantages
No Federated Mandatory Incident Reporting Policies
Some States Have Mandatory Incident Reporting Policies
Mandatory Incident Reporting Policies Vary Across States
Voluntary Incident Reporting Policies
Low Number of Incidents Reported in Voluntary Reporting Frameworks
Documentation of AI Incidents but Not Comprehensive
Missing Incidents
Inconsistent Data
Incomparable Data Points for Analysis
Characteristics of Incident Reporting in Transportation
Inconsistent Data Collection Across Reporting Frameworks
Incomparable Data Points for Analysis
Identify Root Causes for Safety Issues
Oversight Committee Provides Accountability
Investigative Board
Reporting Incidents to External Entities
Consistent Data Collection on Serious Incidents
Mandatory Reporting on Serious Incidents
Evidence-based Safety Measures
Use of Incident Data to Inform Safety Policies
Detection of Incidents During Vehicle Usage
Citizen Reporting Portal Provided
Characteristics of Incident Reporting in Cybersecurity
Comprehensive Data Collection on Covered Incidents
Federated Mandatory Incident Reporting for Critical Infrastructures
Incident Reporting Frameworks Are Domain-Specific
Incomparable Data Points for Analysis

Lessons Learned

The authors' analysis revealed important lessons for an AI incident reporting policy framework:

- **Limited incident reporting frameworks are inadequate.** The incident reporting initiatives they studied emphasized citizen, voluntary, or mandatory reporting, often focusing on one or two of these categories. However, when used in isolation, each framework has limitations.

- **Inconsistent data collection creates meaningless data.** Relying on state initiatives or domain-specific guidelines may lead to inconsistent AI incident data, hindering statistical analysis and accurate depiction of AI harm.

- **There is a need for a federated AI incident reporting framework.** The absence of a federated AI incident reporting policy has resulted in fragmented and inconsistent reporting initiatives, impacting incident data collection in healthcare.

- **Incident investigation supports effective safety policies.** An investigative safety board can conduct root-cause analyses of AI incidents, provide feedback to AI actors, assist U.S. policymakers in crafting regulations, and educate the public on AI safety.

Turning Lessons into Action

Based on the observations discussed above and the nature of AI as a general-purpose technology, here are the authors' recommendations to address the current gap in AI incident reporting.

- **Establish clear policies for a federated hybrid reporting framework.** U.S. policymakers should establish a comprehensive AI incident reporting framework to gather data across sectors. Incidents should be reported to an external entity for transparency and accountability. The hybrid reporting framework should be supported by:

    - **Mandatory reporting:** Relevant AI actors should be mandated to report covered incidents in a timely manner.

    - **Voluntary reporting:** Capturing AI incidents beyond the mandatory scope.

    - **Citizen reporting:** Easily accessible reporting framework for the public to report AI incidents.

- **Develop a standardized and authoritative classification system.** The AI incident reporting framework should include standardized disclosures and address unique domain characteristics like privacy and regulatory requirements.

- **Create an independent AI incident investigation agency.** When a significant AI incident occurs, an independent board should investigate and provide safety recommendations.

- **Explore automated data collection mechanisms.** Automated data collection mechanisms could help obtain intricate technical and contextual information from AI incidents.

For more information:

- Download the report: https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/

- Contact us: cset@georgetown.edu