Designing Incident Reporting Systems for Harms from General-Purpose AI

Kevin Wei[1, 2], **Lennart Heim**[1, 2]

[1] RAND
[2] GovAI (Centre for the Governance of AI)
kwei@rand.org

Abstract

We introduce a conceptual framework and provide considerations for the institutional design of AI incident reporting (IR) systems, i.e., processes for collecting information about safety- and rights-related events caused by general-purpose AI. As general-purpose AI systems are increasingly adopted, they are causing more real-world harms and displaying the potential to cause significantly more dangerous incidents—events that did or could have caused harm to individuals, property, or the environment. Through a literature review, we develop a framework for understanding the institutional design of AI IR systems, which includes seven dimensions: policy goal, actors submitting and receiving reports, type of incidents reported, level of risk materialization, enforcement of reporting, anonymity of reporters, and post-reporting actions. We then examine nine case studies of incident reporting in safety-critical industries to extract design considerations for AI IR in the United States. We discuss, among other factors, differences in systems operated by regulatory vs. non-regulatory government agencies, near miss reporting, the roles of mandatory reporting thresholds and voluntary reporting channels, how to enable safety learning after reporting, sharing incident information, and clarifying legal frameworks for reporting. Our aim is to inform researchers and policymakers about when particular design choices might be more or less appropriate for AI incident reporting.

Full version with exec. summary, extended discussion, & appendices — https://arxiv.org/abs/2511.05914

1 Introduction

General-purpose artificial intelligence (GPAI) systems, including large language models (LLMs), have recently contributed to a number of high-profile cases of harm. For instance, GPAI systems have helped perpetrate a $25.6 million financial scam (Chen and Magramo 2024), assisted in planning explosives attacks (Palmer 2025; Winter, Blankstein, and Planas 2025), created explicit deepfakes (Weatherbed 2024), accidentally deleted all of a company's code (Lee 2025), been demonstrated to be capable of blackmail and deception (Lynch et al. 2025), and spread election disinformation in the United States (US) and across the world (Seitz-Wald and Memoli 2024; RoW 2024). Increased AI capabilities, more agentic AI systems (Chan et al. 2023), and wider adoption of GPAI also suggest that GPAI will soon contribute to more, and more severe, safety incidents and rights incidents: events in which AI systems cause or nearly cause harm to people, property, the environment, legal or human rights, infrastructure, or the public interest.

Normal accident theory implies that in complex systems such as GPAI, severe incidents are inevitable over time (Bianchi, Cercas Curry, and Hovy 2023; Maas 2018); the risk of system failures and accidents rises as GPAI systems are becoming increasingly complex (Cook 2000; Maas 2018; Amodei et al. 2016; Jatho et al. 2023; Zaharia et al. 2024). Governance interventions occurring prior to model deployment such as audits also may not succeed in preventing all AI incidents (O'Brien, Ee, and Williams 2023). For instance, emergent capabilities in LLMs may arise unexpectedly after deployment (Zoph et al. 2022) and create unanticipated incident types (Casper et al. 2024; Woodside 2024). Moreover, because GPAI can be deployed across domains, GPAI-caused harms might involve a wide range of forms, hazards, settings, and affected parties (Jatho et al. 2023).

These dynamics increase the importance of post-deployment governance interventions—e.g., incident reporting, post-deployment evaluations, and other risk management practices—that can uncover new risks (Gailmard et al. 2025) and enable greater visibility into GPAI deployments (Schuett et al. 2024; EC 2025). In particular, incident reporting (IR) has been implemented in many safety-critical industries such as nuclear power, civilian aviation, healthcare, and pharmaceuticals. The experiences of these domains suggest that incident reporting could be an important mechanism for managing safety and rights risks from AI systems (Guha et al. 2023; Leveson 2011) by enabling learning and promoting accountability (WHO 2005).

AI incident reporting initiatives have yet to be implemented on a wide scale (Section 2). Moreover, discussions around incident reporting in the AI governance literature are still developing, and little guidance exists concerning AI incident reporting systems' institutional design—i.e., the construction of organizations, rules, and norms, that shape the tasks and responsibilities between actors (Klijn and Koppenjan 2006; Koppenjan and Groenewegen 2005).

We fill this gap in the literature by 1) introducing a framework for conceptualizing the institutional design of AI inci-

dent reporting systems, and 2) reviewing nine case studies of safety-critical industries to provide the first, systematic examination of design considerations for AI incident reporting in the US. Our scope is limited to institutional design, and we exclude, e.g., the design of AI systems themselves or the operational-level details of IR; our scope is also limited only to safety and rights incidents *caused by* GPAI, not on security incidents in which GPAI systems could be compromised by external actors (though these can be related). We further exclude discussions of AI whistleblowing (see Bullock and Arnold 2025; Hilton et al. 2024). We hope to inform US researchers and policymakers who are considering incident reporting as a mechanism for mitigating harms from GPAI.

2 The AI Incident Reporting Landscape

Incident reporting systems allow companies, users, victims of harm, and others with knowledge of incidents to convey such information to institutions with responsibility for or oversight over AI products. This information has a variety of downstream uses, such as improving systems to prevent reoccurrences of past harms (Turetsky, Nussbaum, and Tatar 2020), surfacing novel risks (NAIAC 2023; Dunbar 2014), evaluating safety mitigations (NIOSH 2005), or estimating risks for insurance policies (Kvist, Dattani, and Wang 2025). Appendix A lists more use cases for incident information.

In this article, we adopt an effects-based working definition of "incident:" incidents are events that either resulted in real-world harm (harm events) or that could have but did not ultimately result in harm (near misses). As of July 2025, no consensus definition for AI "incidents" has emerged, and we do not suggest that our definition is ideal or operationalizable. Rather, it aims to capture most events of interest while remaining consistent with safety science definitions.

We now review the literature on and current initiatives for GPAI incident reporting. Appendix A has more background.

Literature Review. The AI governance literature is supportive of incident reporting (Goodman 2024; Dixon and Frase 2024; DSIT 2023; Schuett et al. 2023). Existing literature has analyzed incident databases or operational-level factors such as documentation (McGregor 2021; Longpre et al. 2025; Cattell, Ghosh, and Kaffee 2024; OECD 2025); sources have also proposed a variety of incident reporting systems for GPAI with different or even conflicting design choices and policy goals (see Appendix A). To date, however, no comprehensive analysis exists in the literature concerning when particular institutional design choices may be more or less appropriate, which is the focus of this article.

Governmental AI Incident Reporting Initiatives. As of November 2025, China, the European Union (EU), California, and South Korea are the only jurisdictions that have adopted incident reporting mandates specific to general-purpose AI. In China, an August 2023 "generative AI" regulation requires that AI service providers create channels for users to report problems, and that companies report any "illegal content" to the government (CAC 2023). In the EU, the AI Act will (once implemented) require AI service providers to report "serious incidents" to national regulators (EU AI Act 2024, Art. 73). And in California, SB-53 was enacted in September 2025 and requires frontier developers to report certain safety and security incidents to the state government, as well as creating a hotline for public reporting (Wiener 2025). Policymakers in many other jurisdictions—including the US federal and other state governments (Bores 2025; Warner and Tillis 2024; NAIAC 2025)—have also called for or are considering proposals for AI incident reporting;[1] Appendix F has more examples of government IR initiatives.

Non-Governmental AI Incident Reporting Initiatives. Non-governmental private actors have also established systems for reporting AI incidents. These non-governmental systems take the form of voluntary, crowdsourced incident databases that solicit public submissions of alleged incidents, filter those submissions, and make the curated databases publicly available; the most prominent are the Artificial Intelligence Incident Database (AIID 2024a; McGregor 2021); the AI, Algorithmic, and Automation Incidents and Controversies Repository (AIAAIC 2024); and the AI Vulnerability Database (AVID 2024). Appendix A has more examples of non-governmental IR initiatives.

3 The Institutional Design of Incident Reporting Systems

Incident reporting systems are well-established in industries such as aviation and agriculture, so it is possible to study from other industries and assess whether their learnings can be applied to AI (Guha et al. 2023; West and Kak 2024; Gailmard et al. 2025). We thus adapt a three-step case study method from Raji et al. (2022), Ayling and Chapman (2022), and Stein et al. (2024) to identify design considerations for AI incident reporting.

First, we select nine safety-critical industries with robust incident reporting regimes as case studies, identified via seed articles (see Appendix B). Then, through a background literature review of IR in these industries and of the AI governance literature, we develop a framework for the institutional design of IR systems that consists of seven dimensions (Table 1). Finally, we purposively select IR systems from our nine case study industries and categorize them according to our framework. By identifying best or common practices in these industries, we extract design considerations for AI incident reporting systems and discuss when particular institutional design choices may be appropriate (Section 4).

Table 1 presents our framework for conceptualizing the institutional design of IR systems (full definitions in Appendix C). Our working definitions for the incident "lifecycle," as represented by the "level of risk materialization" dimension, are visualized in Figure 1. In particular, we make a distinction between AI issues (or AI flaws) and AI incidents. Issues are system *conditions* (hazards) that once exposed to an external environment (situations) become prerequisites for incidents, while incidents are *events* that could have caused harm (near misses) or did cause harm (harm event). Note that incidents can escalate into emergencies or crises (Gor and Iliadis 2025), which we do not discuss here.

[1] Some sectoral incident reporting mandates already exist, e.g., for autonomous vehicles (NHTSA 2023); however, these are not generally applicable to general-purpose AI systems.

Dimension	Definition
Policy Goal	The policy aim that the incident reporting system attempts to achieve: safety *learning* or *accountability* for harm.
Actors Submitting & Receiving Reports	Possible actors include users, victims of harm, third-party individuals or organizations, companies, industry employees, and governments at various levels. See Appendix C.2 for details.
Type of Incidents Reported	The type of incident reported in the system: *safety*, *rights*, or *security* incidents.
Level of Risk Materialization	The level of risk materialization reported in the system: *hazards*, *situations*, *near misses*, or *harm events*. See Figure 1.
Enforcement of Reporting	The procedures that incentivize actors to submit incident reports: *voluntary* or *mandatory* (by law).
Anonymity of Reporters	The actors who have access to the reporter's identity: *open*, *confidential*, or *anonymous*.
Post-Reporting Actions	The actions taken by the party receiving incident reports, after reports are received: *information sharing*, *information disclosure*, *audit*, or *regulatory action*.

Table 1: Seven dimensions of the institutional design of IR systems. Options for each dimension are italicized.

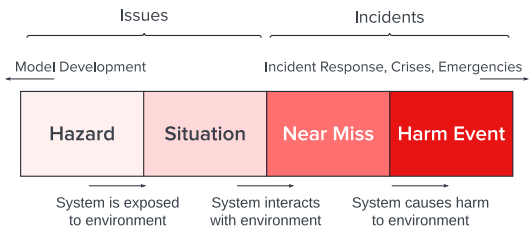

Figure 1: Lifecycle of an (AI) incident

4 Design Considerations for AI Incident Reporting Systems

Applying the framework in Section 3, we review incident reporting systems from nine safety-critical industries in the US: nuclear power, aviation, pesticides, pharmaceuticals, cybersecurity, dams, rail, workplace safety, and healthcare. We discuss design considerations from these nine industries for AI incident reporting. Our discussion is organized by design dimension; Appendix D contains full results of our review, and extended discussion is available in the full version of this article at https://arxiv.org/abs/2511.05914.

4.1 Policy Goal of the Incident Reporting System

Addressing Different Policy Goals of Incident Reporting. Policymakers and system operators may wish to achieve multiple policy goals with IR systems, such as safety learning (helping stakeholders adapt from past incidents and improving processes to prevent future harms) or corporate accountability (holding actors responsible for past harms via corrective actions or penalties). Although systems can in theory be oriented toward both learning and accountability, dual-goal systems are rare in practice because these goals often point toward opposing design choices (WHO 2005; Mills 2010). For instance, a voluntary reporting system (for learning) that encouraged responsible actors to report directly to a regulatory agency would be ineffective if reporters were deterred by the prospects of fines (for accountability). Thus, pursuing multiple goals via incident reporting systems could require the creation of multiple single-goal systems.

4.2 Actors Submitting & Receiving Reports

Shortcomings of Existing AI Incident Databases. Third-party incident databases such as the AIID, AIAAIC, or AVID are good first steps for creating visibility into the AI risk landscape (McGregor, Paeth, and Lam 2022; Whittlestone and Clark 2021). However, these databases lack stakeholder buy-in and often lack information necessary for safety learning (Richards, Benn, and Zilka 2025). Moreover, the stated goal of these databases is generally to facilitate safety learning; it is uncertain that databases alone can advance accountability beyond raising public awareness of AI-caused harms (Rodrigues, Resseguier, and Santiago 2023; Richards, Benn, and Zilka 2025).

Buy-in from both industry and government is needed for incident databases to contribute meaningfully to safety learning (Wolff 2014). An example is the Dam Directory of the National Performance on Dams Program, a database of dam safety incidents that contains materials drawn from many federal programs, dam engineers, professional organizations, and private collections; the program also maintains a real-time incident notification database with guidelines developed by state, federal, and industry dam safety engineers (McCann n.d.). Similar levels of buy-in have yet to be achieved by current AI incident databases—e.g., most top contributors to AIID are from civil society (AIID 2024b).

Because of these shortcomings of independent databases, the AI governance literature has called for official or centralized incident reporting systems (Brundage et al. 2022; Shevlane et al. 2023; Schuett 2023). Eight of the nine safety-critical industries we examined have implemented reporting regimes that include other systems beyond independent incident databases (Table 18 in Appendix D).

Facilitating Data Collection by Expanding Coverage. Reporting "coverage" refers to the audiences from whom incident reports are solicited and for whom an IR system is designed. Increasing coverage means that more information is collected from different parties, thus generating a more complete picture of the incident landscape (Wolff 2014; Mehran et al. 2004). Higher coverage is particularly critical when incident information is not concentrated but rather spread between different parties. Different parties may both have information about new, unreported incidents, as well as have corroborating or additional information about incidents already reported by others.

An illustrative case study of an incident reporting regime that has significant coverage is that of agricultural pesticides, which can cause health or environmental harm both to im-

mediate users and to downstream consumers (Damalas and Eleftherohorinos 2011). The complexity of pesticide distribution and the distributed nature of harm from pesticides has resulted in an enormous web of reporting and data sharing that involves, in various capacities: agriculture workers, poison control centers, doctors, medical labs, hospitals, most US state governments, the Department of Agriculture, the Environmental Protection Agency, the Bureau of Labor Statistics, the World Health Organization (WHO), the Intergovernmental Forum on Chemical Safety, and the UN Food and Agriculture Organization (Calvert et al. 2010). Similarly, AI systems can cause harm in a variety of ways—importantly, in ways not anticipated by traditional safety science such as catastrophic misuse, psychological manipulation, mental health harms, or civil liberties violations (Bengio et al. 2025; Weidinger et al. 2022; Slattery et al. 2025). Research into LLMs has found an impressive breadth of adoption (McElheran et al. 2024), which could foreshadow risk models and distribution chains that are at least as complicated as those in agriculture (see Hopkins et al. 2025). Information about AI incidents may, similar to pesticide incidents, be in the hands of many different actors due to this complexity (EC 2025) Thus, expanding reporting coverage may be important to facilitate a broader understanding of AI issues and incidents; these factors may justify reporting systems for diverse segments of society, especially as lower-severity incidents become more common.

Coverage can be increased by establishing multiple IR systems each targeted at different parties likely to have knowledge of incidents. Many design options can satisfy this criteria—e.g., multiple industry employee reporting systems, multiple industry organization reporting systems for organizations at different parts of the supply chain or for service providers in different verticals, or even one primary reporting system with many reporting formats for different parties.

Regulatory vs. Non-Regulatory Governmental System Operators. Whether governmental incident reporting systems are operated by regulatory vs. non-regulatory agencies can affect policy goals and post-reporting actions. Throughout our case study industries, regulators generally oversee reporting by companies, industry employees, citizens, third parties, or product users (Table 18)—in particular, regulators usually oversee reporting when industry actors are mandated to report incidents to the government. Additionally, systems that allow members of the public, third parties, or product users to report to regulators have been shown to improve product safety (Geier and Geier 2004), especially if regulators can take enforcement actions as a result of those reports (Raji et al. 2022). Public reporting hotlines may also have secondary benefits such as providing consumer information or access to resources (Calvert et al. 2010). Some industries such as aviation have multiple IR systems, some run by regulators and others by non-regulators.

On the other hand, systems where incidents are reported to non-regulatory agencies are generally non-punitive and oriented toward learning—since accountability could be difficult to achieve without regulatory authority. Thus, non-regulatory agencies usually help coordinate reporting for companies or industry employees. Because these agencies cannot take punitive actions against reporters, systems administered by non-regulators could also engender more trust in reporters (Mills and Reiss 2017). Moreover, regulators may sometimes have perverse incentives to avoid solicitation of incident reports if they believe that such reports may reflect negatively on their reputations; non-regulators tend to be more insulated from these political pressures may thus avoid perverse incentives (Etienne 2015; Christensen 2017).

The gold standard for a non-regulator managed voluntary reporting system is the Aviation Safety Reporting System (ASRS), whose administration the Federal Aviation Administration (FAA) entirely outsourced to the National Aeronautics and Space Administration (NASA). Because NASA does not regulate airlines, its designation as administrator helped promote trust and confidence in ASRS (Mills 2010)—especially as NASA can guarantee reporter anonymity and offer some liability protections (FAA 2021) (more on ASRS in Section 4.5). Similarly, the National Transportation Safety Board (NTSB) is a cross-cutting agency responsible for incident investigation; it has no regulatory authority, allows participation by outside experts and parties, and is insulated from the political process. These factors have all made the NTSB process more collaborative and made stakeholders more likely to participate to gain access to incident information (Fielding, Lo, and Yang 2010).

User-to-Company Reporting Systems. AI model developers and service providers may benefit from establishing internal systems for accepting and investigating reports or complaints from product users (McGregor et al. 2024). Currently, not all model developers have harm reporting channels for users or members of the public, and existing forms do not appear to always capture information about what harm or incident has occurred. User reports can be a valuable source of safety information (Sarkar and Rajagopalan 2018), so the absence of such reporting presents a gap that prevents companies from gaining important safety information and thus effectively learning from incidents.

4.3 Type of Incidents Reported

Distinguishing AI Safety, Security, and Rights Incidents. Operators of AI incident reporting systems may benefit from recognizing one possible categorization of incidents in terms of safety, security, and rights (defined in Appendix C); incidents in different categories may have different risk profiles and require different responses. Notably, safety and security are traditionally distinct fields. The goal of safety mitigations in AI is to protect external actors (i.e., users or other actors external to the AI systems) from harms that could be caused by such systems, whereas the goal of security is to protect *AI systems* from external actors (Khlaaf 2023; Roumani and Nwankpa 2020; Qi et al. 2024). Sharing information about security vulnerabilities (and perhaps also AI misuse) carries risks (Albakri, Boiten, and De Lemos 2018; Shevlane and Dafoe 2020; Grotto and Dempsey 2021), whereas sharing information about safety hazards is critical to learning. Security incidents can often be patched, whereas AI safety incidents might not be as easily addressed at a model level (Cattell, Ghosh, and Kaffee 2024; Kim, Kotha,

and Raghunathan 2024). And existing AI security frameworks differ widely from safety frameworks (Grotto and Dempsey 2021; MITRE 2024; Kumar et al. 2019).

Rights harms can be distinct from both security and safety harms. The notion of "harm" in safety science traditionally does not encompass harms such as civil liberties or human rights violations, and most safety IR systems do not support reporting rights incidents (Table 18). This limitation might be because historically, rights harms were primarily caused by people, not technical systems—but as GPAI systems become increasingly autonomous (Chan et al. 2023), they may cause new types of rights harms (Dobson et al. 2023; Hoffman and Frase 2023) or magnify existing rights harms or inequities (Critch and Russell 2023; Shelby et al. 2023).

On the other hand, safety, rights, and security incidents (in AI) can also occur conjointly (Qi et al. 2024; Johnston 2004). More research is needed to understand the confluence between different incident types, as well as how incident reporting and response may differ between incident types.

4.4 Level of Risk Materialization

Reporting AI Near Misses. Near misses—events that could have but ultimately did not cause harm—are a valuable source of data for safety learning, especially since the vast majority of safety incidents are not harm events but rather near misses. For instance, one hospital study found that less than 1% of reported safety incidents caused major harm, 18% caused minor/temporary harm, and 82% resulted in no harm (Mansouri et al. 2016); another estimate pins near misses as occurring at up to 300 times as often as harm events (Shojania et al. 2001). Reducing the number of near misses can also reduce the number of harm events (Jones, Kirchsteiger, and Bjerke 1999). If AI issues and incidents occur at rates proportionate to those in healthcare or other industries, then AI near misses may similarly be valuable for safety learning. Note that near misses and harm events are frequently reported via the same reporting system (Cheng et al. 2011; Manheim 2021; Table 18).

Governments and industry organizations can consider enabling AI near miss reporting by industry employees, users, third parties, or citizens—either through reporting systems specific to near misses or by allowing near misses to be reported along with harm events or issues. Near miss reporting systems tend to be voluntary, but they could also be mandatory; at the very least, there is consensus in the literature that self-reporting near misses should be non-punitive given that no harm was caused (Coyle 2005). Many industries have implemented near miss reporting systems—including aviation, nuclear, energy, chemical, and construction (Gnoni et al. 2022; Macrae 2014)—and some of these may be appropriate as guides for the AI context (Shrishak 2023).

Reporting of AI Issues vs. AI Incidents. AI issues may also be useful sources of information for safety learning, and the AI governance literature has recently begun to address "flaw disclosures" (Longpre et al. 2025; Cattell, Ghosh, and Kaffee 2024). Whether issues should be reported via the same reporting channels as incidents may depend on whether the reporting parties could have access to knowledge about either issues or incidents. Some incident reporting systems, such as those in nuclear energy and civilian aviation, appear to permit the reporting of both issues and incidents (Table 18). With AI systems, however, some hazards or situations may be discoverable only by third parties like experts or red-teamers, which may necessitate reporting systems for AI issues that are different from those designed for users, the public, or other actors. System administrators will also need to ensure that issue reporting does not overwhelm reporting systems, especially if the reports are primarily about product complaints rather than safety or rights issues (Havinga, Bancroft, and Rae 2021).

4.5 Enforcement of Reporting

Mandatory Reporting Thresholds. In some safety-critical industries, government-mandated incident reporting has seen some success. The EU AI Act has already taken a step in this direction (EU AI Act 2024, Art. 73), though its incident definition remains somewhat ambiguous. A phased reporting mandate—requiring that organizations report incidents shortly after discovery but permitting reports to be later amended with details and investigative results—may also be appropriate for high-severity AI incidents. Because LLM systems are complex, an initial report can notify government actors and determine if an official response is necessary while leaving organizations time to investigate (e.g., as in EU AI Act 2024, Art. 73(5)).

Reporting mandates generally require reports to be submitted to government agencies or other centralized actors, and mandates can be accountability-oriented since they may result in regulatory actions (WHO 2005). For instance, Kesari (2023) finds that mandatory cybersecurity incident reporting to state Attorneys General reduced consumer complaints of identity theft by 10.1% on average, possibly by deterring firms from engaging in unsafe practices.

Mandatory reporting generally applies to (high-severity) harm events (e.g., cases of death or serious injury): hospitals must report severe incidents to state agencies (CDPH 2021), airlines must report certain accidents and collisions to the FAA (FAA 2024), and medical device manufacturers must report drug reactions and device malfunctions to the Food and Drug Administration (FDA 2022). One analysis estimates that 90% of reports in the FDA's central database, MedWatch, are submitted by device manufacturers under mandatory requirements (Rajan, Kramer, and Kesselheim 2015). Thus, mandatory reports can be important mechanism for regulatory visibility.

Defining Reporting Requirements and Thresholds. Incident reporting systems usually need to develop clear, well-scoped reporting thresholds and definitions to be practically useful. Ideally, reporting thresholds capture all or most new hazards and incidents helpful for safety learning, but they cannot be so low that systems become inundated with reports that may or may not be useful (Johnson 2003). Approaches to defining incidents include encouraging reporting for any possible issue or incident, providing lists of reportable incident types, or establishing thresholds based on particular incident outcomes, system behaviors, or procedural violations. Which approach is appropriate for AI incident reporting is unclear; we offer examples of incident defini-

tions in Appendix E, but additional research is needed to operationalize AI incident definitions and taxonomies.

Vagueness in reporting thresholds hampers safety learning (Stavropoulou, Doherty, and Tosey 2015) and accountability while allowing industry actors to dodge compliance. For instance, when the National Highway Traffic Safety Administration (NHTSA)'s requirements were unclear as to what types of safety defect-related documents automobile manufacturers were required to turn over, manufacturers employed a range of strategies designed to evade responsibility. During NHTSA investigations, manufacturers denied that defects existed, responded to NHTSA requests for information with misleading and confusing language, and denied that the issues identified were safety-related—despite, in one case, a manufacturer later issuing a voluntary recall for the exact issues identified by NHTSA (Pecht et al. 2005). Similarly, Kesari (2023) also found that when state laws exempted breaches involving encrypted data from cybersecurity notification requirements, companies did not report breaches where data was stolen along with the encryption key. States that closed this loophole saw a 13.1% decrease in the number of data breach consumer complaints received.

Near miss reporting systems similarly need clear definitions (Gnoni et al. 2022) and to collect sufficient supplementary information to be useful for learning, e.g., information about the user, or interactions/communications between users or systems (Thoroman, Goode, and Salmon 2018).

Voluntary Reporting Systems. Mandatory reporting requirements alone may be insufficient to achieve safety learning—because high-severity harm events are rare, significant learning can occur from information about issues and near misses, and such information may not be centralized (Sections 4.2, 4.4). Voluntary systems can fill those gaps by allowing citizens, users, third-parties, companies, and industry employees to submit reports to centralized actors. Commentators have suggested that AI incident reporting adopt the model of the voluntary systems of the FDA (NAIAC 2023) or—more commonly—of the FAA (Shrishak 2023; Croxton et al. 2024).

The rest of this subsection examines the FAA's voluntary reporting programs (primarily ASRS), which are often considered the gold standard of voluntary reporting systems. ASRS's success is attributed to multiple factors. First, reports to the ASRS (run by NASA) are de-identified and—with certain exceptions—cannot be used by the FAA in regulatory enforcement actions (FAA 2021); these guarantees ensure that ASRS is not viewed as punitive while creating incentives for reporting (Cohen and Bagley 2020). Additionally, the FAA also intentionally aimed to achieve industry buy-in to ASRS by involving stakeholders early in its design process (ASRS 2001; Mills and Reiss 2017).

Ultimately, the FAA's systems have successfully enabled safety learning (Connell 2004; Mills and Reiss 2014). Since its inception in 1975, ASRS has received over 2,000,000 reports (Marfise and Hooey 2023), regularly generates feedback on safety hazards (Mills 2010), and surfaced unique information unavailable via other sources (Connell 2004).

The ASRS model, however, has not succeeded elsewhere. Inspired by ASRS, the Federal Railroad Administration has established the Confidential Close Call Reporting System (C^3RS), also administered by NASA (Ranney et al. 2019). But C^3RS failed to achieve industry buy-in: only 23 of 800 railroad companies in the US participated (GAO 2022), and some even withdrew participation because they perceived C^3RS to be ineffective (Jeffries and Buttigieg 2023).

It is unclear whether the FAA's voluntary systems are an appropriate model for AI incident reporting. As in rail, buy-in from AI developers may be difficult to achieve. Unlike aviation, current competitive dynamics in AI may inhibit voluntary information reporting (Beers 2025). For instance, developers have been reluctant to share data about their models, training data, and other technical features (Bommasani et al. 2024), but such information may be critical to enable safety learning. Aviation incidents are also industry-specific whereas harms from general-purpose AI systems are likely to traverse different verticals and involve significantly more actors than in the FAA's reporting scheme, which complicates reporting structures. ASRS aside, FAA's other voluntary programs like the Aviation Safety Action Program are administered in partnership with labor unions (Mills 2010), which are virtually non-existent in the AI industry. Finally, FAA is known to have established a highly collaborative relationship with industry via its voluntary reporting programs (Mills 2010), but such relationships in AI may raise concerns about regulatory capture (Wei et al. 2024).

4.6 Anonymity of Reporters

Anonymous or Confidential Reporting. Whether parties submitting incident reports should be offered anonymity or confidentiality depends heavily on the identity of the parties and their perceptions about the possibility of retaliation if they report. Anonymity may be beneficial to reduce individuals' fears of reprisal or reputational risk (van der Schaaf and Kanse 2004; Durant 2020; Beers 2025), to encourage reporting from parties who fall outside the chain of incident responsibility (e.g., third parties), or in independent databases. On the other hand, most mandatory systems in practice appear to be open or confidential when the reporting parties are industry organizations (Table 18).

4.7 Post-Reporting Actions

Facilitating Safety Learning After Reporting. Incident reporting is the first, but not the only step, toward safety learning and accountability. To improve safety, organizations must investigate, classify, and analyze incidents before implementing and monitoring safety interventions (Adole 2020; Briggs, Jeske, and Coventry 2017; Drupsteen, Groeneweg, and Zwetsloot 2013). However, this learning lifecycle is not yet mature in AI. AI incident databases have inspired research into incident types and taxonomies (e.g., McGregor, Paeth, and Lam 2022), and they can help us understand the risk landscape (Whittlestone and Clark 2021) and raise awareness of AI risks (Feffer, Martelaro, and Heidari 2023). However, centralized repositories are necessary for transparency, accountability, and analysis (Mandel and Runciman 2014; Lupo 2023). In addition, lack of transparency in many general-purpose AI systems (Bommasani et al. 2024) may impede safety learning research, and the

community has not yet developed consensus classification taxonomies, investigation methods, or evaluation/intervention processes (Paeth et al. 2024). Additional research is needed to adapt and operationalize the post-reporting learning lifecycle in the context of general-purpose AI systems.

Information Sharing Between IR Systems. After an incident report has been submitted, information must be aggregated and/or routed to relevant stakeholders to facilitate safety or accountability. Successful information sharing requires both identifying the correct actors and ensuring that information can be easily shared (e.g., via standardization and interoperability between systems).

Experience from other industries indicates that incident reporting systems at the national and international level are more concerned with incidents that are predictive of greater risks, high-severity incidents or incidents that could escalate into emergencies or crises, or incidents from which system- or industry-wide learning is possible (Barach and Small 2000; IAEA 2022; Novak 1985; NAIAC 2023). Local or industry-specific systems will be better positioned to handle incidents that are limited in scope or generalizability, and user reporting systems or third-party reporting systems to industry organizations may want to set lower thresholds so that low-severity incidents may be captured for learning (Frey et al. 2002; Webster 2016).

Standardization and interoperability between systems are also important to facilitate the flow of information (Shane 2024). A counterexample is the U.S. cybersecurity IR regime, in which 22 federal US agencies have implemented at least 45 sometimes-duplicative incident reporting requirements (DHS 2023; Kosseff 2016). Such fragmentation makes data aggregation difficult and hinders learning (Wood and Nash 2005). On the other hand, reporting systems can consider closer integration to avoid increasing the burden of filing reports, which can disincentivize reporting (Lubomski et al. 2004; Guffey, Culwick, and Merry 2014). Reporting standards may also need to carefully consider privacy policies, which must accommodate information aggregation (Dixon and Frase 2024), ensure that relevant actors can access information (Kolt et al. 2024), and also protect against (perceptions of) identity disclosure and retaliation (Section 4.5). Nascent efforts are attempting to create standards for GPAI incident documentation (Longpre et al. 2025; OECD 2025; Ezell, Roberts-Gaal, and Chan 2025), but industry adoption may pose a challenge.

Legal Liability and Regulatory Frameworks. Individuals and companies commonly cite legal uncertainty and fear of liability as top reasons for deciling to report (Nagamatsu, Kami, and Nakata 2009; IC IG 2023; Fukuda et al. 2010; Carlfjord, Öhrn, and Gunnarsson 2018). Policymakers who wish to facilitate incident reporting in the context of GPAI can consider clarifying legal frameworks for reporting up front, setting clear reporting incentives (Glendinning 2001; Vredenburgh 2002; Briggs, Jeske, and Coventry 2017), and communicating these guidelines to reporting parties. Some issues for policymakers to consider include what types of liability attach to reporting, whether incident reports are discoverable in court, the precision of the scope of reporting requirements, the relevant standards to apply where security incidents are intertwined (Johnson 2014), and how antitrust law interacts with safety information sharing (Anthropic 2023). In addition, which parties have access to incident data is a perennial source of concern, and determining with whom to share information will require balancing legal and competitive concerns with report receivers' interests in visibility and accountability (IC IG 2023).

Note that some legal incentives are commonly used in different systems. In mandatory reporting, penalties for failing to make reports may be effective (Grepperud 2005; Yew and Hadfield-Menell 2022). In voluntary reporting, limited liability protections for reporters are also common and can incentivize reporting: in a survey of healthcare providers, for instance, 72% of physicians indicated that they would be more likely to submit incident reports if reports were protected from legal discovery (Harper and Helmreich 2005).

Finally, legal loopholes that allow companies to avoid reports or disclosures can hinder incident reporting. Product manufacturers, for example, have obtained broad protective orders or confidential settlements in court to avoid public disclosure of defects and product safety issues (Engstrom et al. 2024; Egilman et al. 2020; Cohen and Bagley 2020; Saver 2017). Firms fearing that incident reports or documentation will be used in subsequent litigation may also intentionally keep less documentation, making IR requirements less useful (Schwarcz, Wolff, and Woods 2023). AI incident reporting regimes may wish to take note of these problems.

5 Limitations

The scope of our work is limited. Importantly, we do not conduct a full cost-benefit analysis of whether incident reporting systems are desirable in the context of GPAI systems; it is possible that goals of learning or accountability could be better achieved by other governance practices. Our case studies are also US-centric, and some lessons may not be easily transferred to non-US jurisdictions. Moreover, AI security incidents may require different processes and frameworks than safety and rights incidents (Qi et al. 2024), which are our primary focus here.

Our scope is also restricted to the institutional design of AI incident reporting systems. We do not address, for instance, many implementation features of incident reporting systems such as documentation or safety culture.

6 Conclusion

Incident reporting systems can help create a safer AI ecosystem and hold organizations responsible for harms from AI. Incident reporting systems' success is rooted in their institutional structures, and this paper provides the first systematic examination of how IR systems may be designed in the context of general-purpose AI. Through nine case studies of safety-critical industries, we provide institutional design considerations for GPAI incident reporting systems, and we discuss when particular design choices may be more or less appropriate based on stakeholders' goals and other factors. We hope to inform to US stakeholders interested in establishing incident reporting systems at a time when GPAI is seeing increased adoption across the economy.

Ethical Considerations Statement

This project was determined not to be human subjects research after an initial review by the RAND Human Subjects Protection Committee. No further review was required for this work.

Acknowledgements

KW was the primary author of this paper, and LH was responsible for supervision as well as review and editing; this work was initiated while KW and LH were affiliated with the Centre for the Governance of AI (KW as a Summer Fellow) and completed while KW and LH were affiliated with RAND. The authors are grateful for conversations with and feedback from (in random order): Karson Elmgren, Tommy Shaffer Shane, Morgan Simpson, Alan Chan, Markus Anderljung, Casey Mahoney, Emma Bluemke, Ben Garfinkel, Shaun Ee, Leonie Kessler, Fabian Ulmer, Gaurav Sett, Jason Greenlowe, Noam Kolt, Lisa Soder, Mauricio Baker, Toni Lorente, Cristian Trout, Michael Aird, Alexis Carlier, Ren Bin Lee Dixon, Francis Rhys Ward, and four anonymous AAAI reviewers. Funding for this work was provided by gifts from RAND supporters and by the Centre for the Governance of AI.

References

Adole, A. 2020. Accident and Incident Investigation. In Friend, M.; Stolzer, A.; and Aguiar, M., eds., *Safety Management Systems: Applications for the Aviation Industry*, 125–144. Lanham, MD: Benham Press. ISBN 978-1-64143-361-7.

AIAAIC. 2024. AIAAIC Repository. https://perma.cc/YX6A-FT2E. Accessed: 2023-10-03.

AIID. 2024a. Artificial Intelligence Incident Database (AIID). https://perma.cc/Z9TQ-HER2. Accessed: 2023-10-03.

AIID. 2024b. Submissions Leaderboard. https://perma.cc/9FRW-K2K8. Accessed: 2024-03-30.

Albakri, A.; Boiten, E.; and De Lemos, R. 2018. Risks of Sharing Cyber Incident Information. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, ARES '18, 1–10. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6448-5.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. arXiv:1606.06565.

Anthropic. 2023. Comment on FR Doc # 2023-07776 AI Accountability Policy Comment. *National Telecommunications and Information Administration*. https://perma.cc/6FJA-6RTA.

ASRS. 2001. ASRS: The Case for Confidential Incident Reporting Systems. Technical Report Pub. 60, NASA Aviation Safety Reporting System.

AVID. 2024. AI Vulnerability Database (AVID). https://perma.cc/ELL5-ZGU6. Accessed: 2023-10-03.

Ayling, J.; and Chapman, A. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, 2(3): 405–429.

Barach, P.; and Small, S. D. 2000. Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. *BMJ : British Medical Journal*, 320(7237): 759–763.

Beers, K. 2025. Beyond Corporate Promises: How Government Can Follow Through on AI Preparedness. https://perma.cc/B46A-RM4L.

Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; Heidari, H.; Ho, A.; Kapoor, S.; Khalatbari, L.; Longpre, S.; Manning, S.; Mavroudis, V.; Mazeika, M.; Michael, J.; Newman, J.; Ng, K. Y.; Okolo, C. T.; Raji, D.; Sastry, G.; Seger, E.; Skeadas, T.; South, T.; Strubell, E.; Tramèr, F.; Velasco, L.; Wheeler, N.; Acemoglu, D.; Adekanmbi, O.; Dalrymple, D.; Dietterich, T. G.; Felten, E. W.; Fung, P.; Gourinchas, P.-O.; Heintz, F.; Hinton, G.; Jennings, N.; Krause, A.; Leavy, S.; Liang, P.; Ludermir, T.; Marda, V.; Margetts, H.; McDermid, J.; Munga, J.; Narayanan, A.; Nelson, A.; Neppel, C.; Oh, A.; Ramchurn, G.; Russell, S.; Schaake, M.; Schölkopf, B.; Song, D.; Soto, A.; Tiedrich, L.; Varoquaux, G.; Yao, A.; Zhang, Y.-Q.; Albalawi, F.; Alserkal, M.; Ajala, O.; Avrin, G.; Busch, C.; Carvalho, A. C. P. d. L. F. d.; Fox, B.; Gill, A. S.; Hatip, A. H.; Heikkilä, J.; Jolly, G.; Katzir, Z.; Kitano, H.; Krüger, A.; Johnson, C.; Khan, S. M.; Lee, K. M.; Ligot, D. V.; Molchanovskyi, O.; Monti, A.; Mwamanzi, N.; Nemer, M.; Oliver, N.; Portillo, J. R. L.; Ravindran, B.; Rivera, R. P.; Riza, H.; Rugege, C.; Seoighe, C.; Sheehan, J.; Sheikh, H.; Wong, D.; and Zeng, Y. 2025. International AI Safety Report. Research Report DSIT 2025/001, UK Department of Science, Innovation & Technology, AI Action Summit.

Bianchi, F.; Cercas Curry, A.; and Hovy, D. 2023. Viewpoint: Artificial Intelligence Accidents Waiting to Happen? *J. Artif. Int. Res.*, 76.

Bommasani, R.; Klyman, K.; Kapoor, S.; Longpre, S.; Xiong, B.; Maslej, N.; and Liang, P. 2024. The Foundation Model Transparency Index v1.1: May 2024. arXiv:2407.12929.

Bores, A. 2025. A6453A RAISE Act. https://perma.cc/284Y-G8ME.

Briggs, P.; Jeske, D.; and Coventry, L. 2017. The Design of Messages to Improve Cybersecurity Incident Reporting. In Tryfonas, T., ed., *Human Aspects of Information Security, Privacy and Trust*, 3–13. Cham: Springer International Publishing. ISBN 978-3-319-58460-7.

Brundage, M.; Mayer, K.; Eloundou, T.; Agarwal, S.; Adler, S.; Krueger, G.; Leike, J.; and Mishkin, P. 2022. Lessons learned on language model safety and misuse. https://perma.cc/G4M6-59CM.

Bullock, C.; and Arnold, M. 2025. Protecting AI Whistleblowers. *Lawfare*. https://perma.cc/6E2Z-BANM. Accessed: 2025-07-25.

CAC. 2023. Translation: Interim Measures for the Management of Generative Artificial Intelligence Services. *China Law Translate (Translation); Cyberspace Administration*

of China (CAC) (Original). https://perma.cc/UK9V-N3NX. Accessed: 2023-09-30.

Calvert, G. M.; Mehler, L. N.; Alsop, J.; De Vries, A. L.; and Besbelli, N. 2010. Surveillance of Pesticide-Related Illness and Injury in Humans. In Krieger, R., ed., *Hayes' Handbook of Pesticide Toxicology (Third Edition)*, 1313–1369. New York: Academic Press. ISBN 978-0-12-374367-1.

Carlfjord, S.; Öhrn, A.; and Gunnarsson, A. 2018. Experiences from ten years of incident reporting in health care: a qualitative study among department managers and coordinators. *BMC Health Services Research*, 18: 113.

Casper, S.; Schulze, L.; Patel, O.; and Hadfield-Menell, D. 2024. Defending Against Unforeseen Failure Modes with Latent Adversarial Training. arXiv:2403.05030.

Cattell, S.; Ghosh, A.; and Kaffee, L.-A. 2024. Coordinated Disclosure for AI: Beyond Security Vulnerabilities. arXiv:2402.07039.

CDPH. 2021. DPH-11-023 Adverse Events Reporting. *California Department of Public Health (CDPH)*. https://perma.cc/KF4N-KDRE. Accessed: 2024-04-28.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; Lin, M.; Mayhew, A.; Collins, K.; Molamohammadi, M.; Burden, J.; Zhao, W.; Rismani, S.; Voudouris, K.; Bhatt, U.; Weller, A.; Krueger, D.; and Maharaj, T. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666.

Chen, H.; and Magramo, K. 2024. Finance worker pays out $25 million after video call with deepfake 'chief financial officer'. *CNN*. https://perma.cc/KF4N-KDRE. Accessed: 2024-02-06.

Cheng, L.; Sun, N.; Li, Y.; Zhang, Z.; Wang, L.; Zhou, J.; Liang, M.; Cui, X.; Gao, G.; and Yuan, Q. 2011. International comparative analyses of incidents reporting systems for healthcare risk management. *Journal of Evidence-Based Medicine*, 4(1): 32–47.

Christensen, I. 2017. The elements of a commercial human spaceflight safety reporting system. *Acta Astronautica*, 139: 228–232.

Cohen, I. G.; and Bagley, N. 2020. Private Rights and the Public Interest in Drug and Medical Device Litigation. *JAMA Internal Medicine*, 180(2): 299–300.

Connell, L. J. 2004. Cross-Industry Applications of a Confidential Reporting Model. In Phimister, J. R.; Bier, V. M.; and Kunreuther, H., eds., *Accident Precursor Analysis and Management: Reducing Technological Risk Through Diligence*, 139–146. Washington, DC: The National Academies Press. ISBN 978-0-309-09216-6 |.

Cook, R. 2000. How Complex Systems Fail. https://perma.cc/2L3L-8VE6. Accessed: 2024-03-21.

Coyle, G. A. 2005. Designing and Implementing a Close Call Reporting System. *Nursing Administration Quarterly*, 29(1): 57.

Critch, A.; and Russell, S. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. arXiv:2306.06924.

Croxton, J.; Robusto, D.; Thallam, S.; and Calidas, D. 2024. How to Create an AI Incident Reporting System. https://perma.cc/Y6HY-NAR6.

Damalas, C. A.; and Eleftherohorinos, I. G. 2011. Pesticide Exposure, Safety Issues, and Risk Assessment Indicators. *International Journal of Environmental Research and Public Health*, 8(5): 1402–1419.

DHS. 2023. Harmonization of Cyber Incident Reporting to the Federal Government. Technical report, United States Department of Homeland Security.

Dixon, R. B. L.; and Frase, H. 2024. An Argument for Hybrid AI Incident Reporting. Technical report, CSET.

Dobson, S.-J.; Margolis, P.; Ciclitra, K.; and Altaf, Y. 2023. Intangible risks of modern products. https://perma.cc/GJV3-8RQU.

Drupsteen, L.; Groeneweg, J.; and Zwetsloot, G. I. 2013. Critical Steps in Learning From Incidents: Using Learning Potential in the Process From Reporting an Incident to Accident Prevention. *International Journal of Occupational Safety and Ergonomics*, 19(1): 63–77.

DSIT. 2023. Emerging processes for frontier AI safety. Technical report, United Kingdom Department for Science, Innovation & Technology.

Dunbar, M. 2014. Hazard Identification and Risk Assessment. *Safety First*, 34–36.

Durant, J. L. 2020. Ignorance loops: How non-knowledge about bee-toxic agrochemicals is iteratively produced. *Social Studies of Science*, 50(5): 751–777.

EC. 2025. Code of Practice for General-Purpose AI Models: Safety and Security Chapter. *European Commission*. https://perma.cc/U6FM-RU8Y.

Egilman, A. C.; Kesselheim, A. S.; Krumholz, H. M.; Ross, J. S.; Kim, J.; and Kapczynski, A. 2020. Confidentiality Orders and Public Interest in Drug and Medical Device Litigation. *JAMA Internal Medicine*, 180(2): 292–299.

Engstrom, N. F.; Engstrom, D. F.; Gelbach, J. B.; Peters, A.; and Schaffer-Neitz, A. 2024. Secrecy by Stipulation. SSRN:4811151.

Etienne, J. 2015. The Politics of Detection in Business Regulation. *Journal of Public Administration Research and Theory*, 25(1): 257–284.

European Parliament; and Council of the European Union. 2024. 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. (EU AI Act). *Official Journal of the European Union*. https://perma.cc/Z5UR-MUY4.

Ezell, C.; Roberts-Gaal, X.; and Chan, A. 2025. Incident Analysis for AI Agents. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 865–878.

FAA. 2021. AC 00-46F: Aviation Safety Reporting Program. *United States Federal Aviation Administration (FAA)*. https://perma.cc/K5B3-T299.

FAA. 2024. Aeronautical Information Manual: Official Guide to Basic Flight Information and ATC Procedures. *United States Federal Aviation Administration (FAA)*. https://perma.cc/UXT8-FAN7. Accessed: 2024-04-28.

FDA. 2022. Providing Submissions in Electronic Format — Postmarketing Safety Reports. *United States Food and Drug Administration (FDA)*. https://perma.cc/V2GJ-NSFH. Accessed: 2024-04-28.

Feffer, M.; Martelaro, N.; and Heidari, H. 2023. The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements. arXiv:2310.06269.

Fielding, E.; Lo, A. W.; and Yang, J. H. 2010. The National Transportation Safety Board: A Model for Systemic Risk Management. SSRN:1695781.

Frey, B.; Buettiker, V.; Hug, M. I.; Waldvogel, K.; Gessler, P.; Ghelfi, D.; Hodler, C.; and Baenziger, O. 2002. Does critical incident reporting contribute to medication error prevention? *European Journal of Pediatrics*, 161(11): 594–599.

Fukuda, H.; Imanaka, Y.; Hirose, M.; and Hayashida, K. 2010. Impact of system-level activities and reporting design on the number of incident reports for patient safety. *BMJ Quality & Safety*, 19(2): 122–127.

Gailmard, L.; Spence, D.; Lawrence, C.; and Ho, D. E. 2025. Known Unknowns and Unknown Unknowns: Designing a Scalable Adverse Event Reporting System for AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1004–1017.

GAO. 2022. Federal Railroad Administration: Better Communication of Safety Information Could Improve the Close Call System. Technical Report GAO-23-105287, United States General Accounting Office (GAO).

Geier, D. A.; and Geier, M. R. 2004. A review of the Vaccine Adverse Event Reporting System database. *Expert Opinion on Pharmacotherapy*, 5(3): 691–698.

Glendinning, P. M. 2001. Employee safety incentives: A best practices survey of human resource practitioners. *Professional Safety*, 46(2): 22–24.

Gnoni, M. G.; Tornese, F.; Guglielmi, A.; Pellicci, M.; Campo, G.; and De Merich, D. 2022. Near miss management systems in the industrial sector: A literature review. *Safety Science*, 150: 105704.

Goodman, E. P. 2024. AI Accountability Policy Report. Technical report, United States National Telecommunications and Information Administration (NTIA).

Gor, G.; and Iliadis, N. 2025. What Is an Artificial Intelligence Crisis and What Does It Mean to Prepare for One? https://perma.cc/F6MA-FNLR.

Grepperud, S. 2005. Medical Errors: Mandatory Reporting, Voluntary Reporting, or Both? *European Journal of Law and Economics*, 20(1): 99–112.

Grotto, A.; and Dempsey, J. X. 2021. Vulnerability Disclosure and Management for AI/ML Systems: A Working Paper with Policy Recommendations. https://perma.cc/6JLT-NW67.

Guffey, P. J.; Culwick, M.; and Merry, A. F. 2014. Incident Reporting at the Local and National Level. *International Anesthesiology Clinics*, 52(1): 69.

Guha, N.; Lawrence, C.; Gailmard, L. A.; Rodolfa, K.; Surani, F.; Bommasani, R.; Raji, I.; Cuéllar, M.-F.; Honigsberg, C.; Liang, P.; and Ho, D. E. 2023. AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *George Washington Law Review*.

Harper, M. L.; and Helmreich, R. L. 2005. Identifying Barriers to the Success of a Reporting System. In Henriksen, K.; Battles, J. B.; Marks, E. S.; and Lewin, D. I., eds., *Advances in Patient Safety: From Research to Implementation*, volume 3. Agency for Healthcare Research and Quality.

Havinga, J.; Bancroft, K.; and Rae, A. 2021. Hazard reporting: How can it improve safety? *Safety Science*, 142: 105365.

Hilton, J.; Kokotajlo, D.; Kumar, R.; Nanda, N.; Saunders, W.; Wainwright, C.; Ziegler, D.; Anonymous; Anonymous; Anonymous; Anonymous; Anonymous; and Anonymous. 2024. A Right to Warn about Advanced Artificial Intelligence (Open Letter). https://perma.cc/6M2V-PPTD.

Hoffman, M.; and Frase, H. 2023. Adding Structure to AI Harm. Technical report, Center for Security and Emerging Technology.

Hopkins, A.; Cen, S. H.; Ilyas, A.; Struckman, I.; Videgaray, L.; and Madry, A. 2025. AI Supply Chains: An Emerging Ecosystem of AI Actors, Products, and Services. arXiv:2504.20185.

IAEA. 2022. IAEA Nuclear Safety and Security Glossary. Text, International Atomic Energy Agency (IAEA).

IC IG. 2023. Joint Report on the Implementation of the Cybersecurity Information Sharing Act of 2015. Technical Report AUD 2023 002, United States Office of the Inspector General of the Intelligence Community (IC IG).

Jatho, E. W.; Mailloux, L. O.; Williams, E. D.; McClure, P.; and Kroll, J. A. 2023. Concrete Safety for ML Problems: System Safety for ML Development and Assessment. arXiv:2302.02972.

Jeffries, I.; and Buttigieg, P. 2023. Letter from the Association of American Railroads to the U.S. Secretary of Transportation. https://perma.cc/9DAA-2LKW.

Johnson, C. 2003. *Failure in Safety-Critical Systems: A Handbook of Accident and Incident Reporting*. Glasgow, Scotland: University of Glasgow Press. ISBN 0-85261-784-4.

Johnson, C. W. 2014. Tools and Techniques for Reporting and Analysing the Causes of Cyber-Security Incidents in Safety-Critical Systems. In *Proceedings of the 9th IET International Conference on System Safety and Cyber Security*, 3.2.1–3.2.1. Manchester: The Institution of Engineering and Technology. ISBN 978-1-84919-940-7.

Johnston, R. G. 2004. Adversarial safety analysis: Borrowing the methods of security vulnerability assessments. *Journal of Safety Research*, 35(3): 245–248.

Jones, S.; Kirchsteiger, C.; and Bjerke, W. 1999. The importance of near miss reporting to further improve safety performance. *Journal of Loss Prevention in the Process Industries*, 12(1): 59–67.

Kesari, A. 2023. Do Data Breach Notification Laws Work? *N.Y.U. Journal of Legislation and Public Policy*, 26(1): 173–237.

Khlaaf, H. 2023. Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems. *Trail of Bits*. https://perma.cc/HX8L-Y3YV.

Kim, T.; Kotha, S.; and Raghunathan, A. 2024. Jailbreaking is Best Solved by Definition. arXiv:2403.14725.

Klijn, E.-H.; and Koppenjan, J. F. M. 2006. Institutional design: Changing institutional features of networks. *Public Management Review*, 8(1): 141–160.

Kolt, N.; Anderljung, M.; Barnhart, J.; Brass, A.; Esvelt, K.; Hadfield, G. K.; Heim, L.; Rodriguez, M.; Sandbrink, J. B.; and Woodside, T. 2024. Responsible Reporting for Frontier AI Development. arXiv:2404.02675.

Koppenjan, J.; and Groenewegen, J. 2005. Institutional design for complex technological systems. *International Journal of Technology, Policy and Management*, 5(3): 240–257.

Kosseff, J. 2016. Positive Cybersecurity Law: Creating a Consistent and Incentive-Based System. *Chapman Law Review*, 19: 401.

Kumar, R. S. S.; Brien, D. O.; Albert, K.; Viljöen, S.; and Snover, J. 2019. Failure Modes in Machine Learning Systems. arXiv:1911.11034.

Kvist, R.; Dattani, R.; and Wang, B. 2025. Underwriting Superintelligence. *Underwriting Superintelligence*. https://perma.cc/P25V-69AC. Accessed: 2025-07-25.

Lee, C. M. 2025. Replit's CEO apologizes after its AI agent wiped a company's code base in a test run and lied about it. *Business Insider*. https://perma.cc/V7C9-335S. Accessed: 2025-07-24.

Leveson, N. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press. ISBN 978-0-262-01662-9.

Longpre, S.; Klyman, K.; Appel, R. E.; Kapoor, S.; Bommasani, R.; Sahar, M.; McGregor, S.; Ghosh, A.; Blili-Hamelin, B.; Butters, N.; Nelson, A.; Elazari, D. A.; Sellars, A.; Ellis, C. J.; Sherrets, D.; Song, D.; Geiger, H.; Cohen, I.; McIlvenny, L.; Srikumar, M.; Jaycox, M. M.; Anderljung, M.; Johnson, N. F.; Carlini, N.; Miailhe, N.; Marda, N.; Henderson, P.; Portnoff, R. S.; Weiss, R.; Westerhoff, V.; Jernite, Y.; Chowdhury, R.; Liang, P.; and Narayanan, A. 2025. Position: In-House Evaluation Is Not Enough. Towards Robust Third-Party Evaluation and Flaw Disclosure for General-Purpose AI. In *Proceedings of the 42nd International Conference on Machine Learning*.

Lubomski, L. H.; Pronovost, P. J.; Thompson, D. A.; Holzmueller, C. G.; Dorman, T.; Morlock, L. L.; Dickman, F.; Fahey, M.; and Wu, A. W. 2004. Building a better incident reporting system: Perspectives from a multisite project. *Journal of Clinical Outcomes Management*, 11(5): 275–280.

Lupo, G. 2023. Risky Artificial Intelligence: The Role of Incidents in the Path to AI Regulation. *Law, Technology and Humans*, 5(1): 133–152.

Lynch, A.; Wright, B.; Larson, C.; Troy, K. K.; Ritchie, S. J.; Mindermann, S.; Perez, E.; and Hubinger, E. 2025. Agentic Misalignment: How LLMs Could be an Insider Threat. https://perma.cc/8U2C-W527.

Maas, M. M. 2018. Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 223–228. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6012-8.

Macrae, C. 2014. *Close Calls: Managing Risk and Resilience in Airline Flight Safety*. Springer. ISBN 978-1-137-37612-1.

Mandel, C.; and Runciman, W. 2014. System for Reporting and Analysing Incidents. In Lau, L.; and Ng, K.-H., eds., *Radiological Safety and Quality: Paradigms in Leadership and Innovation*, 203–221. Dordrecht: Springer Netherlands. ISBN 978-94-007-7256-4.

Manheim, D. B. 2021. Results of a 2020 Survey on Reporting Requirements and Practices for Biocontainment Laboratory Accidents. *Health Security*, 19(6): 642–651.

Mansouri, M.; Aran, S.; Harvey, H. B.; Shaqdan, K. W.; and Abujudeh, H. H. 2016. Rates of safety incident reporting in MRI in a large academic medical center. *Journal of Magnetic Resonance Imaging*, 43(4): 998–1007.

Marfise, E.; and Hooey, B. 2023. NASA Aviation Safety Reporting System (ASRS). *United States Federal Aviation Administration, Orlando Flight Standards District Office, Certified Flight Instructor Special Emphasis Quarterly Webinar*. https://perma.cc/28S3-AY7P.

McCann, M. W. n.d. National Performance of Dams Program: An Archive of Dam Information and Experience. Technical report, National Performance of Dams Program (NPDP).

McElheran, K.; Li, J. F.; Brynjolfsson, E.; Kroff, Z.; Dinlersoz, E.; Foster, L.; and Zolas, N. 2024. AI adoption in America: Who, what, and where. *Journal of Economics & Management Strategy*, 33(2): 375–415.

McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15458–15463.

McGregor, S.; Ettinger, A.; Judd, N.; Albee, P.; Jiang, L.; Rao, K.; Smith, W.; Longpre, S.; Ghosh, A.; Fiorelli, C.; Hoang, M.; Cattell, S.; and Dziri, N. 2024. To Err is AI : A Case Study Informing LLM Flaw Reporting Practices. arXiv:2410.12104.

McGregor, S.; Paeth, K.; and Lam, K. 2022. Indexing AI Risks with Incidents, Issues, and Variants. arXiv:2211.10384.

Mehran, R.; Leon, M. B.; Feigal, D. A.; Jefferys, D.; Simons, M.; Chronos, N.; Fogarty, T. J.; Kuntz, R. E.; Baim, D. S.; and Kaplan, A. V. 2004. Post-Market Approval Surveillance. *Circulation*, 109(25): 3073–3077.

Mills, R. W. 2010. The Promise of Collaborative Voluntary Partnerships: Lessons from the Federal Aviation Administration | IBM Center for The Business of Government. Technical report, IBM Center for the Business of Government.

Mills, R. W.; and Reiss, D. R. 2014. Secondary learning and the unintended benefits of collaborative mechanisms: The Federal Aviation Administration's voluntary disclosure programs. *Regulation & Governance*, 8(4): 437–454.

Mills, R. W.; and Reiss, D. R. 2017. The role of trust in the regulation of complex and high-risk industries: the case of the U.S. Federal Aviation Administration's voluntary disclosure programs. In *Trust in Regulatory Regimes*, 37–59. Edward Elgar Publishing. ISBN 978-1-78536-557-7.

MITRE. 2024. MITRE ATLAS. https://perma.cc/UTP3-GD5G. Accessed: 2024-03-22.

Nagamatsu, S.; Kami, M.; and Nakata, Y. 2009. Healthcare safety committee in Japan: mandatory accountability reporting system and punishment. *Current Opinion in Anesthesiology*, 22(2): 199.

NAIAC. 2023. RECOMMENDATION: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting. *The National Artificial Intelligence Advisory Committee (NAIAC)*. https://web.archive.org/web/20250112122101/https://ai.gov/wp-content/uploads/2023/12/Recommendation_Improve-Monitoring-of-Emerging-Risks-from-AI-through-Adverse-Event-Reporting.pdf. Accessed: 2025-07-24.

NAIAC. 2025. NAIAC Insights for the Administration of President Donald J. Trump. *The National Artificial Intelligence Advisory Committee (NAIAC)*. https://perma.cc/3Q2L-RK9Q.

NHTSA. 2023. Second Amended Standing General Order 2021-01: Incident Reporting for Automated Driving Systems and Level 2 Advanced Driver Assistance Systems. *United States National Highway Traffic Safety Administration (NHTSA)*. https://perma.cc/9KWV-QPWJ.

NIOSH. 2005. Pesticide-Related Illness and Injury Surveillance: A How-To Guide for State-Based Programs. *United States National Institute for Occupational Safety and Health (NIOSH)*. https://perma.cc/55KS-NZRX.

Novak, S. 1985. The IAEA Incident Reporting System and its possible role in the improvement of the safety and availability of nuclear power plants. In *Proceedings of an International Symposium on Advances in Nuclear Power Plant Availability, Maintainability and Operation*. Vienna: International Atomic Energy Agency. ISBN 92-0-050085-4.

O'Brien, J.; Ee, S.; and Williams, Z. 2023. Deployment corrections: An incident response framework for frontier AI models. Technical report, Institute for AI Policy and Strategy.

OECD. 2025. Towards a common reporting framework for AI incidents. Technical Report 34, Organisation for Economic Cooperation and Development, Paris.

Paeth, K.; Atherton, D.; Pittaras, N.; Frase, H.; and McGregor, S. 2024. Lessons for Editors of AI Incidents from the AI Incident Database. arXiv:2409.16425.

Palmer, A. 2025. FBI says Palm Springs bombing suspects used AI chat program to help plan attack. *CNBC*. https://perma.cc/82PG-52H5. Accessed: 2025-07-24.

Pecht, M.; Ramakrishnan, A.; Fazio, J.; and Nash, C. 2005. The role of the U.S National Highway Traffic Safety Administration in automotive electronics reliability and safety assessment. *IEEE Transactions on Components and Packaging Technologies*, 28(3): 571–580.

Qi, X.; Huang, Y.; Zeng, Y.; Debenedetti, E.; Geiping, J.; He, L.; Huang, K.; Madhushani, U.; Sehwag, V.; Shi, W.; Wei, B.; Xie, T.; Chen, D.; Chen, P.-Y.; Ding, J.; Jia, R.; Ma, J.; Narayanan, A.; Su, W. J.; Wang, M.; Xiao, C.; Li, B.; Song, D.; Henderson, P.; and Mittal, P. 2024. AI Risk Management Should Incorporate Both Safety and Security. arXiv:2405.19524.

Rajan, P. V.; Kramer, D. B.; and Kesselheim, A. S. 2015. Medical Device Postapproval Safety Monitoring. *Circulation: Cardiovascular Quality and Outcomes*, 8(1): 124–131.

Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 557–571. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9247-1.

Ranney, J. M.; Davey, M.; Morell, J.; Zuschlag, M.; and Kidda, S. 2019. Confidential Close Call Reporting System (C3RS) Lessons Learned Evaluation – Final Report. Technical Report DOT/FRA/ORD-19/01, United States Federal Railroad Administration (FRA).

Richards, I.; Benn, C.; and Zilka, M. 2025. From Incidents to Insights: Patterns of Responsibility following AI Harms. arXiv:2505.04291.

Rodrigues, R.; Resseguier, A.; and Santiago, N. 2023. When Artificial Intelligence Fails: The Emerging Role of Incident Databases. *Public Governance, Administration and Finances Law Review*, 8(2): 17–28.

Roumani, Y.; and Nwankpa, J. 2020. Examining Exploitability Risk of Vulnerabilities: A Hazard Model. *Communications of the Association for Information Systems*, 46(1).

RoW. 2024. Rest of World's 2024 AI elections tracker. https://perma.cc/4DSM-24KS. Accessed: 2025-07-28.

Sarkar, S.; and Rajagopalan, B. 2018. Consumer safety complaints and organizational learning: evidence from the automotive industry. *International Journal of Quality & Reliability Management*, 35(10): 2094–2118.

Saver, R. S. 2017. Deciphering the Sunshine Act: Transparency Regulation and Financial Conflicts in Health Care. *American Journal of Law & Medicine*, 43(4): 303–343.

Schuett, J. 2023. Risk management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, 1–19.

Schuett, J.; Anderljung, M.; Carlier, A.; Koessler, L.; and Garfinkel, B. 2024. From Principles to Rules: A Regulatory

Approach for Frontier AI. In Bullock, J.; Chen, Y.-C.; Himmelreich, J.; Hudson, V. M.; Korinek, A.; Young, M. M.; and Zhang, a. B., eds., *The Oxford Handbook of AI Governance*, Oxford Handbooks. Oxford, New York: Oxford University Press. ISBN 978-0-19-757932-9.

Schuett, J.; Dreksler, N.; Anderljung, M.; McCaffary, D.; Heim, L.; Bluemke, E.; and Garfinkel, B. 2023. Towards best practices in AGI safety and governance: A survey of expert opinion. Technical report, Centre for the Governance of AI.

Schwarcz, D.; Wolff, J.; and Woods, D. W. 2023. How Privilege Undermines Cybersecurity. *Harvard Journal of Law & Technology*, 36(2).

Seitz-Wald, A.; and Memoli, M. 2024. Fake Joe Biden robocall tells New Hampshire Democrats not to vote Tuesday. *NBC News*. https://perma.cc/4XAJ-WVQ6. Accessed: 2024-02-08.

Shane, T. S. 2024. AI incident reporting: Addressing a gap in the UK's regulation of AI. Technical report, Centre for Long-Term Resilience.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 723–741. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0231-0.

Shevlane, T.; and Dafoe, A. 2020. The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 173–179. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-7110-0.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; Ho, L.; Siddarth, D.; Avin, S.; Hawkins, W.; Kim, B.; Gabriel, I.; Bolina, V.; Clark, J.; Bengio, Y.; Christiano, P.; and Dafoe, A. 2023. Model evaluation for extreme risks. arXiv:2305.15324.

Shojania, K. G.; Duncan, B. W.; McDonald, K. M.; Wachter, R. M.; and Markowitz, A. J. 2001. Making health care safer: a critical analysis of patient safety practices. Evidence Report/Technology Assessment No. 43, United States Department of Health and Human Services (HHS).

Shrishak, K. 2023. How to deal with an AI near-miss: Look to the skies. *Bulletin of the Atomic Scientists*, 79(3): 166–169.

Slattery, P.; Saeri, A. K.; Grundy, E. A. C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2025. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622.

Stavropoulou, C.; Doherty, C.; and Tosey, P. 2015. How Effective Are Incident-Reporting Systems for Improving Patient Safety? A Systematic Literature Review. *The Milbank Quarterly*, 93(4): 826–866.

Stein, M.; Gandhi, M.; Kriecherbauer, T.; Oueslati, A.; and Trager, R. 2024. Public vs Private Bodies: Who Should Run Advanced AI Evaluations and Audits? A Three-Step Logic Based on Case Studies of High-Risk Industries. arXiv:2407.20847.

Thoroman, B.; Goode, N.; and Salmon, P. 2018. System thinking applied to near misses: a review of industry-wide near miss reporting systems. *Theoretical Issues in Ergonomics Science*, 19(6): 712–737.

Turetsky, D.; Nussbaum, B.; and Tatar, U. 2020. Success Stories in Cybersecurity Information Sharing. In *Proceedings of the 2018 Cybersecurity Conference*. Albany, New York: University at Albany.

van der Schaaf, T.; and Kanse, L. 2004. Biases in incident reporting databases: an empirical study in the chemical process industry. *Safety Science*, 42(1): 57–67.

Vredenburgh, A. G. 2002. Organizational safety: Which management practices are most effective in reducing employee injury rates? *Journal of Safety Research*, 33(2): 259–276.

Warner, M.; and Tillis, T. 2024. S. 4230 Secure A.I. Act of 2024. https://perma.cc/7F7W-FU8K.

Weatherbed, J. 2024. Trolls have flooded X with graphic Taylor Swift AI fakes. *The Verge*. https://perma.cc/HK8D-N4CH. Accessed: 2024-02-06.

Webster, C. S. 2016. Safety in unpredictable complex systems – a framework for the analysis of safety derived from the nuclear power industry. *Prometheus*, 34(2): 115–132.

Wei, K.; Ezell, C.; Gabrieli, N.; and Deshpande, C. 2024. How Do AI Companies "Fine-Tune" Policy? Examining Regulatory Capture in AI Governance. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1539–1555.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 214–229. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.

West, S. M.; and Kak, A. 2024. Lessons from the FDA for AI. Technical report, AI Now Institute.

Whittlestone, J.; and Clark, J. 2021. Why and How Governments Should Monitor AI Development. arXiv:2108.12427.

WHO. 2005. WHO Draft Guidelines for Adverse Event Reporting and Learning Systems: From Information to Action. Technical Report WHO/EIP/SPO/QPS/05.3, World Health Organization (WHO).

Wiener, S. 2025. SB-53 Transparency in Frontier Artificial Intelligence Act. https://perma.cc/HY54-YDXU.

Winter, T.; Blankstein, A.; and Planas, A. 2025. Driver in Las Vegas Cybertruck explosion used ChatGPT to plan blast, authorities say. *NBC News*. https://perma.cc/97A7-VCF7. Accessed: 2025-07-24.

Wolff, J. 2014. Models for Cybersecurity Incident Information Sharing and Reporting Policies. SSRN:2587398.

Wood, K. E.; and Nash, D. B. 2005. Mandatory State-Based Error-Reporting Systems: Current and Future Prospects. *American Journal of Medical Quality*, 20(6): 297–303.

Woodside, T. 2024. Emergent Abilities in Large Language Models: An Explainer. Technical report, Center for Security and Emerging Technology.

Yew, R.-J.; and Hadfield-Menell, D. 2022. A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 823–830. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9247-1.

Zaharia, M.; Khattab, O.; Chen, L.; Davis, J. Q.; Miller, H.; Potts, C.; Zou, J.; Carbin, M.; Frankle, J.; Rao, N.; and Ghodsi, A. 2024. The Shift from Models to Compound AI Systems. https://perma.cc/4CCT-CURN.

Zoph, B.; Raffel, C.; Schuurmans, D.; Yogatama, D.; Zhou, D.; Metzler, D.; Chi, E. H.; Wei, J.; Dean, J.; Fedus, L. B.; Bosma, M. P.; Vinyals, O.; Liang, P.; Borgeaud, S.; Hashimoto, T. B.; and Tay, Y. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.