

Issue Brief

An Argument for Hybrid AI Incident Reporting

Lessons Learned from Other
Incident Reporting Systems

Authors

Ren Bin Lee Dixon

Heather Frase



CSET CENTER *for* SECURITY *and*
EMERGING TECHNOLOGY

March 2024

Executive Summary

AI incidents have been occurring with growing frequency since AI capabilities began advancing rapidly in the last decade. Despite the number of incidents that have emerged during the development and deployment of AI, there is not yet a concerted U.S. policy effort to monitor, document, and compile AI incidents and use the data to enhance our understanding of AI harm and inform AI safety policies in order to foster a robust AI safety ecosystem. In response to this critical gap, the objectives of this paper are to:

- Examine and assess **existing AI incident reporting initiatives**—both databases and government initiatives.
- Elicit lessons from **incident reporting databases from other sectors**.
- Provide **recommendations** based on our analysis.
- Propose a **federated* and standardized hybrid reporting framework** that consists of
 - **Mandatory reporting:** Organizations must report certain incidents as directed by regulations, usually to a government agency.
 - **Voluntary reporting:** Individuals and groups are permitted and encouraged to report incidents, often with clear guidelines and policies, and usually to a government agency or professional group.
 - **Citizen reporting:** This is similar to voluntary reporting, but incidents are reported by the public, journalists, and organizations acting as watchdogs.

*For the purpose of this paper, we define a federated framework as a centralized framework prescribed by a singular authoritative government body or the federal government. The framework stipulates a set of minimum requirements that can be adapted and implemented across government agencies or non-governmental organizations.

When discussing incident reporting in this paper, we emphasize reporting to an independent external organization (e.g., a government agency, professional association, oversight body, etc.).

A survey of existing AI incident collection efforts identified only two citizen-reporting organizations actively capturing AI incidents. Additionally, a review of AI legislative initiatives globally revealed China, the European Union, Brazil, and Canada have enacted or proposed guidelines for mandatory AI incident reporting. Currently, there aren't any significant legislative initiatives for establishing an AI incident reporting policy framework in the United States. The available U.S. governmental documents that mention reporting AI incidents are recommendations and guidelines for implementing reporting mechanisms but not necessarily toward an external entity.

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 has an established incident reporting framework that includes investigative boards for identifying root causes, which are then 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 other soft laws.

Our analysis of the two AI incident reporting databases, emerging government initiatives related to AI incident reporting, and the various incident reporting systems in the healthcare, transportation, and cybersecurity sectors revealed disadvantages and advantages. These insights offered several important lessons that can be applied to an AI incident reporting policy framework, as discussed in the following:

- **Limited incident reporting frameworks are inadequate.** Across the board, the incident reporting initiatives examined in this paper often emphasized either citizen, voluntary, or mandatory reporting, typically focusing on one or two of these reporting categories. In isolation, each of these three frameworks has limitations.
- **Inconsistent data creates meaningless data.** Relying on state initiatives or domain-specific guidelines will likely produce uneven or inconsistent data that

might not be adequate for aggregating AI incident data for statistical analysis or accurately depicting the many dimensions of AI harm.

- **There is a need for a federated AI incident reporting framework.** The absence of a federated AI incident reporting policy framework has impacted incident data collection efforts in the healthcare sector, resulting in fragmented and inconsistent reporting initiatives.
- **Incident investigation supports effective safety policies.** An investigative safety board can be useful for conducting root-cause analysis of significant AI incidents and providing feedback to help AI actors improve their design and development, enable policymakers to craft effective regulations, and educate the public on AI safety.*

Based on the observations discussed above and the nature of AI as a general-purpose technology, we make the following recommendations to address the current gap in AI incident reporting.

- **Establish clear policies for a federated hybrid reporting framework.** Policymakers should establish a federated and comprehensive AI incident reporting policy framework to gather incident data across sectors and applications. AI incidents should be reported to an independent external entity (e.g., government agency, professional association, oversight body, etc.) to promote transparency and accountability in AI incident management. A hybrid reporting framework is supported by:
 - **Mandatory reporting:** Relevant AI actors should be mandated to report covered incidents in a timely manner.

*UNESCO defines AI actors as any actor involved in at least one stage of the AI system lifecycle, and can refer both to natural and legal persons, such as researchers, programmers, engineers, data scientists, end-users, business enterprises, universities and public and private entities, among others. See: “Recommendation on the Ethics of Artificial Intelligence,” UNESCO (2021), 10, <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

- **Voluntary reporting:** Voluntary reporting frameworks should also be established alongside the mandatory framework to capture AI incidents outside the mandatory jurisdiction.
- **Citizen reporting:** An easily accessible reporting framework should be made available for the public and all other stakeholders to report and document AI incidents.
- **Develop a standardized and authoritative classification system.** The AI incident reporting framework should include a standardized set of disclosed information plus accommodations for the unique characteristics of distinct domains, such as privacy concerns and other regulatory requirements.
- **Create an independent AI incident investigation agency.** When a significant AI incident occurs, an independent board should investigate the root cause and provide evidence-based safety recommendations.
- **Explore automated data collection mechanisms.** Automated data collection mechanisms could be highly advantageous to obtain technical and contextual information from AI incidents.

Further research will be needed to explore the necessary content and considerations for implementing a comprehensive reporting framework that is applicable across sectors and applications. We will explore this in a follow-up paper and will not delve into it in this paper.

The ability to mitigate AI harms and manage their aftermath competently can shape public conversations about AI usage. An AI incident reporting framework must be integrated as an essential component of AI safety rather than developed as an afterthought in AI legislative initiatives. The present moment offers a prime opportunity to establish an AI incident reporting framework with relatively low stakes. However, this window is rapidly closing as AI becomes more prevalent across applications and sectors. A federated, comprehensive, and standardized framework will prevent data gaps and enhance data quality. Adopting a hybrid framework that includes mandatory, voluntary, and citizen reporting will improve data fidelity, providing a more accurate representation of the emerging trends in AI harm and risk.

Table of Contents

Executive Summary	1
Introduction	6
A Brief Overview of Incident Reporting	10
Current AI Incident Reporting	11
Government Initiatives for AI Incident Reporting.....	15
Lessons from Incident Reporting Policies in Healthcare, Transportation, and Cybersecurity Sectors	20
Discussion.....	31
Recommendations	35
Conclusion	39
Authors.....	41
Acknowledgements	41
Endnotes	42

Introduction

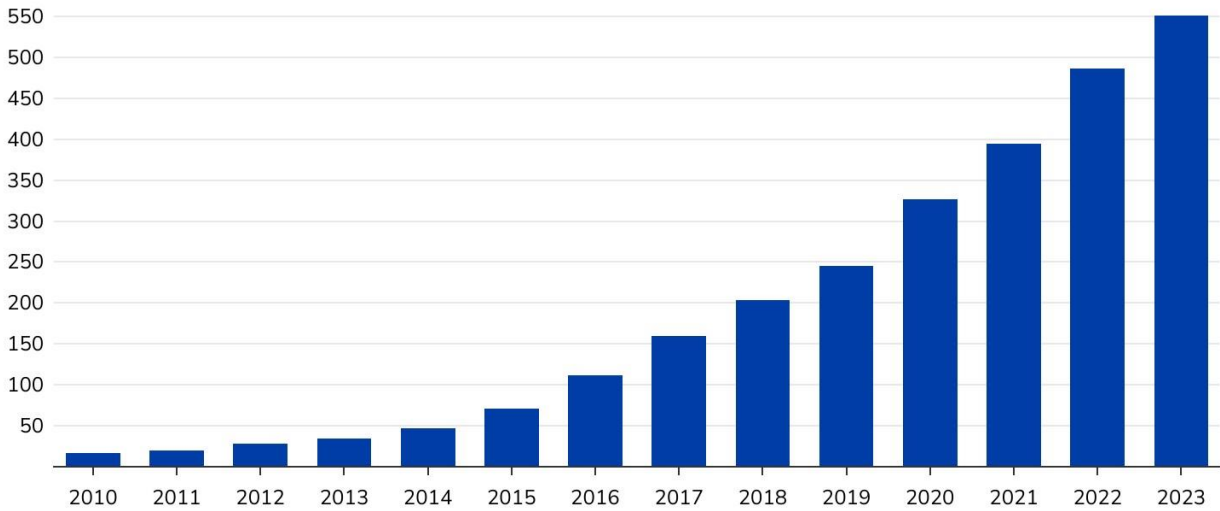
The potential capabilities and benefits that AI can bring are immense and wide-reaching. Notably, the technology has contributed to significant advancements in foundational scientific research that otherwise would have taken years or decades to achieve. For instance, in bioscience and medicine, researchers used AI to accurately predict protein structures that could accelerate new medical discovery and efficiently identify potential compounds that could attack an antibiotic-resistant superbug.¹ In clean energy, a reinforcement learning algorithm has successfully controlled nuclear fusion plasma in a tokamak (a machine that controls and contains heated hydrogen plasma), accelerating toward clean energy generated from nuclear fusion.²

That said, as AI applications and systems have become more prevalent across sectors and industries, the number of AI incidents has increased. An AI incident can be generally described as an event where an entity experienced tangible or intangible harm that can be directly linked to a consequence of the behavior of an AI system.³ Several notable recent incidents—such as biased outcomes in AI systems used to predict recidivism, in facial recognition technology, hiring decisions, and welfare allocation decisions—have drawn significant public attention to the issue of AI harm.⁴ The data captured in the AI Incident Database shows a rapid growth in AI incidents since 2010 (see Figure 1).*

*AI incidents are submitted by community members to the AI Incident Database before expert review and inclusion into the database. Data is gathered from the October 2023 database backup, the most recent available at time of writing.

Figure 1. Cumulative AI Incidents by Year

Cumulative AI Incidents by Year



AI Incidents are submitted by community members to the AI Incident Database, before expert review and inclusion into the database.

Source: AIID

A Snapshot of AI Incidents from the AI Incident Database

Between May and June 2023, volunteers tracking AI incidents submitted thirteen AI incidents to the AIID. The following examples are a snapshot of some of the incidents, to highlight the types of impacts they can have:

- The National Eating Disorders Association in the United States took down its artificial intelligence chatbot “Tessa,” which developers designed to provide healthy eating tips. Contrary to its developers’ expectations, the chatbot reportedly offered bad eating advice that could harm people seeking help.⁵
- A fake AI-generated image showing a building near the Pentagon exploding circulated widely on social media. A financial news site reported the image, which caused a brief dip in the U.S. stock market before experts dispelled the image as fake.⁶
- The U.S. National Highway Traffic Safety Administration has been investigating several road accidents and fatalities linked to Tesla Autopilot. These incidents have increased significantly from 2019 to 2023, reaching more than 736 reported crashes.⁷

Presently, dedicated AI incident reporting databases are primarily citizen reporting. It is ad hoc, in the beginning stages, dependent upon volunteers, and funded by private donations.* Despite the number of reported incidents that have emerged from the development and deployment of AI, there is not yet a concerted policy effort to document and compile AI incidents and use these data to enhance our understanding of AI harm, inform AI safety policies, and in general foster a robust AI safety ecosystem.

*The purpose of this report was to identify the gaps in federated AI incident reporting, thus we only focused on databases that collect AI incidents indiscriminately. We acknowledge that there may be AI incidents that are captured in sector-relevant databases, such as medical instruments, vehicle accidents, and hiring systems.

In response to this critical gap, the objective of this paper is to:

- Examine and assess existing AI incident reporting initiatives—both databases and government initiatives.
- Elicit lessons from incident reporting databases from other sectors.
- Provide recommendations based on our analysis.
- Propose a federated, comprehensive, and consistent reporting framework that consists of mandatory, voluntary, and citizen reporting.

Implementing an operational AI incident reporting framework will require in-depth assessments to determine, for example, the type of incidents to be reported, a comprehensive and adaptable classification system, the types of data that should be collected, and how data will be shared and used. Additional research will be needed to uncover this information, which will be the focus of a follow-up paper and will not be discussed here.

Overall, promoting a robust safety ecosystem for safeguarding society in the face of AI advancement is imperative to enable us to harness its full potential while minimizing the risk of associated harm. Establishing an AI incident reporting policy framework can enhance our understanding of AI harm, contributing to greater AI safety and risk mitigation efforts.

A Brief Overview of Incident Reporting

Incident reporting has been an integral component of safety practices across different sectors to document data when harm has occurred—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 us gain deeper insights into the root causes, uncover trends, and prevent past failures from reoccurring. These insights, in turn, serve as a foundation for developing more accurate and effective policies that foster a robust safety framework and ecosystem.

In this paper, we focus on incident reporting to an independent external organization (e.g., government agency, professional association, oversight body, etc.). We do not consider internal incident collections (e.g., company software bug tracking, internal help desk tickets, etc.) or third-party reports that companies collect on their products (e.g., customer complaints, customer support emails, etc.). Incident reporting to third-party organizations can fall into three main categories:

- **Mandatory reporting:** Organizations must report certain incidents as directed by regulations, usually to a government agency.
- **Voluntary reporting:** Individuals and groups are permitted and encouraged to report incidents, often with clear guidelines and policies, and usually to a government agency or professional groups.
- **Citizen reporting:** This is similar to voluntary reporting, but incidents are reported by the public and organizations acting as watchdogs.

The choice between implementing voluntary or mandatory reporting systems varies across sectors and industries, as do the disclosed information and criteria for reporting. Mandatory reporting requires obligated actors to report covered incidents, and non-compliance may result in legal consequences. Conversely, in voluntary reporting, organizations and relevant actors are encouraged but not required to report incidents. Both policy approaches have advantages and disadvantages, and various sectors have adopted different policies to address their needs.

Evaluating the effectiveness of different incident reporting policy frameworks is crucial to gauging how comprehensively and reliably they can capture the full dimensions and extent of AI harm.

Current AI Incident Reporting

Reporting and tracking AI incidents will play a crucial role in understanding AI harms, facilitating the development of effective policy tools and measures to mitigate these harms, and reducing the potential risks associated with AI. Present AI incident reporting databases are primarily based on citizen reporting. Citizen reporting is invaluable, but insufficient to exhaustively and reliably document incidents. One challenge is that the data gathered from different incident reporting frameworks have different structures, making it difficult to compare data points across the different databases and time-consuming to analyze and extract meaningful information. Furthermore, noticeable gaps exist within the dataset. Since public citizens voluntarily report incidents, there is no guarantee that all incidents are captured. Policies for establishing mandatory and voluntary incident reporting will be necessary to address these data deficiencies.

A survey of existing AI incident reporting databases that track and collect information on AI incidents yielded less than a handful of key players:

- AI Incident Database (AIID)⁸
- AI, Algorithmic, and Automation Incidents and Controversies Repository (AIAAIC)⁹
- AI Vulnerability Database (AVID)¹⁰
- AI Litigation Database¹¹

The AVID emphasizes identifying vulnerabilities* in AI systems, while the AI Litigation Database focuses on documenting AI-related legal cases.¹² As a result, AIID and AIAAIC are the only two key players in AI incident databases that attempt to actively capture all publicly available data related to AI harms and issues. Independently founded and operated, these two databases rely on public submissions of media reports covering AI incidents. However, they have developed different classification

*The AVID defines vulnerability as any weakness in AI systems that can cause incidents.

frameworks, which hamper comparable and complete documentation of all AI incidents.

AI, Algorithmic, and Automation Incidents and Controversies Repository

The AI, Algorithmic, and Automation Incidents and Controversies Repository is an independent collection of incidents and controversies about AI and AI-related technologies that started in June 2019.¹³ As of July 2023, the Repository had more than 1,100 entries on incidents and controversies relating to AI, algorithms, and automation.¹⁴

The AIAAIC Repository is maintained by an editorial team consisting of contributors who identify incidents and process public submissions of media reports using a six-step framework: detect, assess, classify, summarize, approve, and publish. Reports are assessed based on relevance, impact, credibility, and volume before being added to the Repository. The AIAAIC Repository displays incidents on a Google sheet where users with various access levels can view data, modify data, and provide comments, making the Repository a live database.

The repository excludes reports involving certain technologies and issues, such as geopolitical issues, legislations and standards, and quantum computing. Notably, artificial general intelligence and artificial superintelligence—both AI-related topics—are also on the exclusion list, presumably because they do not currently exist or are considered hypothetical concepts. The AIAAIC analysis of harm reflects an organizational viewpoint in which the negative impacts caused by AI systems occur either internally or externally of the organization that developed or deployed the AI system. External harms are negative impacts on individual users or stakeholders, society, and the environment, whereas internal harms affect the business reputation, operations, finances, and compliance of the organization that developed or deployed the AI system.

AI Incident Database

The AI Incident Database (AIID) started in May 2018 and was launched publicly in November 2020.¹⁵ The database is sponsored by the UL Research Institutes, an independent safety science organization with a global reach.¹⁶ The Responsible AI

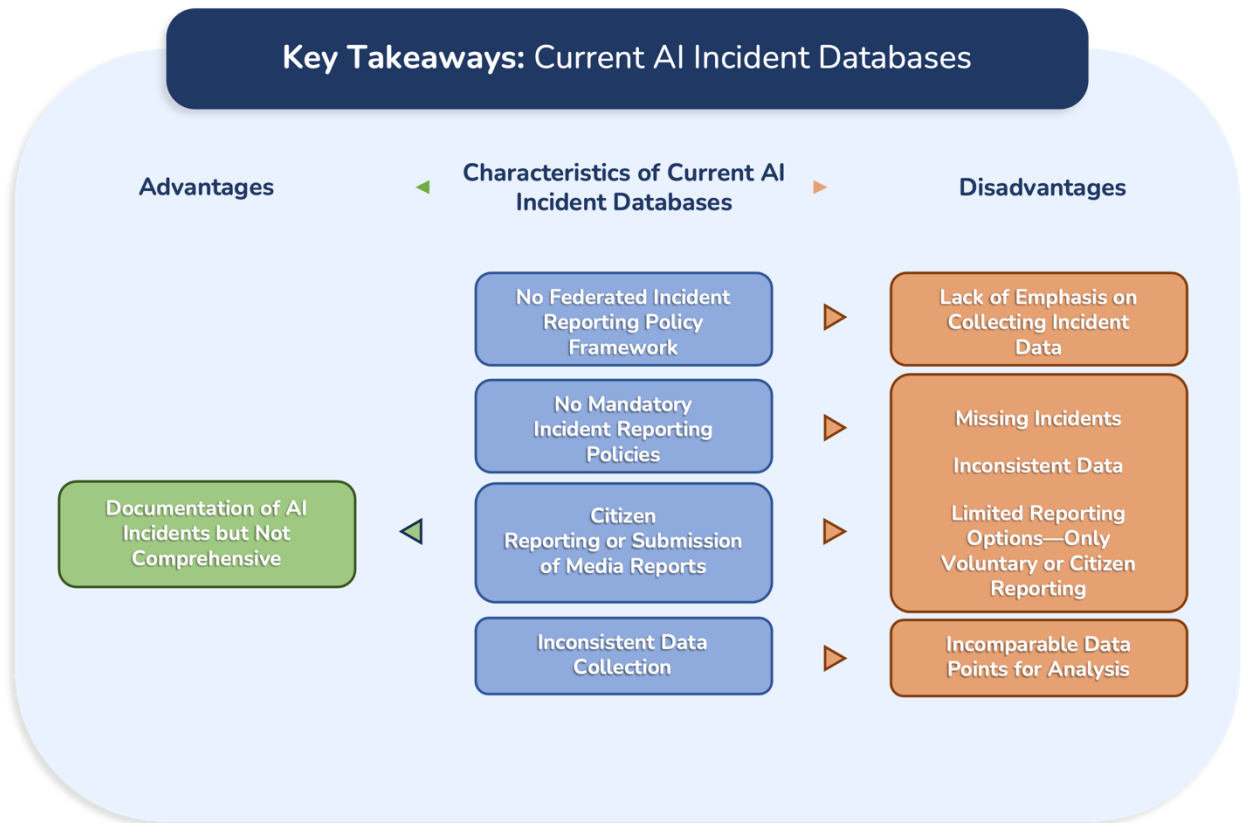
Collaborative, an organization chartered to oversee the incident database, edits the content in the AIID. The AIID collects and catalogs AI incidents through the public submission of media reports covering a wide variety of AI incidents.

As of September 2023, there were 2,813 incident reports connected to 547 unique incidents in the database. Each incident may have one or multiple reports published on the database, reported by various media outlets and providing diverse viewpoints on incidents. Submissions are reviewed and indexed by internal and volunteer editors before being published on the online database. Incidents are currently annotated with two taxonomies: the Goals, Methods, and Failures taxonomy and the Center for Security and Emerging Technology's AI Harm Taxonomy.¹⁷

Key Takeaways: Current AI Incident Databases

While AIID's and AIAAIC's work is valuable in setting the early foundation and infrastructure for documenting AI incidents, both initiatives have developed separate taxonomies and classifications for defining AI harms. For the most part, the databases emphasize and collect different information for each incident. Their conflicting definitions of harm and risks obstruct drawing parallels between the databases that could have contributed to comparable research in AI safety. See Figure 2 for an overview of the key takeaways from current AI incident databases.

Figure 2. Key Takeaways: Current AI Incident Databases



Government Initiatives for AI Incident Reporting

After reviewing the existing AI incident reporting databases, we surveyed emerging AI government initiatives from around the world to assess their provisions for reporting AI incidents and identify potential gaps to address. Of the countries we surveyed, China was the only country that has promulgated AI-related rules that include provisions for incident reporting, while the European Union, Brazil, and Canada have proposed legislative initiatives that include provisions for incident reporting.

China

In the last two years, China has released a series of AI-related rules to address the emerging harms and risks associated with AI. In 2022, the *Provisions on the Management of Algorithmic Recommendations in Internet Information Services* and the *Provisions on the Administration of Deep Synthesis Internet Information Services* came into effect to regulate algorithmic recommenders and deepfakes.¹⁸ In 2023, its *Interim Measures for the Management of Generative Artificial Intelligence Services* came into effect to mitigate the rising concerns over generative AI.¹⁹ Within these three AI-related rules, AI service providers are required to report any violations to relevant authorities, as well as establish reporting mechanisms for the public to lodge complaints and provide feedback on their services.

European Union, Brazil, and Canada

Meanwhile, legislative proposals from the European Union, Brazil, and Canada include requirements for specific AI actors (AI developers, research labs, companies, organizations, and operators) to report incidents to relevant authorities.

The crux of the *Proposal for Laying Down Harmonised Rules on Artificial Intelligence* (EU AI Act) is its risk-based approach that classifies AI systems into unacceptable-risk, high-risk, limited-risk, and minimal-risk systems.²⁰ Regarding AI incident reporting, the EU AI Act requires developers of high-risk AI systems to report any serious incidents or malfunctioning to the corresponding authorities in the Member States where they occurred. A serious incident or malfunctioning constitutes a violation of fundamental rights in the European Union. Reports must be made immediately or within 15 days of

the incident. Conversely, providers of minimal-risk systems are encouraged to follow voluntary codes of conduct instead of mandatory obligations.

High-Risk AI Systems Under the Proposed EU AI Act

1. Biometric identification and categorization of natural persons
2. Management and operation of critical infrastructure
3. Education and vocational training
4. Employment, workers management, and access to self-employment
5. Access and enjoyment of essential private and public services and benefits
6. Law enforcement
7. Migration, asylum, and border control management
8. Administration of justice and democratic processes

Source: *Proposal for Laying Down Harmonised Rules on Artificial Intelligence, Annex III*

Brazil's draft legislation on AI regulation reflects the EU AI Act, using similar provisions. The Brazilian proposal entails reporting obligations imposed on both providers and operators to inform authorities of severe incidents that pose risks to human life, critical infrastructure, property, environmental damage, and infringements upon fundamental human rights.²¹ Similarly, in 2022, Canada's proposed *Artificial Intelligence and Data Act* outlined requirements for individuals responsible for high-impact systems to notify the Ministry of Innovation, Science, and Economic Development in situations where the system has caused substantial harm or presents a significant likelihood of causing such harm.²²

United States

In 2023, the U.S. government announced several AI policy initiatives that included AI incident reporting. *Executive Order 14110* directed the identification, collection, and investigation of AI incidents emerging from the healthcare sector and incidents related to intellectual property.²³ Additionally, the Executive Order instructed the Department of Homeland Security to establish an AI Safety and Security Board to provide the government with recommendations for incident response related to AI usage in critical infrastructure. Following the Executive Order, the National AI Advisory Committee released its recommendations for piloting an adverse AI event reporting system.²⁴ NAIAC focused its recommendations on reporting the most concrete and severe events, such as those involving national security risk, substantial injury and damage, and death to existing regulatory authorities.

Before the Executive Order was published, the U.S. National Institute of Standards and Technology (NIST) released the *AI Risk Management Framework* that includes a set of voluntary guidelines for sharing incident data among relevant AI actors and affected communities.²⁵ In July of the same year, the White House announced it had secured voluntary commitments from the seven leading AI companies in the United States—Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI—that include enabling third-party discovery and reporting vulnerabilities in their AI systems. The last two initiatives did not specify reporting incidents to an external entity.²⁶

Key Takeaways: Government Initiatives on AI Incident Reporting

Despite outlining obligations for AI developers and providers to report incidents, the legislative initiatives from China, the European Union, Brazil, and Canada did not include clear recommendations for implementing consistent federated incident reporting frameworks and data collection. A plausible implication could be that organizations that collect AI incidents might emphasize collecting different data types, and discrepancies might appear in their data management.

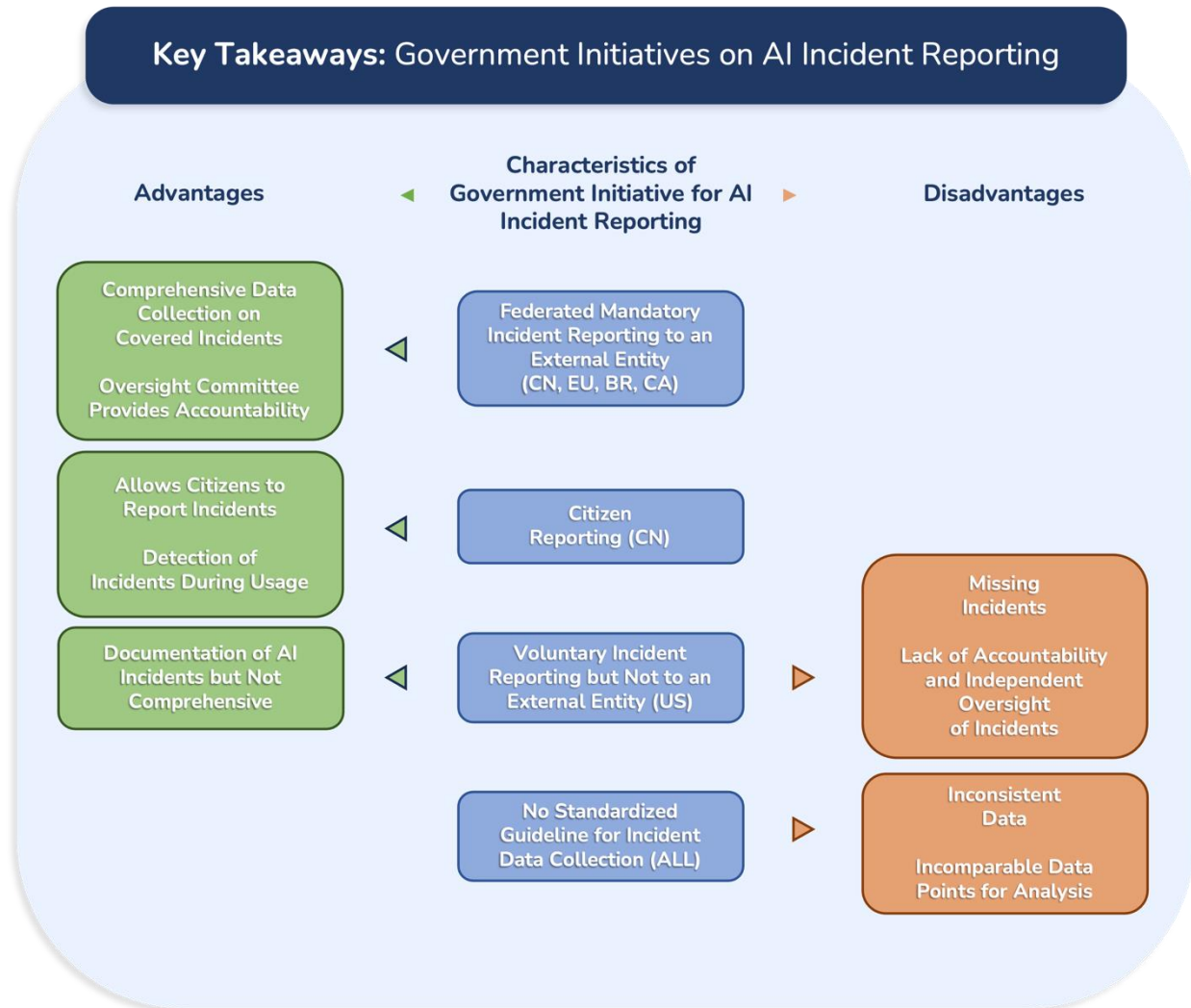
The rules from China were the only legislative initiative that addressed citizen reporting on AI incidents. The European Union, Brazil, and Canada proposals did not extend incident reporting provisions to other stakeholders, such as the public, that could potentially experience AI harm. Nevertheless, the reporting obligations outlined

in the rules from China emphasize mitigating illegal activities and rule violations, which could limit the scope of incidents that get reported. For instance, system vulnerabilities and adversarial attacks might not fall within this scope, and data collection on these incidents would be absent, leaving a key dimension of AI harms missing from the overall picture.

The Executive Order demonstrates the U.S. government's intent to capture AI incidents. However, the Executive Order primarily covers AI incidents involving critical infrastructure, IP, and healthcare, which excludes AI incidents that occur in other domains. NAIAC's recommendations to report AI incidents to existing regulatory authorities could result in incidents falling in between regulatory gaps, given the general-purpose nature of most AI capabilities. Moreover, NAIAC's narrow focus on reporting tangible harm excludes intangible harm that is equally impactful and commonly occurring, such as bias and discriminatory decisions resulting in differential treatment.

The current voluntary approach in the United States does not guarantee all relevant AI actors will implement reporting mechanisms, or utilize one that is comprehensive and consistent across organizations. Furthermore, reporting guidelines outlined in both documents do not suggest collecting and reporting AI incidents to an independent external entity. The reporting mechanisms proposed in these guidelines suggest that incident data is funneled back into the companies and organizations of the AI actors, impeding transparency and accountability when AI harm occurs and limiting information-sharing on AI vulnerabilities. It is undetermined what regulations and policies on reporting AI incidents the U.S. government will announce in the coming years. For an overview of the key takeaways from government initiatives on AI incident reporting, see Figure 3.

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



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




Lessons from Incident Reporting Policies in Healthcare, Transportation, and Cybersecurity Sectors

To understand which type of incident reporting policy framework would work best for recording AI incidents, we look to learn lessons from incident reporting in other fields. We chose three high-risk sectors that have established incident reporting frameworks:

- Healthcare
- Transportation
- Cybersecurity

In each sector, we analyze its reporting structure, policy evolution, and the impact of its distinct policy approach on the outcomes of its incident reporting initiatives. Table 1 provides an overview of these three sectors' various incident reporting systems.

Table 1. Overview of Incident Reporting in Healthcare, Transportation, and Cybersecurity

	 Healthcare	 Transportation	 Cybersecurity
	<i>Who governs and collect incident reports?</i>		
Governing Authority	<ul style="list-style-type: none"> • State governments • Independent organizations 	<ul style="list-style-type: none"> • National Transportation Safety Board • National Highway Traffic Safety Administration 	<ul style="list-style-type: none"> • Cybersecurity and Infrastructure Security Agency • Domain-specific agencies
	<i>Mandatory/Voluntary/Citizen</i>		
Reporting Structure	<ul style="list-style-type: none"> • Mandatory • Voluntary 	<ul style="list-style-type: none"> • Mandatory • Voluntary • Citizen 	<ul style="list-style-type: none"> • Mandatory • Voluntary
	<i>What events are reported?</i>		
Reporting Scope	<ul style="list-style-type: none"> • Serious harm • Death • Near-misses 	<ul style="list-style-type: none"> • Significant accidents in aviation, highway, marine, railroads, pipelines, and accidents involving hazardous materials 	<ul style="list-style-type: none"> • Cyber incidents in specific domains • Significant cyber incidents
	<i>Who must/can report?</i>		
Reporting Entity	<ul style="list-style-type: none"> • Healthcare practitioners • Members of the organizations 	<ul style="list-style-type: none"> • Civilians, transport operators, manufacturers, government agencies, etc. 	<ul style="list-style-type: none"> • Providers of critical infrastructure • Federal agencies • Financial institutions
	<i>How is information collected?</i>		
Disclosure Format	<ul style="list-style-type: none"> • Various disclosure formats 	<ul style="list-style-type: none"> • Investigations • Automated data collection mechanism, e.g. black box • Online reporting 	<ul style="list-style-type: none"> • Specified disclosure formats • Specified reporting timeframe
	<i>How is the initial incident report used or augmented?</i>		
Post Reporting	<ul style="list-style-type: none"> • Review and implement safety measures 	<ul style="list-style-type: none"> • Issue safety recommendations • Enforce safety standards 	<ul style="list-style-type: none"> • Provide assistance for victims • Identify trends across sectors • Share information with network defenders • Warn other potential victims
	<i>Can the public access incident records?</i>		
Public Database	<ul style="list-style-type: none"> • No 	<ul style="list-style-type: none"> • Yes 	<ul style="list-style-type: none"> • Mixed

Healthcare: Inconsistent Incident Reporting

Background on Healthcare Incident Reporting

The impetus for systematic tracking and learning from medical errors gained momentum following the publication of the pivotal book *To Err is Human* by the Institute of Medicine in 2000.²⁷ The book's revelation that up to 98,000 hospital deaths resulted from medical errors annually brought attention to the issue as a pressing public concern that demanded immediate policy intervention. The Centers for Disease Control and Prevention estimates about 1 in 31 hospital patients each day acquires healthcare-associated infection, costing billions of dollars in added expenses to the U.S. healthcare system—and around half of those cases could be preventable.²⁸ Agencies and organizations established numerous incident reporting frameworks to mitigate medical errors' financial and human costs. Medical errors were tracked at various levels: organizational-based, state-based, and national-based. These frameworks range from mandatory reporting to voluntary, and are endorsed by a variety of independent and government-led organizations.

Despite the severity of medical errors and their significant occurrence rate, incident reporting in healthcare is incongruous across the United States. There is considerable variation in the types of events that are mandatory to report, along with disclosure requirements. Below, we examined a selection of the more commonly used frameworks.

- The **National Quality Forum (NQF)** is a nonprofit organization that developed Serious Reportable Events (SREs) in healthcare, which is a set of definitions and standards that some states have used to implement their own mandatory reporting systems.²⁹
- The **Agency for Healthcare Research and Quality (AHRQ)** is a government agency that possesses a repository of patient safety incidents reported voluntarily by entities registered under its Patient Safety Organization Program.³⁰

- **The Joint Commission** is an accreditation organization that launched its sentinel event* reporting system to document adverse patient incidents and encourage accredited organizations to report these events.³¹

The absence of a clear, federally mandated central reporting system for serious healthcare events has contributed to inconsistent efforts in documenting adverse outcomes in patient safety.³² The NQF developed a set of voluntary standards (SREs) that states can adopt in their incident reporting frameworks.³³ More than half the states and the District of Columbia have implemented mandatory reporting based on the SREs standards, and yet there are still discrepancies in how they utilize, implement, and view the reporting of different patient safety events. The critical variations in implementing SREs within states' mandatory reporting frameworks are:

- **State-defined lists** do not include any of the language within NQF's SREs, but may use NQF's standards or others as a launching pad.
- **Modified NQF lists** reference the SREs but add, remove, or modify NQF's events or definitions. A list can be classified as "modified" even by removing one SRE.
- **NQF's SREs** are used entirely and exactly as written for creating legislation.

Voluntary reporting systems can suffer from underreporting, resulting in databases that do not accurately capture the full spectrum of prevalent safety issues. As a result, the precision of incident trend analysis is also diminished.³⁴ The AHRQ manages the Network of Patient Safety Databases, a repository of patient safety incidents reported voluntarily by entities registered under the Patient Safety Organization Program. However, there are currently only 103 registered providers in the Program, which is a tiny fraction of the 6,129 hospitals in the United States.³⁵ Due to the limited number of registered providers and the voluntary nature of their reporting system, the AHRQ admits that their database "does not contain a representative sample of patient safety concerns and cannot be used to calculate the actual incidence or prevalence of patient safety events."³⁶

*A sentinel event is a patient safety event that results in death, permanent harm, or severe temporary harm.

Additionally, there is a staggering difference between the number of incidents recorded in a mandatory reporting system and a voluntary system when comparing The Joint Commission's sentinel event reporting (voluntary) and the New York Patient Occurrence Reporting and Tracking System (NYPORTS) (mandatory). An analysis from 2005 found that NYPORTS recorded 11,028 adverse events between 1998 and 2003, while the national voluntary reporting system run by The Joint Commission collected a mere 176 incidents from the state within a similar timeframe.³⁷

The significant gap in reported incidents between mandatory and voluntary systems raises questions about the efficacy of a voluntary framework as a reliable mechanism for improving safety practices and its ability to represent occurring harms accurately. Voluntary reporting systems are likely to miss valuable data needed to inform and improve safety measures.³⁸ To underline the usefulness of mandatory reporting, New York State—which requires mandatory reporting—has used data from its NYPORTS database to formulate protocols that reduce incident occurrences. For instance, data analysis of wrong-patient/wrong-site events (the severe error of performing a medical procedure on the wrong patient or performing surgery on the wrong place of the body) led to new protocols in 2001 that helped reduce such incidents from 25 events in 2002 to 17 events in 2003 in New York State.³⁹

Transportation: Investigative Data Collection

Transportation-related safety issues in the United States are primarily overseen by the National Transportation Safety Board (NTSB) and the National Highway Traffic Safety Administration (NHTSA).

All significant accidents and crashes in aviation, highways, marine, railroads, pipelines, and hazardous materials must be reported to the NTSB, which then carries out investigations to identify the root causes.⁴⁰ The NTSB utilizes information gathered from automated data-collecting sensors and event recorders in aircraft, cars, and vessels to assist in their investigations. These automatic data collection mechanisms record crucial technical and contextual information that can help identify the root causes of incidents.

The NTSB utilizes the acquired data and results from its investigations and research to construct its Most Wanted List: a compilation of safety recommendations to prevent

accidents, reduce injuries, and save lives.⁴¹ For instance, presently, the NTSB has highlighted the need for standardized alcohol and drug testing to prevent impairment-related crashes on highways, and a ban on personal electronic devices while driving to prevent distracted driving. The NTSB also advocates for these recommendations in state legislation, proposes regulatory amendments, suggests procedural adjustments by operators, and urges professional associations to inform their members about relevant safety issues.

While the NTSB doesn't have the authority to require recipients to implement their safety recommendations, the NHTSA can enforce safety standards and regulations.⁴² Using its authority, the NHTSA has issued mandatory reporting on certain incidents, such as the 2021 order requiring manufacturers and operators of vehicles equipped with automated driving systems and advanced driver assistance systems to report crashes.⁴³

Between 2016 and 2021, the NHTSA investigated 42 crashes that likely involved driving assistance systems.⁴⁴ Since 2021, however, the NHTSA recorded a total of 522 crashes involving various levels of automated driving systems just from data collected between July 2021 to May 2023.⁴⁵ The substantial increase in incident reports following the announcement of the 2021 order was likely supported by the mandatory reporting approach. As vehicles with various levels of automated driving systems become more commonly used on public roads, understanding the potential safety issues and trends in automated driving systems can be enhanced by the number of incidents reported and data collected.

Apart from issuing orders for specific incident reporting obligations, the NHTSA also provides a citizen reporting portal on their website, where individuals can report safety concerns related to their vehicle, tires, car seat, or equipment.⁴⁶ Data collected from this portal helps the NHTSA detect safety issues from vehicle usage, launch investigations on possible defects, and initiate safety recalls when necessary. Also observed in the transportation sector is the Aviation Safety Reporting System, which is a voluntary reporting system emphasizing human performance in the aviation industry. The ASRS receives reports on both unsafe occurrences and hazardous situations, submitted by pilots, air traffic controllers, dispatchers, cabin crew, maintenance technicians, unmanned aircraft systems crew, and others.

Cybersecurity: A Shift to Mandatory Reporting

In recent years, the U.S. government has begun implementing a succession of new regulations and guidelines to make reporting cyber incidents mandatory in various domains. Previously, cyber incident regulations primarily focused on infrastructure resilience and data privacy to manage cyber incidents and keep organizations accountable. There was no widely established federal policy framework mandating reporting of cyber incidents, and the limited emphasis on collecting incident data also meant that learning from previous incidents was less of a priority.⁴⁷

NIST, in collaboration with MITRE, launched the U.S. National Vulnerability Database in 2005, to provide a collection and knowledge base of cybersecurity vulnerability incidents.⁴⁸ The NVD was developed upon and synchronized with the Common Vulnerability Enumeration (CVE) list, which is a voluntary reporting framework operated by MITRE that was launched publicly in 1999.⁴⁹ In 2016, NIST released its “Guide to Cyber Threat Information Sharing,” which recommended sharing Cyber Threat Information—that is, any information that can help an organization identify, assess, monitor, and respond to cyber threats, including findings from analyses of incidents—to improve cybersecurity within organizations.⁵⁰ Surveys showed a growing trend of organizations using CTI, but the use of CTI faced several challenges stemming from the absence of a federal mandate for an incident reporting policy framework.⁵¹ Organizations struggled to find reliable and comprehensive sources of CTI, and it was unclear what information could be shared, how it could be shared, and whether their information-sharing practices were compliant.⁵²

Definition of Cyber Incidents

Cyber incident. An event occurring on or conducted through a computer network that actually or imminently jeopardizes the integrity, confidentiality, or availability of computers, information or communications systems or networks, physical or virtual infrastructure controlled by computers or information systems, or information resident thereon. For purposes of this directive, a cyber incident may include a vulnerability in an information system, system security procedures, internal controls, or implementation that a threat source could exploit.

Significant cyber incident. A cyber incident that is (or a group of related cyber incidents that together are) likely to result in demonstrable harm to the national security interests, foreign relations, or economy of the United States, or to the public confidence, civil liberties, or public health and safety of the American people.

Source: Presidential Policy Directive 41 (PPD-41): United States Cyber Incident Coordination.⁵³

The rapid emergence of new technologies coupled with the advancement of AI resulted in the proliferation of CyberAI threats that brought a new urgency to the field of cybersecurity.⁵⁴ As these technologies become increasingly integral to both public and private sectors, the U.S. government has recognized the existence of the crucial gap in cyber incident reporting and has been actively formulating initiatives to address it.

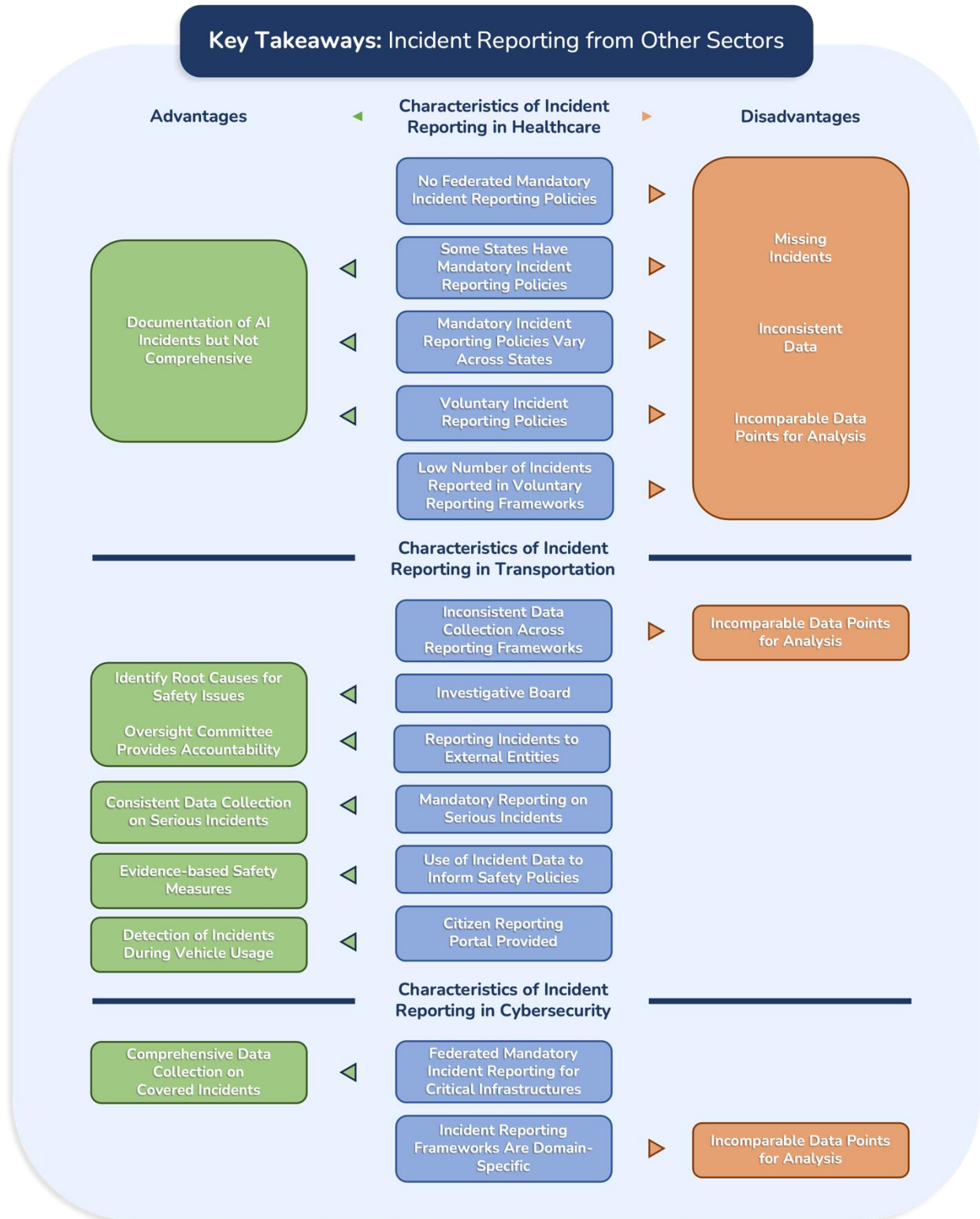
This shift signified a departure from soft laws—such as standards—and reflects the U.S. government's commitment to improving its understanding of cyber incidents and bolstering its response and resilience to future threats. These emerging proposals and regulations pertain to various entities, including financial service providers, critical infrastructure providers, and public companies.

- The **Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCI)** was signed into law in March 2022, directing the Cybersecurity and Infrastructure Security Agency (CISA) to develop and implement mandatory incident reporting.⁵⁵ CIRCI requires providers of critical infrastructure to report substantial cyber incidents to CISA within 72 hours, while ransomware attacks where payment occurred must be reported within 24 hours.
- The **Federal Information Security Modernization Act of 2014 (FISMA)** requires federal Executive Branch civilian agencies to alert CISA on cybersecurity incidents involving their information and information systems.⁵⁶ FISMA includes guidelines to align incident reporting disclosure information and a one-hour notification timeframe; it also moved root cause analysis to the end of the incident-handling process to allow agencies to notify response teams sooner.
- The **Securities and Exchange Commission** has adopted new requirements for public companies to report and disclose security breaches or incidents.⁵⁷
- The **Office of the Comptroller of the Currency—Treasury, the Board of Governors of the Federal Reserve System, and the Federal Deposit Insurance Corporation** have issued a rule requiring banking organizations to notify their federal regulator of cyber incidents within 36 hours.⁵⁸

- The **National Credit Union Administration Board** has approved a final rule requiring federally insured credit unions to notify them of cyber incidents within 72 hours.⁵⁹

Cyber incidents have been around for decades, and it is only recently that federated incident reporting policy frameworks are being established to document them systematically, improve information sharing and situational awareness of incidents among response teams, and shorten incident response time. The emerging trend of mandatory cyber incident reporting policies could plausibly set a strong foundation and incentive for the early adoption of incident reporting policies in the field of AI.

Figure 4. Key Takeaways: Incident Reporting From Other Sectors



Discussion

Our analysis of the two AI incident reporting databases, emerging government initiatives related to AI incident reporting, and the various incident reporting systems in the healthcare, transportation, and cybersecurity sectors revealed disadvantages and advantages. These insights offered several important lessons that can be applied to an AI incident reporting policy framework, as discussed in the following:

- Limited incident reporting frameworks are inadequate.
- Inconsistent data collection creates meaningless data.
- There is a need for a federated AI incident reporting framework.
- Incident investigation supports effective safety policies.

Limited Incident Reporting Frameworks Are Inadequate

Across the board, the incident reporting initiatives examined in this paper often emphasized either citizen, voluntary, or mandatory reporting, typically focusing on one or two of these reporting categories. In isolation, each of these three frameworks has limitations. Adopting a hybrid framework that incorporates all three reduces the limitations.

Our assessment of reporting frameworks in the healthcare sector demonstrates that relying on voluntary reporting alone may result in low numbers of reported incidents, and potentially miss incident data. The low numbers may be attributed to the lack of incentive for entities and organizations to report incidents without a reporting obligation. Consequently, voluntary reporting is unlikely to be a reliable and sufficient method for capturing an impactful and comprehensive AI incident landscape.

The legislative initiatives examined in this paper have embraced mandatory AI incident reporting in their AI policies, underscoring a consensus on the importance of collecting and documenting AI incidents. However, limiting incident reporting regulations to mandatory obligations may miss out on incidents that don't fall within regulatory scopes. Supplementing a mandatory reporting framework with voluntary and citizen reporting (similar to those outlined by China) can help identify out-of-scope incidents

and detect novel incidents that emerge during usage. This could be particularly useful as AI systems are usually trained under controlled conditions that do not fully reflect the real-world context in which they are deployed.

Employing a hybrid incident reporting framework to collect and document AI incidents will be crucial for documenting a wide array of AI harms and harm dimensions. As AI continues to advance and become more prevalent, AI harms can be expected to grow both in scale and severity.⁶⁰ The information gathered from incident reports will be essential for policymakers and researchers to gain a more thorough insight into the potential risks associated with AI, and to develop effective safety regulations to reduce the reoccurrence of AI harm.

Inconsistent Data Collection Creates Meaningless Data

As AI systems are developed and deployed throughout a wide range of sectors and applications, their impacts will extend across regulatory jurisdictions and geographical boundaries. Relying on state initiatives or domain-specific guidelines will likely not be adequate for aggregating AI incident data that can accurately depict the many dimensions of AI harm. This is evident in our discussion of incident reporting in the healthcare sector, where states have adopted the NQF's Serious Reportable Events differently. The differences have made it difficult to aggregate a national dataset on medical incidents to identify healthcare safety trends and systemic issues. Such incongruencies could significantly undermine efforts to identify harmful trends, system vulnerabilities, and the safety measures needed to mitigate risks associated with AI.

On the other hand, the mandatory incident reporting policies in cybersecurity delineated clear instructions for information disclosure, and specific notification timeframes. Clear and intentional guidelines like these may enable timely reporting, improve information sharing, and engender greater data quality and quantity for understanding and mitigating AI incidents. Furthermore, a standardized disclosure guideline would greatly assist the development of a robust taxonomy and classification framework on AI harms that can enhance information sharing and research on AI safety by enabling comparable data points for analysis. The definitions and classification of AI harms will be foundational when developing an AI incident reporting framework to

accurately capture the data that will promote our understanding of the various dimensions of emerging AI harms and risks.

There Is a Need for a Federated AI Incident Reporting Framework

Implementing a federated framework for AI incident reporting is essential as AI is developed and deployed across sectors and applications. A federated approach provides a centralized framework prescribed by a singular authoritative government body or the federal government. The framework stipulates a set of minimum requirements that can be adapted and implemented across government agencies and non-governmental organizations. A federated AI incident reporting framework can promote comprehensive and consistent collecting, documenting, and sharing of AI incident data. Conversely, relying on individual regulatory agencies or sector-specific frameworks could result in fragmented efforts and inconsistent data.

Legislative initiatives from China, the European Union, Brazil, and Canada suggest a growing consensus on reporting AI incidents to mitigate rising concerns about AI harms. The U.S. government has not yet announced significant legislative initiatives outlining a federated AI incident reporting policy framework that includes reporting to external oversight entities. Presently, the U.S. approach to AI incident reporting is generally limited to voluntary and citizen reporting, and maintains its strategy of directing government organizations to regulate AI incidents within their domains. This could increase the risk of engendering fragmented incident reporting frameworks, such as those observed in the healthcare sector. The value and emphasis different authorities will place on establishing an AI incident reporting framework and database will likely vary. This diversity can impact data collection in each domain, making it difficult—if not impossible—to aggregate, analyze, and understand trends in AI incidents across sectors. As a result, developing comprehensive measures to mitigate AI harms becomes more challenging.

In the healthcare sector, the absence of a federated AI incident reporting policy framework impacted incident data collection efforts. Incident reporting initiatives were fragmented and inconsistent, making it difficult to identify comparable data points for analysis. In contrast, the transportation sector and the cybersecurity sector have clear policies and standardized rules for reporting incidents. The NTSB and NHTSA have

established incident reporting and investigative mechanisms, facilitating a robust system for documenting incidents, identifying root causes, and developing evidence-based safety policies. In the cybersecurity sector, until recently there were uncertainties about how incident information could be shared safely and compliantly, which hindered incident response efforts. However, this will be changing as the U.S. government has now made incident reporting in cybersecurity mandatory. This move demonstrates the U.S. government's commitment to and endorsement of collecting, documenting, and sharing data on cybersecurity incidents.

Incident Investigation Supports Effective Safety Policies

A safety investigation board or research team has been advantageous in identifying root causes of transportation incidents. This has led to the implementation of evidence-based, life-saving measures and safety regulations in the U.S. transportation sector. The NTSB's and NHTSA's approach to using incident investigation facilitates a direct link between data collection and policy responses, leveraging in-depth investigations to support informed decision-making. An investigative safety board would be equally useful for conducting root-cause analysis of significant AI incidents and providing feedback to help AI actors improve their design and development, and enable policymakers to craft effective regulations and educate the public on AI safety. A safety investigation board or safety research team could contribute valuable technical and contextual data to understanding AI harm.

Recommendations

Policies promoting the establishment of a federated AI incident reporting framework would ensure a more comprehensive collection of AI incidents data and facilitate the development of an authoritative classification system for extracting meaningful data and trends on AI harm. This data would support in-depth research on AI safety and system vulnerabilities, enhance our ability to understand potential AI risks, and equally important, help policymakers develop more effective safety regulations and practices to mitigate AI harm.

Based on the observations discussed above and the nature of AI as a general-purpose technology, we make the following recommendations to address the current gap in AI incident reporting.

- Establish clear policies for federated hybrid AI incident reporting.
- Develop a standardized and authoritative classification system.
- Create an independent AI incident investigation agency.
- Explore automated data collection mechanisms.

Establish Clear Policies for Federated Hybrid AI Incident Reporting

Policymakers should establish a federated AI incident reporting policy framework to gather incident data across sectors and applications, involving a hybrid of mandatory, voluntary, and citizen reporting. It should include clear guidelines on implementing a hybrid of mandatory, voluntary, and citizen reporting policies. Reporting should be made to an independent external committee (government agency, professional association, oversight body) to promote transparency and accountability in AI incident management. The incident reporting policy framework should be incorporated into national legislative AI proposal packages to ensure a comprehensive implementation across sectors and applications.

Mandatory Reporting

Relevant AI actors should be mandated to report covered incidents promptly. The rise in regulatory mandates for reporting cyber incidents signals the U.S. government's commitment to enhancing resilience and safety in cyber technology. Policymakers should leverage this shift and advocate for similar support in implementing mandatory incident reporting in AI. Mandatory reporting can promote consistent AI incident reporting, prevent data gaps across sectors and applications, and provide a comprehensive knowledge base on AI harm that can inform research on AI safety and risks. An in-depth assessment will be necessary to define covered incidents involving the types of harm, scale, and AI actors to make the reporting obligation proportionate.

Voluntary Reporting

Voluntary reporting frameworks should also be established alongside the mandatory framework to capture AI incidents outside the mandatory jurisdiction. AI actors should be permitted and encouraged to report AI incidents that fall outside regulatory scopes voluntarily, usually to a government agency or professional groups. This would have lesser compliance obligations on AI actors compared to mandatory reporting. Though voluntary, the data collected from voluntary reporting can enhance the overall data fidelity of documented AI incidents. Supplementing the mandatory reporting framework with a voluntary option may also reduce resistance to implementing a mandatory framework.

Citizen Reporting

Similarly, aligned with the values of democratic governance, an easily accessible reporting framework should be made available for citizen reporting to document AI incidents. While AI system providers and operators should be required to report AI incidents, other stakeholders and the public should also be able to report AI harm they may have experienced. When designing a citizen reporting system, special attention should be given to vulnerable populations and underrepresented communities—groups disproportionately affected by biased AI systems. Relevant stakeholders should be included meaningfully in the development process of the reporting system to ensure their needs and concerns are adequately addressed and incorporated.

Develop a Standardized and Authoritative Classification System

The AI incident reporting framework should include a standardized set of disclosed information plus accommodations for the unique characteristics of distinct domains, such as privacy concerns and other regulatory requirements. Standardizing the disclosure system can promote greater consistency in the collected data, allow comparable analyses, and reduce the risk of missing crucial information from incidents across different domains. Implementing a standardized disclosure system can also contribute to developing a robust classification framework on AI harm, providing a common foundation for identifying AI harm and thus enhancing our analysis on the subject.

Create an Independent AI Incident Investigation Agency

When a significant AI incident occurs, an independent board should investigate the root cause and objectively analyze the incident.* This will provide extra scrutiny over significant AI incidents, keep AI actors accountable, and retrieve valuable technical and contextual data about the incidents. AI actors should be compelled to design AI systems with mechanisms supporting investigations and data collection.

Furthermore, establishing an investigative agency will help ensure appropriate response measures for significant AI incidents. This agency will play a vital role in addressing and mitigating adverse consequences resulting from AI use. Outcomes from these investigations will provide key insights into significant incidents, enabling the agency to recommend safety regulations to reduce the risk of similar incidents from reoccurring.

Explore Automated Data Collection Mechanisms

Automated data collection mechanisms—such as flight recorders—can provide crucial technical and contextual information that facilitates root-cause analysis of accidents, one of the methods used by the NTSB to collect information on transportation

*The definition of a significant AI incident should be determined during the development of a standardized taxonomy framework.

incidents. Comparable mechanisms for AI systems should be explored. Obtaining technical and contextual information from AI incidents would be highly advantageous. For example, such a mechanism could capture critical information about the system's environment, or a "snapshot" of the model's technical data during an incident. This information could address concerns pertinent to the issues of explainability in AI harms.

At the same time, automated data collection mechanisms in AI systems could raise concerns about proprietary data and security issues. These issues should be addressed thoroughly to avoid pushback from companies. Additionally, automated data-collection mechanisms do not replace other incident-reporting systems. Rather, they supplement the data collected with additional technical and contextual information.

Conclusion

The present moment offers a prime opportunity to establish an AI incident reporting framework with relatively low stakes. However, this window is rapidly closing as AI becomes more prevalent across applications and sectors. A federated, comprehensive, and standardized framework will prevent data gaps and enhance data quality. Adopting a hybrid framework that includes mandatory, voluntary, and citizen reporting will improve data fidelity, providing a more accurate representation of the emerging trends in AI harm and risk. Further research will be necessary to determine the details for operationalizing such an AI incident reporting policy framework.

An AI incident reporting framework must be integrated as an essential component of AI safety rather than developed as an afterthought in AI legislative initiatives. Clear obligations and disclosure requirements should be outlined from the outset to enable frictionless compliance from relevant AI actors. Likewise, easily accessible and comprehensive reporting platforms should be made available to the public so the database may capture novel, unexpected incidents that may emerge during usage.

Lessons from the healthcare, transportation, and cybersecurity sectors provided a compelling argument for implementing a federated mandatory incident reporting system that will positively affect safety practices. In places where mandatory incident reporting has been implemented, evidence of higher reporting rates has been observed and associated with a more positive safety culture and a significant reduction in adverse events.⁶¹ Being at the early stages of AI harm research, the data gathered from a comprehensive and systematic incident reporting system would greatly assist and expedite our knowledge in this area. Policymakers will be better equipped to propose more precise and effective safety regulations, and researchers will gain greater clarity on both the short- and long-term risks associated with AI.

The ability to mitigate AI harms and manage their aftermath competently can shape public conversations about AI usage. Nuclear plant disasters such as Chernobyl, Three Mile Island, and Fukushima have had adverse effects on global perceptions of nuclear energy.⁶² In the aftermath of these instances, public opinions shifted on nuclear energy, and governments either significantly delayed implementation plans or reinforced their stance against nuclear power. Even the German parliament, which has long stood by

technologically safe nuclear power plants, voted to phase out nuclear power plants shortly after the Fukushima disaster in 2011.⁶³ The country closed its last nuclear power plants in 2023.⁶⁴

Presently, a growing percentage of Americans say they feel more concerned than excited about the increased use of artificial intelligence, rising from 38 percent in 2022 to 52 percent in 2023.⁶⁵ A comprehensive incident reporting system can help mitigate these fears by providing valuable insights that can inform effective safety measures, leading to enhanced AI safety and promoting public trust in the technology.

As more data becomes available from AI incident reporting, improving our understanding of AI harms and risks, the policies for the incident reporting framework should be assessed regularly to determine its robustness and capacity for recording and tracking AI incidents. Such iterative practices should be applied to most AI governance initiatives, as there are still uncertainties surrounding emerging technologies and their impact on society.

Authors

Ren Bin Lee Dixon is an AI policy analyst researching AI policies, governance, and ethics.

Heather Frase, PhD, is a senior fellow at CSET and leads the AI Assessment line of research. Dr. Frase serves on the board of the Responsible AI Collaboration (TheCollab), an organization chartered to advance the AI Incident Database and providing editorial oversight for it.

Acknowledgements

For their comprehensive and valuable reviews, we would like to thank Sean McGregor, Violet Turri, Borhane Blili-Hamelin, Mia Hoffman, Mina Narayanan, Josh Goldstein, Helen Toner, and Zach Arnold. Finally, we would like to thank Christian Schoeberl and Jason Ly for their assistance in providing and designing the figures and tables, and Margarita Konaev and Igor Mikolic-Torreira for their feedback and support.



© 2024 by the Center for Security and Emerging Technology. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.

To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc/4.0/>.

Document Identifier: doi: 10.51593/20230046

Endnotes

¹ Ewen Callaway, “‘The entire protein universe:’ AI predicts shape of nearly every known protein,” *Nature* 608, no. 7921 (July 29, 2022): 15–16, <https://doi.org/10.1038/d41586-022-02083-2>; Gary Liu et al., “Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*,” *Nature Chemical Biology*, May 25, 2023, 1–9, <https://doi.org/10.1038/s41589-023-01349-8>.

² Jonas Degraeve et al., “Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning,” *Nature* 602, no. 7897 (February 2022): 414–19, <https://doi.org/10.1038/s41586-021-04301-9>.

³ Mia Hoffman and Heather Frase, “Adding Structure to AI Harm: An Introduction to CSET’s AI Harm Framework,” Center for Security and Emerging Technology (July 2023), 16, <https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/>.

⁴ Julia Angwin et al., “Machine Bias,” ProPublica, May 23, 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>; Larry Hardesty, “Study finds gender and skin-type bias in commercial artificial-intelligence systems,” MIT News | Massachusetts Institute of Technology, February 11, 2018, <https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>; “Amazon Scraps a Secret A.I. Recruiting Tool That Showed Bias against Women,” CNBC, October 10, 2018, <https://www.cnbc.com/2018/10/10/amazon-scraps-a-secret-ai-recruiting-tool-that-showed-bias-against-women.html>; Nicola Davis, “AI skin cancer diagnoses risk being less accurate for dark skin – study,” *The Guardian*, November 9, 2021, sec. Society, <https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study>; Melissa Heikkilä, “Dutch scandal serves as a warning for Europe over risks of using algorithms,” *POLITICO* (blog), March 29, 2022, <https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/>.

⁵ “Incident 545: Chatbot Tessa gives unauthorized diet advice to users seeking help for eating disorders,” AI Incident Database, accessed November 21, 2023, <https://incidentdatabase.ai/cite/545/>.

⁶ “Incident 543: Deepfake of Explosion Near US Military Administration Building Reportedly Causes Stock Dip,” AI Incident Database, accessed November 21, 2023, <https://incidentdatabase.ai/cite/543/>.

⁷ “Incident 550: Tesla Allegedly on Autopilot Struck High School Student Exiting School Bus,” AI Incident Database, accessed November 21, 2023, <https://incidentdatabase.ai/cite/550/>.

⁸ “Welcome to the AI Incident Database,” AI Incident Database, accessed July 19, 2023, <https://incidentdatabase.ai/>.

- ⁹ “AIAAIC,” AI, algorithmic, and automation incidents and controversies, accessed July 19, 2023, <https://www.aiaaic.org/home>.
- ¹⁰ “AVID,” AI Vulnerability Database (AVID), accessed July 19, 2023, <https://avidml.org/>.
- ¹¹ “AI Litigation Database,” Ethical Tech Initiative, accessed July 19, 2023, <https://blogs.gwu.edu/law-eti/ai-litigation-database/>.
- ¹² Carol Anderson et al., “Response from the AI Risk and Vulnerability Alliance to the NTIA AI Accountability Policy Request for Comment,” AI Risk and Vulnerability Alliance (ARVA) 2023, <https://docs.google.com/document/d/1qLHAdH3On2iMmnBh83vEwHhmFErI3QOF8Rw-uSo1EJQ>.
- ¹³ See AIAAIC.
- ¹⁴ “AIAAIC Repository,” accessed September 19, 2023, https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lt0n_4rzLGxFADMLVW0PYI/edit#gid=1051812323.
- ¹⁵ Responsible AI Collective, “Founding Report,” March 28, 2022, <https://docsend.com/view/a45p7mgh44nu8x7j>; “Join the Responsible AI Collaborative Founding Staff,” AI Incident Database, accessed November 21, 2023, <https://incidentdatabase.ai/blog/join-raic/>.
- ¹⁶ See AIID.
- ¹⁷ “What is the GMF Taxonomy?,” Artificial Intelligence Incident Database, accessed July 19, 2023, <https://incidentdatabase.ai/taxonomy/gmf/>. Data from CSET, “CSET’s Harm Taxonomy for the AI Incident Database,” GitHub, accessed November 1, 2023, <https://github.com/georgetown-cset/CSET-AIID-harm-taxonomy>.
- ¹⁸ China Law Translate, “Provisions on the Management of Algorithmic Recommendations in Internet Information Services,” *China Law Translate* (blog), January 4, 2022, <https://www.chinalawtranslate.com/en/algorithms/>; China Law Translate, “Provisions on the Administration of Deep Synthesis Internet Information Services,” *China Law Translate* (blog), December 12, 2022, <https://www.chinalawtranslate.com/deep-synthesis/>.
- ¹⁹ China Law Translate, “Interim Measures for the Management of Generative Artificial Intelligence Services,” *China Law Translate* (blog), July 13, 2023, <https://www.chinalawtranslate.com/en/generative-ai-interim/>.

²⁰ European Commission, “Proposal for Laying Down Harmonised Rules on Artificial Intelligence,” 2021, <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206>.

²¹ Senador Rodrigo Pacheco (PSD/MG), “Projeto de Lei N° 2338, de 2023” (2023), https://legis.senado.leg.br/sdleg-getter/documento?dm=9347593&ts=1683152235237&disposition=inline&_gl=1*_edqnm*_ga*MTgyMDY0MTcwMS4xNjc5OTM2MTI0*_ga_CW3ZH25XMK*MTY4MzIxNzUzMy4yLjEuMTY4MzlyMDAyMy4wLjAuMA.

²² Minister of Innovation, Science and Industry—Canada, “Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to Other Acts” (2022), <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading>.

²³ Exec. Order No. 14110, “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” 2023–24283 § 88 FR 75191 (November 1, 2023), <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.

²⁴ NAIAC, “RECOMMENDATION: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting” (National Artificial Intelligence Advisory Committee, November 2023), https://ai.gov/wp-content/uploads/2023/12/Recommendation_Improve-Monitoring-of-Emerging-Risks-from-AI-through-Adverse-Event-Reporting.pdf.

²⁵ National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework: AI RMF (1.0),” Gaithersburg, MD: National Institute of Standards and Technology (January 2023), <https://doi.org/10.6028/NIST.AI.100-1>.

²⁶ The White House, “FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI,” The White House, July 21, 2023, <https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>.

²⁷ Institute of Medicine (US) Committee on Quality of Health Care in America; Kohn LT, Corrigan JM, Donaldson MS, editors. *To Err is Human: Building a Safer Health System*. Washington (DC): National

Academies Press (US); 2000. <https://www.ncbi.nlm.nih.gov/books/NBK225187/> 2, Errors in Health Care: A Leading Cause of Death and Injury.

²⁸ “HAI and Antibiotic Use Prevalence Survey,” Centers for Disease Control and Prevention, March 31, 2022, <https://www.cdc.gov/hai/eip/antibiotic-use.html>; Craig Umscheid et al., “Estimating the Proportion of Healthcare-Associated Infections That Are Reasonably Preventable and the Related Mortality and Costs,” *Infection Control and Hospital Epidemiology* 32, no. 2 (February 2011), <https://doi.org/10.1086/657912>.

²⁹ “Serious Reportable Events aka ‘Never Events,’” National Quality Forum (NQF), accessed August 15, 2023, https://www.qualityforum.org/Topics/SREs/Serious_Reportable_Events.aspx.

³⁰ “Home | Patient Safety Organization (PSO) Program,” Agency for Healthcare Research and Quality (AHRQ), accessed August 23, 2023, <https://pso.ahrq.gov/>.

³¹ “Sentinel Event Policy and Procedures | The Joint Commission,” The Joint Commission, accessed August 16, 2023, <https://www.jointcommission.org/resources/sentinel-event/sentinel-event-policy-and-procedures/>.

³² “State-Based Reporting in Healthcare,” National Quality Forum (NQF), accessed July 23, 2023, https://www.qualityforum.org/Projects/State_Based_Reporting/State-Based_Reporting_in_Healthcare.aspx.

³³ National Quality Forum (NQF), “Variability of State Reporting of Adverse Events,” October 2011. https://www.qualityforum.org/Topics/SREs/State_Variability_Fact_Sheet.aspx

³⁴ Agency for Healthcare Research and Quality, “Network of Patient Safety Databases Chartbook, 2021” (Rockville, MD: AHRQ, August 2021). <https://www.ahrq.gov/sites/default/files/wysiwyg/npsd/data/npsd-chartbook-2021.pdf>

³⁵ “Listed PSOs,” Agency for Healthcare Research and Quality (AHRQ), accessed August 16, 2023, <https://pso.ahrq.gov/pso/listed>; AHA Hospital Statistics, “Fast Facts on U.S. Hospitals, 2023” (AHA Hospital Statistics), accessed September 10, 2023, <https://www.aha.org/system/files/media/file/2023/05/Fast-Facts-on-US-Hospitals-2023.pdf>.

³⁶ Agency for Healthcare Research and Quality, “Network of Patient Safety Databases Chartbook, 2021,” 1.

³⁷ Ellen Flink et al., “Lessons Learned from the Evolution of Mandatory Adverse Event Reporting Systems,” in *Advances in Patient Safety: From Research to Implementation (Volume 3: Implementation*

Issues), ed. Kerm Henriksen et al., *Advances in Patient Safety* (Rockville, MD): Agency for Healthcare Research and Quality (US), 2005), <http://www.ncbi.nlm.nih.gov/books/NBK20547/>.

³⁸ Sentinel Event Policy and Procedures. The Joint Commission.

<https://www.jointcommission.org/resources/sentinel-event/sentinel-event-policy-and-procedures/>.

³⁹ Flink et al., “Lessons Learned from the Evolution of Mandatory Adverse Event Reporting Systems.”

⁴⁰ The National Transportation Safety Board, accessed July 14, 2023.

<https://www.nts.gov/Pages/home.aspx>.

⁴¹ “2021–2023 Most Wanted List,” National Transportation Safety Board (NTSB), accessed February 24, 2024, <https://www.nts.gov/Advocacy/Pages/ArchiveMWL.aspx>.

⁴² “Home | NHTSA,” Text, National Highway Traffic Safety Administration, accessed October 21, 2023, <https://www.nhtsa.gov/>.

⁴³ NHTSA, “Standing General Order 2021-01 | Incident Reporting for Automated Driving Systems (ADS) and Level 2 Advanced Driver Assistance Systems (ADAS),” Text, National Highway Traffic Safety Administration, 2021, <https://www.nhtsa.gov/document/sgo-crash-reporting-adas-ads>.

⁴⁴ Faiz Siddiqui, Rachel Lerman, and Jeremy B. Merrill, “Teslas running Autopilot involved in 273 crashes reported since last year,” *The Washington Post*, June 15, 2022, <https://www.washingtonpost.com/technology/2022/06/15/tesla-autopilot-crashes/>.

⁴⁵ NHTSA, “Summary Report: Standing General Order on Crash Reporting for Level 2 Advanced Driver Assistance Systems,” National Highway Traffic Safety Administration (June 2022), <https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/ADAS-L2-SGO-Report-June-2022.pdf>; NHTSA, “Summary Report: Standing General Order on Crash Reporting for Automated Driving Systems,” National Highway Traffic Safety Administration (June 2022), <https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/ADS-SGO-Report-June-2022.pdf>.

⁴⁶ “Report a Safety Problem | NHTSA,” NHTSA, accessed September 15, 2023, <https://www.nhtsa.gov/report-a-safety-problem>.

⁴⁷ Clare M. Patterson, Jason R. C. Nurse, and Virginia N. L. Franqueira, “Learning from cyber security incidents: A systematic review and future research agenda,” *Computers & Security* 132 (September 1, 2023): 103309, <https://doi.org/10.1016/j.cose.2023.103309>.

- ⁴⁸ CVE, “Related Efforts | CVE,” CVE, accessed November 29, 2023, <https://www.cve.org/About/RelatedEfforts>.
- ⁴⁹ CVE, “History | CVE,” CVE, accessed November 29, 2023, <https://www.cve.org/About/History>.
- ⁵⁰ Chris Johnson et al., “Guide to Cyber Threat Information Sharing,” National Institute of Standards and Technology Special Publication (October 2016), <https://doi.org/10.6028/NIST.SP.800-150>.
- ⁵¹ Rebekah Brown and Robert M Lee, “The Evolution of Cyber Threat Intelligence (CTI): 2019 SANS CTI Survey,” 2019, https://a51.nl/sites/default/files/pdf/Survey_CTI-2019_IntSights.pdf
- ⁵² Konstantinos Rantos et al., “Interoperability Challenges in the Cybersecurity Information Sharing Ecosystem,” *Computers* 9, no. 1 (March 2020): 18, <https://doi.org/10.3390/computers9010018>.
- ⁵³ “Presidential Policy Directive -- United States Cyber Incident Coordination,” Obama White House, July 26, 2016, <https://obamawhitehouse.archives.gov/the-press-office/2016/07/26/presidential-policy-directive-united-states-cyber-incident>.
- ⁵⁴ Ben Buchanan et al., “Automating Cyber Attacks,” Center for Security and Emerging Technology (November 2020), <https://doi.org/10.51593/2020CA002>.
- ⁵⁵ CISA, “Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA) Fact Sheet,” Cybersecurity and Infrastructure Security Agency (CISA), (2023), https://www.cisa.gov/sites/default/files/2023-01/CIRCIA_07.21.2022_Factsheet_FINAL_508%20c.pdf.
- ⁵⁶ “Federal Incident Notification Guidelines,” Cybersecurity & Infrastructure Security Agency (CISA), accessed July 6, 2023, <https://www.cisa.gov/federal-incident-notification-guidelines>.
- ⁵⁷ “Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure,” U.S. Securities and Exchange Commission, November 14, 2023, <https://www.sec.gov/corpfin/secg-cybersecurity>.
- ⁵⁸ Federal Register, “Computer-Security Incident Notification Requirements for Banking Organizations and Their Bank Service Providers,” 12 CFR 53 12 CFR 225 12 CFR 304 § (November 23, 2021), <https://www.federalregister.gov/documents/2021/11/23/2021-25510/computer-security-incident-notification-requirements-for-banking-organizations-and-their-bank>.
- ⁵⁹ “Cyber Incident Notification Requirements,” National Credit Union Administration (NCUA), August 14, 2023, <https://ncua.gov/regulation-supervision/letters-credit-unions-other-guidance/cyber-incident-notification-requirements>.

⁶⁰ Nestor Maslej et al., “Artificial Intelligence Index Report 2023,” AI Index Steering Committee (Stanford, CA: Institute for Human-Centered AI, Stanford University, April 2023), https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

⁶¹ A. Hutchinson et al., “Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System,” *BMJ Quality & Safety* 18, no. 1 (February 1, 2009): 5–10, <https://doi.org/10.1136/qshc.2007.022400>; Elena Ramírez et al., “Effectiveness and limitations of an incident-reporting system analyzed by local clinical safety leaders in a tertiary hospital,” *Medicine* 97, no. 38 (September 21, 2018): e12509, <https://doi.org/10.1097/MD.00000000000012509>.

⁶² Ortwin Renn, “Public responses to the Chernobyl accident,” *Journal of Environmental Psychology* 10, no. 2 (June 1990): 151–67, [https://doi.org/10.1016/S0272-4944\(05\)80125-2](https://doi.org/10.1016/S0272-4944(05)80125-2).

⁶³ “Die Beschlüsse des Bundestages am 30. Juni und 1. Juli,” Deutscher Bundestag, June 2011, https://www.bundestag.de/webarchiv/textarchiv/2011/34915890_kw26_angenommen_abgelehnt-205788.

⁶⁴ Catherine Clifford, “Germany has shut down its last three nuclear power plants, and some climate scientists are aghast,” *CNBC*, April 18, 2023, <https://www.cnbc.com/2023/04/18/germany-shuts-down-last-nuclear-power-plants-some-scientists-aghast.html>.

⁶⁵ Alec Tyson and Emma Kikuchi, “Growing public concern about the role of artificial intelligence in daily life,” *Pew Research Center* (blog), August 28, 2023, <https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/>.