Issue Brief

AI Incidents

Key Components for a Mandatory Reporting Regime

Authors
Ren Bin Lee Dixon
Heather Frase

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

January 2025

Executive Summary

In our past publication, "An Argument for Hybrid AI Incident Reporting," we proposed implementing a federated* and comprehensive artificial intelligence incident reporting framework to systematically record, analyze, and respond to AI incidents.[1] The hybrid reporting framework proposes implementing mandatory, voluntary, and citizen reporting mechanisms. This document describes the critical content that should be included in a mandatory AI incident reporting regime and should also inform voluntary and citizen reporting efforts.

In this publication, we define a set of standardized key components of AI incidents that can be used as a reporting template to collect vital AI incident data. These components include, but are not limited to, information about the type of AI incident, the nature and severity of harm, technical data, affected entities and individuals, and the context and circumstances within which the incident unfolded. While intentionally high level, our proposed set of components distills information from existing AI initiatives that track real-world events, harms, and risks related to AI, and incorporates lessons learned from incident reporting systems and practices in the transportation, healthcare, and cybersecurity sectors.

If adopted and used widely and consistently by governments, regulators, professional organizations, developers, and researchers, these reporting components can help enhance AI safety and security measures by:

- Facilitating consistent data collection of AI incidents
- Promoting tracking, monitoring, research, and information sharing of AI incidents
- Enhancing knowledge around AI-related harms and risks
- Ensuring that essential AI incident data is collected to prevent reporting gaps
- Building a foundational framework for agile incident reporting that adapts to AI advancements

To fully utilize the benefits of this list of components, we recommend publishing mandatory AI incident reporting formats based on them and establishing an independent investigative agency to uncover incident data that may not be immediately discernible at the time of reporting. This list can also serve as a template for desirable disclosure guidelines of incident data for voluntary and citizen AI incident reporting systems.

* 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 nongovernmental organizations.

Table of Contents

Introduction

Artificial intelligence incidents have been occurring with wide-ranging adverse impacts. Currently, only a few independent databases document AI incidents.[2] These databases primarily rely upon news reports and other publicly available incident reports from "popular, trade, and academic press."[3] There is not yet any federated* policy framework established to facilitate systematic and comprehensive AI incident reporting practices. As AI continues to expand its capabilities, there is an urgent need to systemically collect AI incident data to enhance our knowledge of AI-related harm and help us develop safe, secure, and trustworthy AI systems.[4]

To address this critical gap, in our previous paper, "An Argument for Hybrid AI Incident Reporting," we proposed establishing policies for a federated, comprehensive, and standardized AI incident reporting framework.[5] We found that implementing consistent and comprehensive reporting and documentation of incidents can help capture more complete and useful data. As more data is gathered, it can reveal vital trends in AI incidents and provide greater clarity on their severity. Over time, this can allow us to more accurately assess the impact and effectiveness of different AI safety and security policies and measures.

The federated AI incident reporting framework we outlined in our previous paper calls for a combination of mandatory, voluntary, and citizen reporting to an independent external entity (such as a government agency, professional association, or oversight body) to promote transparency and accountability in AI incident management. Government agencies, regulators, professional associations, and oversight bodies that will be implementing such a framework to record, document, and monitor AI incidents would benefit from having a consistent set of components that can convey vital information about the AI incident at hand. In this paper, we present precisely such a list that should be used when collecting incident data under a mandatory incident reporting regime, and that may also be used to inform voluntary and citizen AI incident reporting regimes. This list draws on and synthesizes high-level components from various initiatives focused on AI incidents, harms, and risks, and also includes best practices in incident reporting from the transportation, healthcare, and cybersecurity sectors.

Establishing a fundamental list of key components of AI incidents can provide numerous benefits for policymakers, government agencies, oversight committees, civic

* 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 nongovernmental organizations.

organizations, developers, and researchers in recording, documenting, and monitoring AI incident data, such as:

- Facilitating consistent data collection of AI incidents
- Promoting tracking, monitoring, research, and information sharing of AI incidents
- Enhancing knowledge around AI-related harms and risks
- Ensuring that essential AI incident data is collected to prevent reporting gaps
- Building a foundational framework for agile incident reporting that adapts to AI advancements

The outcomes from these actions can aid in enhancing AI safety and security measures. To achieve these benefits, stakeholders can use this standardized list of key components as a template for AI incident reporting. For instance, regulators can design and tailor AI incident reporting systems, templates, and formats to suit the specific needs of their respective domains based on the list of key components. Additionally, the list can serve as the minimum and desirable disclosure requirements of incident data in mandatory and voluntary AI incident reporting systems, respectively.

We begin with an overview of the synthesized list of key components of AI incidents, continue with a discussion of the selection of the components, and conclude with policy recommendations.

Definitions: These definitions were adopted from *Defining AI Incidents and Related Terms*, published by the Organisation for Economic Co-operation and Development (OECD).*

AI incident: An event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems directly or indirectly leads to any of the following harms:

a) injury or harm to the health of a person or groups of people;
b) disruption of the management and operation of critical infrastructure;
c) violations of human rights or a breach of obligations under the applicable law intended to protect fundamental, labor, and intellectual property rights; or
d) harm to property, communities, or the environment.[6]

AI near miss: An event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems could have directly or indirectly led to any of the following harms, but failed to by chance or was intercepted:

a) injury or harm to the health of a person or groups of people;
b) disruption of the management and operation of critical infrastructure;
c) violations of human rights or a breach of obligations under the applicable law intended to protect fundamental, labor, and intellectual property rights; or
d) harm to property, communities, or the environment.[7]

AI hazard: An event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems could plausibly lead to any of the following harms:

a) injury or harm to the health of a person or groups of people;
b) disruption of the management and operation of critical infrastructure;
c) violations to human rights or a breach of obligations under the applicable law intended to protect fundamental, labor, and intellectual property rights; or
d) harm to property, communities, or the environment.[8]

For the sake of brevity, in this paper we will use "AI incident" as an overarching term encompassing both incidents and near misses, except where a distinction is specifically noted. This terminology aligns with the paper's focus on a mandatory reporting framework, which does not extensively discuss concepts such as AI risks, hazards, and issues, as we do not consider them pertinent to such a regime (see Appendix).

Key Components of AI Incidents

To create a list of key components of AI incidents, we analyzed and synthesized a diverse array of high-level components from multiple initiatives discussing AI incidents and related concepts (see Table A1). The process entailed examining existing AI incident databases to evaluate their emphasis on different components, as well as the rationale behind their varied approaches to collection and classification methods. Additionally, we analyzed incident reporting systems from other high-impact sectors, including transportation, healthcare, and cybersecurity, to identify commonalities and gaps that could further bolster our selection of key components of AI incidents (see Table A2).

The key components included in our list are meant to convey the most relevant, meaningful, and useful information about AI incidents. Pragmatic, adaptable, and efficient (rather than sophisticated) taxonomies are more likely to encourage and enhance incident data collection.[9] Therefore, we used the following criteria to guide our analysis and selection, ensuring that the key components of AI incidents are:

- **Easy** to use, understand, and implement across a wide range of sectors and applications by diverse stakeholders, including policymakers, government agencies, and oversight committees
- **Adaptable** to the emerging capabilities of AI and diverse societal factors
- **Functional** for providing data for policymakers to craft safety and security measures and for researchers to analyze AI harms

Table 1 provides an overview of the key components, along with elements, we identified. These components indicate the AI incident information that should be reported and documented. Certain incident data may not necessarily be immediately apparent at the time of reporting and may require further investigation and assessment.

* In this paper, we differentiate between AI near miss and AI hazard, which the OECD definition groups together. For more details, see OECD, *Defining AI Incidents and Related Terms* (Paris: OECD, 2024), 13, https://doi.org/10.1787/d1a8d965-en.

Table 1. Key Components of AI Incidents

Key Components	Elements
Type of event *Did harm occur or nearly occur?*	Incident
	Near miss
Type of harm *What types of harm?*	Physical
	Environmental
	Economic and financial
	Reputational
	Public interest
	Human rights and fundamental rights
	Psychological
Mechanism of harm *What were the contributing factors?*	Technical factors
	Other factors
Severity factors How impactful was the incident?	Remediability
	Level of severity
	Distribution of harm
	Exposed population size
	Duration
	Optionality
	Frequency
Technical information What were the technical dimensions of the implicated AI systems?	AI system card
	AI model card
	Datasheet

Context and circumstances	Goals and application purpose
	Sector
	Start and end date
	Location
	Reporter
	Existing safeguards and policies
Entities and individuals Who was involved and affected?	AI actors
	Affected stakeholders
Incident response	Mitigation, termination, etc.
Ethical impact	Ethical impact assessment using the United Nations Educational, Scientific and Cultural Organization (UNESCO) ethical impact assessment tool or the OECD "AI Principles"[10]

The elements outlined are not exhaustive and can be refined to meet domain-specific needs. Regulators can define the desired granularity of incident data collection through various response formats, such as multiple-choice options, open-text fields, and rating scales. For example, technical factors under "mechanism of harm" could present a list of possible factors, such as bias in training data, data quality issues, and model drift. The "type of harm" could be documented through a combination of multiple-choice categories and a corresponding textual description.

While these key components aim to document AI incidents within a mandatory reporting regime, they can also serve as a nominal template for voluntary and citizen reporting. Both mandatory and voluntary systems, which are directed at AI actors and oversight groups, should include the complete set of key components (with optional inputs for voluntary reporting). Citizen reporting formats may adopt an abbreviated version, focusing on essential components, such as incident description, AI system, and circumstantial details.

Synthesizing Key Components of AI Incidents

From our examination of the AI initiatives and of reporting systems from the transportation, healthcare, and cybersecurity sectors, we distilled and synthesized key components of AI incidents that presented several overarching themes: types of events, harm dimensions, technical data, context and circumstances, and post-incident data.

Types of Events

AI Incidents and Near Misses

The first step in documenting AI incidents under any reporting regime is to determine whether the event meets the criteria and definitions for incidents subject to reporting under that regime. In the context of a mandatory reporting regime, we propose documenting two types of events: incidents and near misses (see Table 2).

Table 2. Key Component: Type of Event

Key Component	Elements	Description
Type of event Did harm occur or nearly occur?	Incident	Harm occurred.
	Near miss	Harm nearly occurred but either was avoided by chance or was intercepted.

We adopt the OECD definition of an AI incident, which is "an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to" harm.[11] For instance, a deepfake image of an explosion at a U.S. federal government building that briefly causes the stock market to dip would be classified as an AI incident.[12] Near misses are similar to AI incidents, except harm either was avoided by chance or was intercepted.[13] A near miss example would be an autonomous vehicle failing to stop at a stop sign, but fortunately no other cars were present at the intersection, avoiding a potential accident.

Even though the OECD integrates "near miss" under its definition of AI hazards, we believe it is crucial to distinguish between AI hazards and near misses. Near misses are characterized as events that could have led to a harmful outcome but did not, either due to chance or because they were averted.[14] In contrast, hazards are described as

unsafe conditions or circumstances that have the potential to cause harm but have not yet resulted in an actual event, often referred to as "accidents waiting to happen." This distinction is vital in defining the scope of AI incident reporting policies and enhancing AI safety and security protocols related to hazards.

AI incidents and near misses should be included within the scope of mandatory reporting. Reporting near misses can enhance incident data collection, as these events exhibit similar characteristics to incidents, apart from their outcomes. In addition to aiding early detection of novel AI risks, tracking near misses could reveal vital conditions that prevented harm from occurring, which can be leveraged to strengthen safety measures.

In contrast, AI hazards—which are unsafe conditions—do not generally warrant mandatory reporting or adhere to the same incident reporting policies and disclosure requirements. For example, voluntary reporting may be more suitable for AI hazards, allowing developers and policymakers to use the information to implement more effective guardrails, thereby reducing the likelihood of harm materializing.

Harm Dimensions

The harm dimensions of an AI incident consist of several key components: types of harm, mechanisms of harm, and severity factors.

Type of Harm

Among the initiatives examined, the OECD offers the most functional and comprehensive categorization of AI-related harms, including physical, environmental, economic or financial, reputational, public interest, human and fundamental rights, and psychological.[15] The categories effectively cover the diverse descriptions of harms found across the initiatives. Documenting the types of harm using these categories can aid researchers and policymakers in conducting systematic analyses of the impact caused by AI incidents. Often, an incident could involve more than one type of harm. For instance, a facial recognition technology that disproportionately misidentifies women, Black, Latino, and Asian shoppers as "likely" shoplifters can result in harm to fundamental rights, psychological harm, and economic losses.[16]

Mechanism of Harm

Reporting the mechanism of harm aims to capture all plausible contributing factors of AI incidents and near misses. The mechanisms of harm that we observed from existing incidents and initiatives can be grouped into two categories: technical factors and other

factors. Technical factors relate to system vulnerabilities, model drift, and system failures.[17] Other factors can include human and contextual factors, such as weak governance, user misuse, and intentional abuse by user. While technical and nontechnical factors can contribute to AI harm simultaneously, we suggest documenting these two factors separately to distinguish between intentional and unintentional harms, as well as to enhance our understanding of human-AI interaction.[18] The data gathered from this component will be crucial for conducting root cause analysis.

Severity Factor

Collecting data on severity factors will be essential for severity assessments that can aid in prioritizing mitigation efforts, formulating appropriate safety and security policies, and enhancing risk assessments.[19] Although extreme harm, such as death or irreversible disruption of critical infrastructure, clearly constitutes significant impact, the severity of harm is often nuanced and highly context dependent.[20] The perceived severity can vary based on social, legal, and temporal context, leading to differing assessments of harm across individuals, organizations, and governments.[21] Developing an AI incident severity rating or metric framework will require further exploration and will not be discussed in this paper.

Table 3 provides an overview of the key components related to the harm dimensions of an AI incident, along with their elements and descriptions.

Table 3. Key Components: Harm Dimensions

Key Components	Elements	Description
Type of harm *What types of harm?*	Physical	Physical injury and death
	Environmental	E.g., soil contamination, water pollution, and air pollution
	Economic and financial	E.g., financial loss or damages, harm to property
	Reputational	May affect organizations and individuals
	Public interest	Involves critical infrastructure and its function and the social fabric of

		society
	Human rights and fundamental rights	Includes established domestic and international laws
	Psychological	E.g., affects mental health
Mechanism of harm *What were the contributing factors?*	Technical factors	Malfunctions, system vulnerabilities, data poisoning, and concept drift
	Other factors	Weak governance, lack of safeguards, user misuse, and user abuse
Severity factors *How impactful was the incident?*	Remediability	The ability to restore those affected to a situation at least equivalent to their situation before the impact[22]
	Level of severity	How acutely the harm impacted those affected
	Distribution of harm	Whether an individual, group, or population was disproportionately affected by the harm
	Exposed population size	A full estimate of the adversely impacted stakeholders
	Duration	How long the harm was experienced by the affected stakeholders
	Optionality	Users' or stakeholders' ability to accept, challenge, correct, or opt out of the system's output
	Frequency	The rate at which affected stakeholders experience harm

Technical Data

Next, we consolidated all technical data under one key component and propose requiring AI actors to submit AI system or model cards and datasheets as part of mandatory reporting obligations (see Table 4).[23]

Table 4. Key Component: Technical Information

Key Component	Elements	Description
Technical information *What were the technical dimensions of the implicated AI systems?*	AI system card	Data, models, code, system
	AI model card	Model details, intended use, evaluation data, and training data
	Datasheet	Training data information on motivation, composition, collection process, recommended use, and more

At the moment, it is challenging to fully capture the vital technical dimensions of AI harm events because such data is neither publicly available nor readily discernible from news reports. If standardized AI system and AI model cards are required in a mandatory reporting framework, researchers and policymakers will have access to vital technical data, which can improve technical standards and risk-mitigation practices throughout the AI life cycle. For instance, the National Telecommunications and Information Administration recommends federal agencies to work with stakeholders to improve standard information disclosures within such artifacts.[24] Drawing from the concept of nutrition labels mandated by the Food and Drug Administration, these artifacts should provide essential details about AI systems, models, and training data.

Context, Circumstances, and Stakeholders

Table 5 lists key components related to the contextual and circumstantial data, along with information about the relevant AI actors and affected stakeholders. These components are designed to capture the broader situational and preexisting conditions surrounding an AI incident. It includes information such as the purpose of the AI model or system, the sector in which the AI was deployed, the start and end date of the event (if applicable), locations, the reporting entity or individual, and any existing safeguards

and policies. The identity of the reporter may or may not be documented depending on the reporting requirements; in certain schemata anonymity is provided. Documenting existing safeguards and policies enables assessments of safety and security measures and identifies gaps for improvement. In addition to gathering background details about the relevant AI actors and affected stakeholders, the data should indicate whether the affected stakeholders were users or nonusers of the implicated AI system(s) to clarify their relationship with the system(s).

Table 5. Key Components: Context and Circumstances

Key Components	Elements	Description
Context and circumstances	Goals and application purpose	E.g., autonomous driving, video generation, face recognition
	Sector	E.g., healthcare, financial, agriculture
	Start and end date (if applicable)	Date of first and last known incidents
	Location	City, state, country
	Reporter (if relevant)	Name and information of person reporting
	Existing safeguards and policies	Safeguards or policies that were already in place before the incident occurred, if applicable
Entities and individuals *Who was involved and affected?*	AI actors	Details of AI provider, operator, and deployer
	Affected stakeholders	Details of affected individuals and entities, if available, including whether they were users or nonusers of the implicated AI system(s)

Post-incident Data

Certain data, such as the incident response and ethical impact, is only available in the aftermath of an AI incident.

Recording the actions taken after an incident occurs provides information on organizational incident response plans, which offers multiple benefits (see Table 6). This practice promotes transparency and accountability among AI actors while facilitating the evaluation of incident response strategies. Additionally, documenting incident responses contributes to the development of best practices for addressing AI-related harms. The National Institute of Standards and Technology (NIST) further recommends tracking incident response time as a safety metric in AI risk management.[25]

Impact assessments have emerged as a relevant policy instrument for promoting safe, secure, and trustworthy AI development and deployment. Numerous regulatory and standard-setting frameworks, including the European Union's AI Act, the Council of Europe's Framework Convention on AI, and the NIST *AI Risk Management Framework*, have either mandated or recommended the use of impact assessments in their guidance.[26] Relevant authorities can adopt these impact assessment schemes, such as the UNESCO ethical impact assessment tool, to assess the ethical impact of AI-related harm.[27]

Table 6. Key Components: Post-Incident

Key Components	Elements	Description
Incident response	Mitigation, termination, etc.	Description of incident response
Ethical impact	Ethical impact assessment using the UNESCO ethical impact assessment tool or the OECD "AI Principles"	Assessment of the ethical ramifications and broader impacts of an event beyond its immediate outcome

Policy Recommendations

The purpose of this paper is to provide for policymakers, government agencies, and oversight committees a consistent set of key components of AI incidents that can be used as a template for reporting, collecting, and documenting AI harm events under a mandatory reporting regime. To fully utilize the benefits of this list, we recommend publishing AI incident reporting formats based on the key components and establishing an independent investigative agency to uncover incident data that may not be immediately discernible at the time of reporting.

Publish AI Incident Reporting Formats

Policymakers, regulators, and oversight committees are encouraged to adopt and tailor the list of key components for AI incident reporting to meet the specific requirements of their respective domains. Standardizing these components is essential for ensuring consistency in incident data collection, which allows for meaningful data aggregation and comparison. Our prior research has highlighted the difficulties posed by inconsistent data collection, which can render the data unusable and lead to inefficiencies.[28] By standardizing and optimizing data collection, we can derive critical insights that will inform the development and deployment of safe, secure, and trustworthy AI systems.

Establish an Independent Investigation Agency

In line with our previous paper, "An Argument for Hybrid AI Incident Reporting," we recommend establishing an independent AI incident investigation agency (similar to the role of the National Transportation Safety Board) to examine significant AI incidents further to uncover information that may not be evident during the time of reporting.[29]

Conducting investigations and root cause analysis is common practice in incident responses across the transportation, healthcare, and cybersecurity sectors. In the context of AI, uncovering circumstantial information and the conditions of human-machine teaming—specifically the user's interaction, trust, reliance, and dependence on an AI system—can reveal how these factors affect incident outcomes.[30] A dedicated independent investigation agency can pinpoint existing safeguards and policies, if any, that were in place but failed to mitigate the incident. Investigations can also obtain more accurate data on exposed population size and other probability factors in AI incidents. This information can indicate the probability of additional occurrences that were previously undetected, unreported, or could potentially happen in the future.

Conclusion

This set of key components should undergo continuous iteration, with regular reviews and updates, to maintain its effectiveness in capturing the dimensions of AI incidents. The database of key components can be published online, allowing the public to submit suggestions or alert researchers and policymakers to novel harms and risks. These key components can also serve as the foundation for developing metrics for measuring incident severity. Eventually, when regulators have more incident data, they can refine the scope of AI incidents that should be included within a mandatory reporting regime or that require further investigation.

Incorporating a standardized and fundamental set of key components is integral to successfully implementing a federated, hybrid AI incident reporting framework. When we better understand how and why AI incidents happen, we can reduce AI-related harm and prevent their reoccurrence.

Appendix

Decomposing AI Incidents

The majority of the key components of AI incidents were distilled and synthesized from the twelve databases and documents we examined. These initiatives either directly gather AI incident data or provide a framework for analyzing AI-related harms and risks (see Table A1).

Table A1. List of Examined AI Initiatives

Title	Type	Objective	Target User
AI Incident Database (AIID)[31]	Database	Identify, define, and catalog AI incidents in a database to support research within this field	Public, researchers, developers
AI, Algorithmic, and Automation Incidents and Controversies Repository (AIAAIC)[32]	Database	Achieve algorithmic transparency and openness by systematically collecting, classifying, and revealing AI-related incidents and issues	Researchers, educators, policymakers, citizens, consumers
AI Vulnerability Database (AVID)[33]	Database	Store instantiations of AI risks related to AI failure mode	Auditors, developers, regulators, policymakers
OECD AI Incidents Monitor (AIM)[34]	Database	Document AI incidents to help policymakers, AI practitioners, and all stakeholders worldwide gain valuable insights into the incidents and hazards that concretize AI risks	Policymakers, AI practitioners, public
OECD Framework for the Classification of AI Systems[35]	Document	Assess the opportunities and risks that different types of AI systems present and inform national AI strategies	Policymakers, regulators, legislators

Defining AI Incidents and Related Terms[36]	Document	Provide a definition of AI incident and related terminology	Public, policymakers
"A Taxonomic System for Failure Cause Analysis of Open Source AI Incidents"[37]	Document	Analyze and annotate AI incidents via the development of a taxonomic system that captures goals, methods and technologies, and failure causes of a technical nature	Researchers
"Classifying AI Systems"[38]	Document	Classify AI systems uniformly and use those classifications to inform consequential decisions about AI technologies, while effectively monitoring risk and bias and managing system inventories	System developers, governing bodies, users
"Adding Structure to AI Harm"[39]	Document	Introduce the CSET AI Harm Framework, a standardized conceptual framework to support and facilitate analyses of AI harm	Researchers, policymakers
Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence[40]	Document	Direct government agencies to advance AI governance and innovation while managing risks from AI, particularly those affecting public rights and safety	Government agencies and departments
"A Framework for Identifying Highly Consequential AI Use Cases"[41]	Document	Determine which AI uses and outcomes to focus regulatory efforts on	Regulators
"Governing General Purpose AI: A Comprehensive Map of	Document	Provide a structured resource for policymakers seeking to understand the multifaceted challenges of general-purpose AI	Policymakers

Unreliability, Misuse and Systemic Risks"[42]		and the potentially far-reaching impact of governing it effectively	
"Evaluating the Social Impact of Generative AI Systems in Systems and Society"[43]	Document	Provide a standard approach in evaluating the impacts of generative AI systems	Researchers
Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile[44]	Document	Provide a cross-sectoral profile of and companion resource for the *AI Risk Management Framework (AI RMF 1.0)* for Generative AI.	Organizations, deployers, operators, users

The AI initiatives and the reporting systems we examined provided a range of information and approaches for analyzing incidents. While the AI initiatives set a valuable foundation for examining AI-related harms and risks, they were not developed specifically for a federated and comprehensive incident reporting framework. The objectives and target audience of these initiatives influenced their methods and frameworks in documenting AI-related harms and risks, resulting in diverse taxonomies, varying sets of AI incident components, and a lack of clarity on reporting requirements. Notably, none of the initiatives encompassed all the components completely. These variations may hinder comparability across various analytical efforts that are crucial for informing AI safety and security measures.[45]

AI Harm Events as a Spectrum

These AI initiatives document and frame AI harms and incidents across a wide spectrum, ranging from incidents where actual harm occurred, to near-miss events where harm was narrowly avoided, to hazardous circumstances that could lead to harm (such as releasing untested AI systems to the public), to broader systemic risks (such as ideology homogenization), to hypothetical risks (eased access to chemical,

biological, radiological, and nuclear weapons).[46] These incidents and circumstances are generally defined as harm events, near misses, issues, and risks.

In the context of mandatory AI incident reporting, we concentrate solely on incidents that should be reported. This excludes discussions of AI issues, hypothetical risks, and systemic risks, as these would likely require a different management approach due to their nature and temporal aspect. While AI issues and risks are important and should be addressed, they often do not demand the same urgency or resolution as incidents where actual harm occurred or was plausibly imminent. Including issues and risks into incident reporting systems can inundate the system with large volumes of reports, overwhelming its capacity and diminishing its overall effectiveness.

This brings the OECD's definitions of AI incidents and related terms into focus. These definitions describe AI incidents and hazards as involving events and circumstances that result in actual or plausible harm.[47] Incidents and events that lead to, or can plausibly lead to, actual harm are reportable and data should be collected that enables researchers and policymakers to understand their underlying factors. This data collection is essential to developing and improving safety and security measures, fostering accountability, and promoting transparency and public trust.[48]

Documenting AI Harm: The Many Impacts of AI

Given AI's wide-ranging applications, the harms from incidents can be extensively diverse. The examined AI initiatives adopted various approaches to frame, classify, and categorize these harms.

CSET's framework for structuring AI harm employs binary categorizations, grouping them into tangible and intangible harms.[49] The Office of Management and Budget (OMB) guidance shares a binary approach to framing the consequences of AI use, focusing on its applications.[50] The guidance classifies AI systems based on their impact on rights and safety depending on their specific use cases. Meanwhile, other initiatives either outline the types of harm individually (such as data privacy, environmental costs, information security, physical injury, public interest, and misinformation) or adopt classifications that reflect the objectives of their publications and target audience. For instance, technical and security initiatives emphasize system vulnerabilities, organizational management initiatives perceive harm as either internal or external to the organization, and initiatives aimed at policymakers underscore the societal impacts of AI use.[51]

The different frameworks offer various advantages, with some emphasizing flexibility and adaptability, while others feature explanatory harm classification to support analysis and policy development. It is noteworthy that the high-level binary approaches presented by CSET and the OMB offer a flexibility that could be advantageous for documenting emerging AI incidents and adapting to "local laws, societal norms, and communal experiences."[52] However, considering the early stages of implementing an AI incident reporting framework, it may be strategic to incorporate a more descriptive harm classification framework to facilitate meaningful aggregate analyses of the collected incident data by regulators, researchers, and policymakers. Among the various frameworks presented in the AI initiatives, the OECD's list of AI harms offers the most comprehensive descriptions, encompassing the diverse impacts of AI incidents discussed in the initiatives.[53] The classification can, and should, be expanded to address the specific needs of each reporting domain and any novel harms that may arise.

Significant but Limited Information

While the AI initiatives we examined offered extensive discussions on the impact of AI incidents, we found that detailed technical data and the severity levels of the incidents were documented less extensively.

Currently, the AI incident databases (AIID, AIAAIC, AVID, and OECD AIM) largely record high-level information about the AI systems involved in incidents, such as the names of AI systems and deployers. However, granular technical details—such as the properties of the AI system and model, evaluation data, and training data—were less prevalent in the databases. Given that AI incident data is mainly obtained from news reports, richer contextual technical details may not be readily discernible. This is likely because AI developers often do not publish, or users cannot access, this information. Even though granular technical data is not thoroughly documented at present, this information is essential for providing a complete picture of an AI incident and, in turn, aiding in identifying incidents' root causes.

Likewise, documenting severity levels of AI incidents is crucial, as it can help determine the appropriate responses. However, the examined AI incident initiatives often lack a systematic scheme for discerning the severity levels of AI incidents, despite addressing the implications and impacts of AI incidents extensively. The exception is the "Framework for Identifying Highly Consequential AI Use Cases" document, which proposes a framework for assessing risks in AI use cases. The framework delineates severity and likelihood factors that can be combined to assess the risks associated with AI use cases. Even though the framework was not specifically developed for assessing

the severity of AI incidents, the relevant severity factors offer valuable insights into severity-related information that should be included in incident reporting.

Incident Components Reported in Other Sectors

To better understand the key components in AI incidents, we looked beyond the field of AI to analyze a selection of incident reporting frameworks from other high-impact sectors. The AI incident initiatives have also referred to incident reporting from other sectors (e.g., aviation) as a template for developing their frameworks and recommendations.[54] In our previous paper, "An Argument for Hybrid AI Incident Reporting," we conducted a macroanalysis of the incident reporting practices from the transportation, healthcare, and cybersecurity sectors.[55] For this paper, we examined the types of incident data collected in these sectors to identify commonalities, gaps, and considerations to provide further context around the key components of AI incidents (see Table A2).

Table A2. List of Reporting Systems from Other Sectors

Entity	Reporting System
National Transportation Safety Board	Aircraft accidents[56]
National Transportation Safety Board	Transportation accidents[57]
National Quality Forum	"List of Serious Reportable Events"[58]
Agency for Healthcare Research and Quality	Common formats overview[59]
The Joint Commission	"Sentinel Event Policy"[60]
Cybersecurity and Infrastructure Security Agency	Incident Reporting System[61]
MITRE Corporation	Common Vulnerabilities and Exposure (CVE) Program[62]

Even though each reporting system gathers its own set of domain-specific incident data, we were able to observe elements that could be useful in documenting and analyzing AI incidents.

Shared Incident Components

Across the incident reporting systems, we identified common components observed in the AI incident reporting initiatives. They were contextual information about incidents and technical data.

Contextual information about incidents (e.g., timing, duration, and involved entities or artifacts) is collected across the incident reporting systems in the transportation, healthcare, and cybersecurity sectors. While it may seem routine, this data is valuable for revealing the environmental conditions, scope, and extent of an incident's impact. In turn, these insights could reveal additional contributing factors. For example, recording data on the duration of an incident could reveal information about its exposure and likely impact.

Incident reporting systems across sectors collect various **technical data**. Cybersecurity's CVE program, in particular, emphasizes documenting detailed technical data.[63] The rigorous collection of this data facilitates information sharing between information technology and cybersecurity professionals, aiding in prioritizing and mitigating cybersecurity vulnerabilities.[64]

Additional Key Components

The incident reporting systems from these three sectors document two components more rigorously than the AI initiatives do. These components were near misses and existing safeguards and policies.

Near misses are incidents where harm did not occur, but there was imminent potential for harm. Near misses are distinct from harm issues or hazards, where there is only a reasonable probability that harm could occur. In healthcare, near misses are health or safety events where harm does not reach the patient. In transportation, accidents that could have happened but did not are called "close calls."[65] Studying near misses can provide valuable insights into harmful incidents. Reports of near misses often constitute the majority of reported healthcare incidents, while actual harmful incidents are a subset of this population.[66] Thus, by analyzing reports of near misses, researchers can expand efforts to identify contributing factors and reduce reoccurrences.

Identifying **existing safeguards and policies** can highlight policy ineffectiveness and provide insight into risk mitigation. The reporting systems from the transportation and healthcare sectors include documenting lifesaving equipment, policies, and procedures that were present or in effect when an incident occurred. Highlighting ineffective safeguards and policies can assist AI developers and policymakers in enhancing AI safety, trustworthiness, and security measures. Integrating this component into AI incident reporting can benefit AI policy development, particularly as AI governance is an evolving area. This integration facilitates the necessary iteration needed for AI policies to adapt to the dynamic nature of the technology.

Measuring Severity: A First Glimpse

A prevalent component observed throughout the examined reporting systems was the **severity** or impact level of incidents. The severity levels reported in the transportation and healthcare systems serve dual purposes: besides indicating the severity of an incident, they aid in ascertaining the appropriate incident response. Incident response examples include investigation, impact assessment, mitigation, and prioritization. Investigations can be particularly beneficial for uncovering root causes in severe incidents or incidents with ambiguous causal factors.

These reporting systems employ distinct severity levels. In the transportation and healthcare systems, severity levels encompass a span of outcomes ranging from fatalities to mild injuries. In cybersecurity, the CVE Program uses the Common Vulnerability Scoring System (CVSS) Calculator to administer a quantitative measurement of severity on a range of zero to 10.[67] The CVSS Calculator provides industries, organizations, and governments with a consistent framework for assessing and quantifying severity.

Establishing a standardized framework for assessing the severity of AI incidents can be advantageous for the reasons listed in the introduction. Such a framework could also aid in affected stakeholder advocacy, policy analysis, and research on AI harm. Developing an AI incident severity assessment framework will require further exploration and not be discussed in this paper.

Authors

Ren Bin Lee Dixon is an AI policy analyst researching AI policies, governance, and ethics. Heather Frase, PhD, is president of Veraitech and faculty at Virginia Tech's National Security Institute.

Acknowledgments

For their comprehensive and valuable reviews, we would like to thank Sean McGregor, Violet Turri, Borhane Blili-Hamelin, Mia Hoffman, Mina Narayanan, and Catherine Aiken. Finally, we would like to thank Margarita Konaev and Igor Mikolic-Torreira for their feedback and support.

Document Identifier: doi: 10.51593/20240023

Endnotes

[1] Ren Bin Lee Dixon and Heather Frase, "An Argument for Hybrid AI Incident Reporting," CSET, March 2024, https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/.

[2] "Welcome to the AI Incident Database," AIID, accessed 2023, https://incidentdatabase.ai/; "AIAAIC," AI, Algorithmic, and Automation Incidents and Controversies, accessed 2023, www.aiaaic.org/home; "AI Vulnerability Database," AVID, accessed 2023, https://avidml.org/; OECD, "OECD AI Incidents Monitor (AIM)," OECD AI Policy Observatory, 2023, https://oecd.ai/en/incidents.

[3] Sean McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database," Proceedings of the AAAI Conference on Artificial Intelligence 35, no. 17 (May 18, 2021): 15458–15463, https://doi.org/10.1609/aaai.v35i17.17817.

[4] Dixon and Frase, "An Argument for Hybrid AI Incident Reporting"; McGregor, "Preventing Repeated Real World AI Failures."

[5] Dixon and Frase, "An Argument for Hybrid AI Incident Reporting."

[6] OECD, Defining AI Incidents and Related Terms (Paris: OECD, 2024), https://doi.org/10.1787/d1a8d965-en.

[7] OECD, *Defining AI Incidents and Related Terms*; World Alliance for Patient Safety, *WHO Draft Guidelines for Adverse Event Reporting and Learning Systems* (Geneva: World Health Organization, 2005), https://iris.who.int/bitstream/handle/10665/69797/WHO-EIP-SPO-QPS-05.3-eng.pdf.

[8] OECD, *Defining AI Incidents and Related Terms.*

[9] Aziz A. Boxwala et al., "Organization and Representation of Patient Safety Data: Current Status and Issues around Generalizability and Scalability," Journal of the American Medical Informatics Association 11, no. 6 (2004): 468–478, https://doi.org/10.1197/jamia.M1317; Carl Macrae, "The Problem with Incident Reporting," BMJ Quality & Safety 25, no. 2 (February 1, 2016): 71–75, https://doi.org/10.1136/bmjqs-2015-004732.

[10] OECD, "OECD AI Principles Overview," OECD AI Policy Observatory, accessed March 6, 2024, https://oecd.ai/en/principles; UNESCO, Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence (Paris: UNESCO, 2023), https://unesdoc.unesco.org/ark:/48223/pf0000386276.

[11] OECD, *Defining AI Incidents and Related Terms*.

[12] "Incident 543: Deepfake of Explosion Near US Military Administration Building Reportedly Causes Stock Dip," AIID, May 29, 2023, https://incidentdatabase.ai/cite/543/.

[13] World Alliance for Patient Safety, *WHO Draft Guidelines*.

[14] World Alliance for Patient Safety, *WHO Draft Guidelines*, 20.

[15] OECD, *Defining AI Incidents and Related Terms*.

[16] "Incident 619: Rite Aid Facial Recognition Disproportionately Misidentified Minority Shoppers as Shoplifters," AIID, April 5, 2011, https://incidentdatabase.ai/cite/619/.

[17] Helen Toner and Zachary Arnold, "AI Accidents an Emerging Threat," CSET, July 2021, https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/; AVID, "AI Vulnerability Database."

[18] Dewey Murdick, "Building Trust in AI: A New Era of Human-Machine Teaming," CSET, July 19, 2023, https://cset.georgetown.edu/article/building-trust-in-ai-a-new-era-of-human-machine-teaming/.

[19] B-Tech, "Identifying and Assessing Human Rights Risks Related to End-Use" (United Nations Human Rights Office of the High Commissioner, September 2020), www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf; NIST, AI Risk Management Framework: AI RMF (1.0) (Washington, D.C.: Department of Commerce, 2023), https://doi.org/10.6028/NIST.AI.100-1.

[20] European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA Relevance" (2024), EUR-Lex, http://data.europa.eu/eli/reg/2024/1689/oj/eng.

[21] Mia Hoffmann and Heather Frase, "Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework," CSET, July 2023, https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/; Benjamin Gregg, Human Rights as Social Construction (Cambridge: Cambridge University Press, 2011); Morad Elsana, "Legal Pluralism and Indigenous Peoples Rights: Challenges in Litigation and Recognition of Indigenous Peoples Rights," University of Cincinnati Law Review 87, no. 4 (May 23, 2019): 1043.

[22] B-Tech, "Identifying and Assessing Human Rights Risks Related to End-Use."

[23] Mitchell et al., "Model Cards for Model Reporting"; Furkan Gursoy and Ioannis A. Kakadiaris, "System Cards for AI-Based Decision-Making for Public Policy" arXiv preprint arXiv:2203.04754 (August 31, 2022), http://arxiv.org/abs/2203.04754; Timnit Gebru et al., "Datasheets for Datasets" arXiv preprint arXiv:1803.09010 (December 1, 2021), https://doi.org/10.48550/arXiv.1803.09010.

[24] National Telecommunications and Information Administration, Artificial Intelligence Accountability Policy Report (Washington, D.C.: Department of Commerce, 2024), 71, https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf; Mitchell et al., "Model Cards for Model Reporting"; Gursoy and Kakadiaris, "System Cards for AI-Based Decision-Making for Public Policy"; Gebru et al., "Datasheets for Datasets."

[25] NIST, *AI RMF (1.0)*.

[26] European Union, Regulation (EU) 2024/1689; NIST, AI RMF (1.0); Committee on Artificial Intelligence, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law: Explanatory Report (Strasbourg: 133rd Session of the Committee of Ministers, May 17, 2024), https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680afb122.

[27] UNESCO, *Ethical Impact Assessment*.

[28] Dixon and Frase, "An Argument for Hybrid AI Incident Reporting."

[29] Dixon and Frase, "An Argument for Hybrid AI Incident Reporting"; Eric Fielding, Andrew W. Lo, and Jian Helen Yang, "The National Transportation Safety Board: A Model for Systemic Risk Management," SSRN, November 14, 2010, https://doi.org/10.2139/ssrn.1695781.

[30] Murdick, "Building Trust in AI."

[31] AIID, "Welcome to the AI Incident Database"; McGregor, "Preventing Repeated Real World AI Failures."

[32] "AIAAIC."

[33] AVID, "AI Vulnerability Database."

[34] OECD, "OECD AIM."

[35] OECD, Stocktaking for the Development of an AI Incident Definition (Paris: OECD, 2023), https://doi.org/10.1787/c323ac71-en.

[36] OECD, *Defining AI Incidents and Related Terms*.

[37] Nikiforos Pittaras and Sean McGregor, "A Taxonomic System for Failure Cause Analysis of Open Source AI Incidents," arXiv preprint arXiv:2211.07280 (November 14, 2022), http://arxiv.org/abs/2211.07280.

[38] Catherine Aiken, "Classifying AI Systems," CSET, November 2021, https://cset.georgetown.edu/publication/classifying-ai-systems/.

[39] Hoffmann and Frase, "Adding Structure to AI Harm."

[40] Director of OMB, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, M-24-10, March 28, 2024, www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf.

[41] "A Framework for Identifying Highly Consequential AI Use Cases" (Special Competitive Studies Project and John Hopkins University Applied Physics Laboratory, November 7, 2023), www.scsp.ai/wp-content/uploads/2023/11/SCSP_JHU-HCAI-Framework-Nov-6.pdf.

[42] Pegah Maham and Sabrina Küspert, "Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks" (Stiftung Neue Verantwortung, 2023).

[43] Irene Solaiman et al., "Evaluating the Social Impact of Generative AI Systems in Systems and Society" arXiv preprint arXiv:2306.05949 (June 12, 2023), http://arxiv.org/abs/2306.05949.

[44] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (Washington, D.C.: Department of Commerce, 2024), https://doi.org/10.6028/NIST.AI.600-1.

[45] Hoffmann and Frase, "Adding Structure to AI Harm."

[46] "Incident 238: Oregon's Screening Tool for Child Abuse Cases Discontinued Following Concerns of Racial Bias," AIID, October 1, 2018, https://incidentdatabase.ai/cite/238/; Maham and Küspert, "Governing General Purpose AI"; NIST, Generative Artificial Intelligence Profile.

[47] OECD, *Defining AI Incidents and Related Terms*.

[48] Dixon and Frase, "An Argument for Hybrid AI Incident Reporting."

[49] Hoffmann and Frase, "Adding Structure to AI Harm."

[50] Director of OMB, *Advancing Governance, Innovation, and Risk Management.*

[51] "Effect (SEP) View," AVID, accessed 2023, https://docs.avidml.org/taxonomy/effect-sep-view; "Classifications and Definitions," AIAAIC, last modified December 13, 2024, www.aiaaic.org/aiaaic-repository/classifications-and-definitions; Maham and Küspert, "Governing General Purpose AI."

[52] Hoffmann and Frase, "Adding Structure to AI Harm," 7.

[53] OECD, *Defining AI Incidents and Related Terms*, 17.

[54] Pittaras and McGregor, "A Taxonomic System for Failure Cause Analysis of Open Source AI Incidents"; Violet Turri and Rachel Dzombak, "Why We Need to Know More: Exploring the State of AI Incident Documentation Practices," in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (New York: Association for Computing Machinery, 2023), 576–583, https://doi.org/10.1145/3600211.3604700.

[55] Dixon and Frase, "An Argument for Hybrid AI Incident Reporting."

[56] "Report an Aircraft Accident to the NTSB," National Transportation Safety Board, accessed January 4, 2024, www.ntsb.gov/Pages/aviationreport.aspx.

[57] "Responding to a Transportation Accident," NTSB, 2023, https://www.ntsb.gov/Documents/2023%20Transportation%20Accident%20Response%20Guide.pdf.

[58] "List of SREs," National Quality Forum, accessed January 4, 2024, www.qualityforum.org/Topics/SREs/List_of_SREs.aspx.

[59] "Common Formats Overview | PSO," Agency for Healthcare Research and Quality, accessed Nov 2024, https://pso.ahrq.gov/common-formats/about.

[60] The Joint Commission, "Sentinel Event Policy," in Comprehensive Accreditation Manual for Hospitals: Update 2 (Oakbrook Terrace, IL: Joint Commission Resources, 2024), www.jointcommission.org/-/media/tjc/documents/resources/patient-safety-topics/sentinel-event/camh_se_20230906_155314.pdf.

[61] "Incident Reporting System | CISA," Cybersecurity and Infrastructure Security Agency, accessed 2024, https://myservices.cisa.gov/irf?id=irf_report.

[62] "CVE Record Format Overview." HTML. 2020. Reprint, CVE Program, accessed 2025. https://github.com/CVEProject/cve-schema.

[63] CVE Program, "CVE Record Format Overview."

[64] David E. Mann and Steven M Christey, "Towards a Common Enumeration of Vulnerabilities" (MITRE Corporation, January 8, 1999), www.cve.org/Resources/General/Towards-a-Common-Enumeration-of-Vulnerabilities.pdf.

[65] "Precursor Safety Data Program," Bureau of Transportation Statistics, November 20, 2024, www.bts.gov/close-call.

[66] 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.0000000000012509.

[67] "Vulnerability Metrics," National Vulnerability Database, last modified June 27, 2024, https://nvd.nist.gov/vuln-metrics/cvss.