Open Problems in AI Incident Governance

Harleen Kaur Sidhu [1] **Rebecca Scholefield** [1] **Nour Annan** [2] **Kevin Hernandez** [3] **Isabel Nieh Hou** [4] **Abdulrahman Alshaikhi** [4] **Ze Shen Chin** [5 6] **Rokas Gipiškis** [5 7]

Abstract

AI systems may produce failures after deployment that pre-deployment safety assessments do not anticipate. Managing these failures requires what we refer to as adequate *AI incident governance*, where having good definitions, taxonomies, monitoring practices, reporting mechanisms, and incident analysis is essential. We examine existing frameworks related to AI incident governance by regulatory bodies and independent efforts, and find that while there are frameworks that describe how individual functions can be performed, there is a lack of consistency within the aspects of definitions, classification, monitoring, and reporting. These inconsistencies apply to the types of incident data that is collected and reported, the ways in which they are categorised, and as a result, the depth, representativeness, and accuracy of analysis that can be performed. We identify open problems at each stage of the incident governance pipeline, and find that the absence of standardised monitoring and reporting requirements constitutes a significant gap. To address this, we propose a set of principles supported by concrete monitoring guidelines and a reporting template to facilitate their implementation.

1. Introduction

Safety assessments conducted prior to the deployment of an artificial intelligence (AI) system, such as model evaluations and red-teaming, allow testing against anticipated failure modes under controlled settings (Shevlane et al., 2023). However, real-world deployment may produce failures that these assessments cannot anticipate, for example, due to emergent behaviours, adversarial attacks, and unanticipated use cases (O'Brien et al., 2023; Shao et al., 2025).

Collecting, reporting, and analysing information about these incidents enables the identification of causal factors, improves accountability, and mitigates future risks of recurrence. Therefore, effective incident governance plays an important role in improving safety and reducing harm caused by AI systems (Wei & Heim, 2026). Despite this, we found that there is a lack of consistency across the AI incident governance ecosystem, leading to differences in how incidents are defined, categorised, monitored, reported, and analysed. This limits the comparability of individual incidents and, as a result, reduces the effectiveness of analysis and learning across the field.

In this paper, we analyse how definitions, taxonomies, monitoring practices, reporting mechanisms, and incident analysis play a role in AI incident governance. Section 2 examines existing definitions of AI incidents. Section 3 reviews the taxonomies through which incidents are classified. Sections 4 and 5 address monitoring and reporting respectively, where we find the central challenges: the lack of robust monitoring procedures and reporting templates. To address these gaps, we survey corporate monitoring policies (Appendix A), propose monitoring and reporting principles (Appendices B and C), operationalise them as monitoring guidelines (Appendix D), and propose a reporting template (Appendix E). Finally, Section 6 considers how incident data can support meaningful analysis.

2. Definitions

The definition of an AI incident has an effect on what gets monitored, reported, classified, and investigated. For instance, incident repositories based on different definitions may capture different types of events, and therefore different information about them. With the rise of AI regulation worldwide, the definition further determines the obligations and liabilities of actors across the AI value chain. "AI incident" is a relatively new term; as recently as 2024 the Organisation for Economic Co-operation and Development (OECD) described it as an "emerging term" (OECD, 2024). Defining what counts as an AI incident has to take into account fundamental questions of semantics, values, and law (Paeth et al., 2024). We first describe the key definitions of

[1]Independent [2]Sorbonne University [3]Rice University [4]Columbia University [5]AI Standards Lab [6]Oxford Martin AI Governance Initiative [7]Vilnius University. Correspondence to: Rokas Gipiškis <rokas@aistandardslab.org>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

AI incidents in the literature and regulations, highlighting their differences and overlaps, and then identify the open research problems that defining AI incidents still poses.

In the development of the AI Incident Database (AIID), McGregor (2021a) frames an AI incident as a situation in which AI systems caused, or very nearly caused, real-world harm. This broad formulation allows for a wide inclusion of incidents. The OECD offers a more delimited definition of an 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, labour and intellectual property rights; (d) harm to property, communities or the environment" (OECD, 2024). The same definition is used in OECD's common framework for AI incident reporting, which is intended to serve as a global benchmark for stakeholders across jurisdictions and sectors (OECD, 2025).

Current repositories largely inherit these definitions. The AIID uses a refined version of the definition by McGregor (2021a) by more clearly specifying the scope (affected entities): "an alleged harm or near harm event to people, property, or the environment where an AI system is implicated" (AI Incident Database). The OECD's AI Incidents Monitor (AIM) adopts the OECD definition. Working toward a mandatory reporting regime, Dixon & Frase (2025) also builds on the OECD definition and proposes a set of standardised components for reporting templates: the type of incident, the nature and severity of harm, technical data, the affected entities and individuals, and the surrounding context. In the report, Dixon & Frase (2025) treats "AI incident" as covering both incidents and near misses. The AI, Algorithmic and Automation Incidents and Controversies (AIAAIC) Repository also includes events that can potentially cause harm. It defines AI incident as "a sudden known or unknown event that becomes public, takes the form of a disruption, loss, emergency, or crisis, and causes or potentially causes harm" (AIAAIC, 2025).

Recent regulation tends to define narrower, harm-related categories rather than the general term. The EU AI Act defines a "serious incident" as "an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following: (a) the death of a person, or serious harm to a person's health; (b) a serious and irreversible disruption of the management or operation of critical infrastructure; (c) the infringement of obligations under Union law intended to protect fundamental rights; (d) serious harm to property or the environment" (European Parliament & Council of the European Union, 2024). The definition is primarily based on realised harms and does not incorporate near misses, similar to OECD (2024). The defined term is the operative trigger for the reporting obligations in Article 73, whose notification deadlines scale with the sub-category of harm (2 days for a serious and irreversible disruption of critical infrastructure, 10 days in the event of death, and 15 days otherwise). California's SB 53 defines a "critical safety incident" as any of the following: "(1) Unauthorised access to, modification of, or exfiltration of, the model weights of a frontier model that results in death or bodily injury. (2) Harm resulting from the materialisation of a catastrophic risk. (3) Loss of control of a frontier model causing death or bodily injury. (4) A frontier model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behaviour and in a manner that demonstrates materially increased catastrophic risk" (California State Legislature, 2025). This definition includes events that increase risk, such as deceptive evasion, that need not be realised harms. As per SB 53, a frontier developer must report a qualifying critical safety incident to the California Office of Emergency Services within 15 days, or to an appropriate authority within 24 hours where it poses an imminent risk of death or serious physical injury, so a clause such as deceptive evasion determines whether a pre-harm event is reportable at all. New York's Responsible AI Safety and Education Act (RAISE) Act, in its final chapter-amended form (New York State Assembly, 2026), adopts California's "critical safety incident" and its definition, departing from an earlier version that tied safety incidents to critical harm thresholds (death or serious injury of 100 or more people, or one billion dollars in damage). None of these regulations defines the general term "incident".

The definition of an AI incident should serve as a clarification for several aspects, including potential versus actual harm, the scope and severity of harm, and the nature of the event. However, an important difference remains between two groups of AI incident definitions. Most policy-oriented definitions, including those of the OECD and the EU AI Act, define incidents in relation to realised harm; while other broader definitions also include events that have the potential of causing harm even if harm is not realised.

Open Problems

1. The scope of AI incidents is inconsistent across definitions — some include potential but unrealised harm, while others exclude it. Should definitions be standardised? Where does a realised incident end and a near miss begin, and should events with potential but unrealised harm count as incidents?
2. How to validate the quality, usefulness and the intended impact of emerging definitions?
3. When does a single AI incident begin and end, and when the same failure recurs, should repeat occurrences count as one incident or many? When the harm from one model is spread across many people, or across many separate deployments of that model, how should it be counted as an incident?

3. Taxonomies

Taxonomies provide the categories and relationships through which incidents are classified and analysed. Their design determines the extent to which incident records can be meaningfully aggregated across systems, jurisdictions, and time periods. This section surveys the taxonomies that are most closely related to AI incidents, organised by what they classify, namely the causes of failure and the resulting harms, together with their use in incident repositories, before turning to the open problems that taxonomy design still poses.

The Goals, Methods, and Failures (GMF) (Pittaras & McGregor, 2023) taxonomy presents a classification system based on high-level AI system goals (e.g. face recognition), methods and technologies used for system implementation (e.g. transformer neural network), and technical failure causes that result in misbehaviour in the applied system (e.g. distributional bias). Each annotation is paired with a confidence modifier ("known" or "potential") to register how firmly a label can be assigned. There are also other taxonomies that focus on causal factors. Perhaps the most prominent example is the MIT AI Risk Repository's causal taxonomy (entity, intentionality, and timing) (Slattery et al., 2026), itself descended from the taxonomy of pathways to dangerous AI (Yampolskiy, 2016). While these were conceived to classify AI risks and pathways, the former is also used to categorise incidents captured in the AI Incident Tracker (MIT AI Risk Repository, 2026).

A second family of taxonomies classifies incidents by the harm they produce. The Center for Security and Emerging Technology (CSET)'s AI Harm Framework (Hoffmann & Frase, 2023) separates tangible from intangible harm, requiring tangible harm to involve observable injury, loss, or damage, distinguishes harms that have occurred from those that may yet occur, and sorts them into categories such as harm to physical health or safety, financial loss, and human-rights violations. It is used in the AI Incident Database as the "CSETv1" taxonomy (AIID, 2026b). The MIT AI Risk Repository's domain taxonomy (Slattery et al., 2026) is likewise harm-oriented, sorting cases into seven domains and 24 subdomains and extending the language-model risk taxonomy of Weidinger et al. (2022). The former is likewise used to categorise incidents captured in the AI Incident Tracker (MIT AI Risk Repository, 2026).

Repositories utilise multiple taxonomies. The AIID does not designate a single taxonomy, instead hosting several: the CSET harm taxonomy, GMF, and the MIT taxonomies. This is based on the assumption that reasonable parties may classify the same incident differently (McGregor, 2021b). In practice none has been applied across the full database.

The key difference in these taxonomies is between classifying why an incident occurred (GMF; the MIT causal taxonomy) and what harm it caused (CSET; the MIT domain taxonomy). Because the available categories overlap and serve different analytical ends, even reasonable parties may classify the same incident differently (McGregor, 2021b; Agarwal & Nene, 2024). More importantly, the resulting plurality differs in kind from the divergence seen among definitions. Inconsistent definitions across jurisdictions create genuine friction, because a definition fixes obligations and must ultimately be reconciled. Multiple taxonomies, by contrast, may be useful as they address different aspects of incidents. They answer different questions and a repository can effectively use several at once (AIID, 2026a).

Open Problems

1. Can incident taxonomies be made mutually exclusive and collectively exhaustive without becoming too coarse to be informative?
2. How can a category set be chosen so that it stays meaningful as the technology changes (e.g. a method- or technology-based taxonomy designed before the transformer era would now classify most incidents as "other") without continual revision that breaks comparability over time?
3. Can comprehensive AI risk taxonomies not designed for incidents be adapted to incident classification, and how should the mismatch between forward-looking risk categories and records of realised incidents be resolved?

4. Monitoring

Monitoring processes enable the detection of when an incident has occurred and facilitate incident response analysis. This section examines what incident monitoring entails, the actors responsible for monitoring, and the core functions that monitoring performs.

4.1. Scope

We refer to incident monitoring as the systematic collection of operational and contextual data that enables organisations

to (a) detect that an incident has occurred, and (b) provide technical and organisational context needed to understand its root causes (Stein et al., 2024; OECD; O'Brien et al., 2023).

The term *monitoring* is used across AI literature to describe a wide range of activities across the AI lifecycle, including practices such as tracking performance metrics during training and observing model behaviour during red-teaming and pre-deployment testing (Yampolskiy, 2025). However, these pre-deployment modes of monitoring are out of scope of this section, which focuses exclusively on capturing information relating to AI incidents *after* an AI system has been deployed into production environments.

4.2. Actors

Across major regulations, international standards, and governance frameworks, primary responsibility for AI monitoring is typically allocated to the organisations that develop, provide, or deploy AI systems (European Commission, 2025a; ISO/IEC, 2023a; NIST, 2023). These actors conduct system-level monitoring of failures and anomalies based on the operational data that they have access to. However, effective monitoring also requires a distributed ecosystem of external inputs, including users, third-party auditors, researchers, and centralised incident-tracking bodies (Stein et al., 2024; Jones et al., 2023; NIST, 2023).

Incident monitoring extends beyond operational surveillance of individual AI systems to tracking incidents across the AI ecosystem. Public AI incident databases and regulatory market-surveillance bodies primarily engage in this form of cross-industry monitoring of media reports and public disclosures. These organisations provide additional avenues for incident report submissions, thereby surfacing incidents that may not be visible through provider-led monitoring, and help ensure accountability, even though they do not replace the need for robust operational monitoring by AI providers and deployers (Rodrigues et al., 2023).

4.3. Functions

Detection. Detection encompasses the identification and analysis of signals that may indicate that an incident has occurred (Pascoe et al., 2024). For AI systems, incident detection relies on the continuous monitoring of digital infrastructure to identify signals indicative of harm, such as anomalies and threshold-crossing events. Effective detection encompasses initial triage, where false positives are distinguished from genuine incidents, and the assignment of severity levels to route alerts to appropriate incident response processes (Yampolskiy, 2025).

The AI ecosystem has yet to develop standardised classification systems for AI incidents, which may limit the comparability of triage processes across providers, sectors, and jurisdictions. Furthermore, mature high-risk domains systematically incorporate near-miss surveillance as a formal component of detection infrastructure (Gnoni & Saleh, 2017). No equivalent widely adopted near-miss surveillance system currently exists for AI incidents. This may prevent the systemic collection of data that enables the identification of hazardous conditions or behaviours.

Some AI incident signals are distributed across multiple organisational boundaries, including partner companies, cloud providers, downstream developers, and applications that integrate model outputs (O'Brien et al., 2023). As no single actor has visibility into all potential indicators of harm, comprehensive incident detection requires a multi-actor monitoring ecosystem.

Providers of AI systems employ multi-layered automated detection. At inference time, general-purpose filters and real-time automated safety classifiers screen inputs and outputs of AI systems for policy-violating or harmful content. Flagged interactions undergo asynchronous monitoring by classifiers that apply more computationally intensive analysis with greater latency tolerance. Offline safety monitors also review aggregated logs on a periodic basis, enabling the identification of failure modes that real-time classifiers may not be able to detect (Phuong et al., 2026; Williams et al., 2026; Stickland et al., 2025).

These automated incident detection mechanisms enable the rapid identification of incident signals that would be infeasible to detect manually at scale. However, the increasing complexity of AI systems and the volume of monitoring signals generated by these detection systems introduce operational challenges. High alert volumes require human operators to regularly triage and validate alerts, creating significant overhead. As AI systems and their corresponding monitoring mechanisms scale, maintaining a balance between sensitivity to incidents and the burden imposed by frequent alerts may prove increasingly difficult.

Additionally, while automated detection based on system-level operational data is necessary, it is not sufficient for the reliable detection of all incidents that occur. Empirical analysis of production incidents in generative AI cloud services found that 38.3% of incidents were reported by humans rather than automated monitors, reflecting systemic gaps in automated coverage (Yan et al., 2025). Some incidents, especially those with diffused effects, may not be detectable by providers and deployers of AI systems. Detecting these incidents is entirely reliant on incident reports made by affected individuals. These incident reports (which differ from the formal incident reporting described in Section 5) range from bug reports made directly to deployers, to police or news reports, which may then be picked up by media sources. Reports of AI-related harm are monitored by

public AI incident databases such as OECD's AIM detect incidents through monitoring reputable media sources for reports of AI-related harm (OECD). The AIAAIC Repository (AIAAIC, 2024) and the AIID also accept public submission of incident reports (AI Incident Database). The AIID uses a submission leaderboard to increase coverage of AI incidents by encouraging users to submit more reports (McGregor, 2021a).

These continuous and diverse monitoring processes help reduce the time during which incidents remain undetected, limiting the opportunity for harms to propagate. Furthermore, improved detection by public AI incident databases enhances the empirical foundations for forecasting and learning, while simultaneously creating external pressure on providers to maintain high monitoring and accountability standards (McGregor, 2021a; Rodrigues et al., 2023).

Preservation of contextual information. Understanding the causes of an incident requires monitoring that preserves both the technical and organisational context in which the system operated (Dixon & Frase, 2025). Technical context includes information about how users interacted with the system as well as actions taken by agents, and changes in task specifications or permissions (Anderljung et al., 2023; Ezell et al., 2025; European Commission, 2025c). These logs may provide insights into how the system's behaviour diverged from its intended operation. The organisational context includes post-deployment modifications in design decisions, documented safeguards, risk-assumption registers, oversight structures, and approval or escalation pathways (NIST, 2023). These records may allow investigators to determine if governance choices or procedural gaps may have contributed to the incident. Together, this data helps to trace the sequence of decisions, interactions, and events that led to the occurrence of an incident (European Parliament & Council of the European Union, 2024).

To ensure that all relevant contextual evidence is captured and retained, it may be necessary to employ continuous, real-time monitoring mechanisms and processes to track and preserve applicable logs and documents (European Commission, 2025c; Ferdaus et al., 2026). However, some information, such as context windows provided by users, may not be sufficiently monitorable by deployers (Paeth et al., 2024). Preserving sufficient contextual information also poses unique challenges for autonomous and distributed systems, such as multi-agent architectures. Existing memory systems for agentic AI may not sufficiently support traceability and incident investigation (Wang et al., 2026). Finally, monitoring that involves logging user data raises concerns about privacy (Stein et al., 2024). Even when organisations employ pseudonymization or aggregation methods to protect sensitive data, research on de-anonymisation attacks demonstrates that sensitive information can still be recovered by malicious actors (Xin et al., 2025; Feretzakis & Verykios, 2024; Lange et al., 2025).

Open Problems

1. Should the triaging processes be standardised across the AI ecosystem, and if so, how should it be standardised?
2. How should different actors share responsibility for AI incident monitoring when incident signals are so fragmented across the AI ecosystem?
3. How can automated detection systems be kept up to date as models and threat vectors change rapidly over time, and how should they be calibrated to manage trade-offs between detection sensitivity and alert fatigue?
4. How can monitoring systems balance the need for comprehensive user data with privacy protection?

5. Reporting

The EU AI Act establishes incident reporting requirements (European Parliament & Council of the European Union, 2024), and the European Commission has published draft report templates to support implementation (European Commission, 2025b). This section compares its approach with frameworks from the OECD (OECD, 2025) and CSET (Dixon & Frase, 2025; 2024).

5.1. Scope

There is debate about the appropriate scope of mandatory reporting. The EU AI Act requires reporting only serious incidents, where AI systems cause harm, including "the death of a person or serious damage to a person's health, to property or the environment" (European Parliament & Council of the European Union, 2024). Others argue for mandatory reporting of all incidents, regardless of severity, and near misses (Dixon & Frase, 2025).

Increasing the scope of mandatory reporting may increase the burden on reporters and receiving authorities, particularly if reports require manual review. However, some argue that reporting all incidents strengthens oversight of emerging risks, including "systemic" harms that appear insignificant in isolation but become severe in aggregate (Paeth et al., 2024).

5.2. Actors

As discussed in Section 2, certain regulations place obligations on incident reporting to public authorities onto those who develop, provide, or deploy AI systems. Separately, some have also argued for supporting members of the public to report incidents voluntarily, including private individuals who experience or observe them, as well as stakeholders such as researchers, journalists, and watchdogs (Dixon & Frase, 2024). However, there is limited work on how to sup-

port voluntary reporting—for example, how barriers posed by limited technical knowledge can be overcome.

5.3. Common elements of incident reports

Timelines. Each framework asks about the timeline of an incident, though details vary. The European Commission distinguishes between start date, end date, and date of detection (European Commission, 2025b). However, these events are not formally defined: "end date" could be interpreted as the date on which a system was fixed, or the date on which harms ceased or were remediated. It is also unclear how precise dates should be—for instance, whether a month and year suffice if exact dates are unknown. A related question is how reports should capture uncertainty about dates. One approach is to require reporters to indicate whether dates are known, estimated, or unknown (Pittaras & McGregor, 2023).

Implicated systems. Frameworks vary in the technical and operational information requested for AI systems that contributed to an incident. The European Commission focuses on identifying the specific product via its EU database ID, serial or batch number, and software and firmware versions (European Commission, 2025b). CSET proposes more detailed technical documentation, including system cards, model cards, and datasheets (Dixon & Frase, 2025). The OECD places greater emphasis on deployment context, including usage context and rights, a system's autonomy, and whether incidents involve multiple interacting systems (OECD, 2025).

This variation reflects a lack of consensus about the role of incident reports (Stein et al., 2024). Reports could aim to provide enough information for causal analysis, or they could simply identify incidents warranting further investigation. Reports may also enable pattern identification—for example, by flagging whether models lacking certain safeguards, designed for particular use cases, or deployed in certain domains are disproportionately implicated in harm. Each approach calls for different kinds of information. A further question is how information requirements should differ across multiple interacting systems, and how reports can capture their contributions.

Causality. Proposed frameworks differ substantially in how they approach causality. Because the European Commission frames reporting as an iterative process, it requires relatively little causal information upfront, asking "What went wrong with the system?" and "What is the likely cause?" and deferring definitive root causes to final reports (European Commission, 2025b). The OECD and CSET frameworks are more analytical, attempting to capture causal information through comprehensive categories from the outset. For example, the OECD combines questions about the nature of causality—direct cause or contributing factor—with more specific categories such as overreliance, intentional misuse, and human error, and allows reporters to expand on failures at the data, model, and system levels (OECD, 2025).

Where reporting is iterative, an open question is how much causal information initial reports should seek. Open-ended questions with insufficient guidance may produce uninformative reports, unless followed by formal investigation. A related question is how to balance the flexibility of open-text fields with the comprehensiveness of structured categories.

Impact. Across frameworks, there is a consistent focus on core harm types—including physical, environmental, and financial harms—and each framework attempts to quantify their severity. Frameworks differ mainly in how many dimensions they propose: CSET, for instance, adds remediability, optionality, and frequency (Dixon & Frase, 2025). In terms of who is impacted, the European Commission focuses narrowly on users, while the OECD covers a broader range of stakeholders, asking reporters to specify whether groups such as "children", "trade unions", or "businesses" were impacted (OECD, 2025).

An open question is what should be mandatory: some fields—such as the number of users harmed or the remediability of an incident—may be difficult to assess at the point of reporting and may be better treated as optional or follow-up submissions.

Open Problems

1. What purposes should incident reports serve, and what information is required as a result?
2. What should be the scope of mandatory incident reporting, and how should voluntary reporting be supported outside this scope?
3. How should information requirements vary across stages of an iterative reporting process?
4. How should reports balance structured categories with open-ended reporting?

6. Incident analysis

Establishing how AI incidents are defined and taxonomised standardises the types of events that are recognised as incidents and the ways in which they are categorised. The need for incidents to be meaningfully analysed and understood, in turn, determines what needs to be monitored and reported. However, the mere collection of AI incident information does not, in itself, enable learning and reduce risk (Turri & Dzombak, 2023). The main goal of incident analysis is to prevent future incidents from occurring, including through improved organisational practices. In this section, we describe single-incident causal analysis and aggregate cross-incident analysis, and then identify the open research problems in AI incident analysis.

Given one event, single-incident causal analysis reconstructs what happened and why. It involves identifying the technical, organisational, and contextual factors that lead to AI incidents. Mylius (2024) demonstrates fault-tree analysis of AI safety incidents, using language models to infer candidate causes from incident reports. Ezell et al. (2025) introduce a causal framework for AI-agent incidents that, drawing on human-factors methods, traces failures through system factors (e.g. training data and learning methods), contextual factors (e.g. prompt injections and tool access), and cognitive errors (e.g. misinterpreting a request), while specifying the activity logs and system information an investigator needs to determine which factors apply.

Given a corpus of incidents, aggregate cross-incident analysis may help in finding patterns, identifying recurring contributors, and understanding emerging risks. Beyond understanding incidents at an individual level, analysis on a population level requires a common taxonomy and reporting system, but can provide insights on broader trends. Analyses of incidents can enable a more comprehensive understanding of historical trends and the current state of the world. Large-scale analyses of public AI incident databases have been used to examine recurring harm categories (May et al., 2024; Velázquez et al., 2024), identify gaps between developer risk assumptions and real-world impacts (Rao et al., 2025), and analyse patterns in accountability and response following AI incidents (Richards et al., 2025).

Additionally, extrapolation of historical trends can also lead to insights into future trends. Forecasting extends pattern identification toward estimating the frequency of future incidents. Incident analysis remains nascent in AI safety, with existing work primarily centred around generalising and causal analysis. However, incident forecasting is an established practice in other safety-critical domains (Zheng & Liu, 2009). It has applications in epidemiology (Desai et al., 2019; Rilkoff et al., 2024; Centers for Disease Control and Prevention, 2026), aviation (Zhang & Mahadevan, 2019), cybersecurity (Almahmoud et al., 2023), and the maritime domain (Kandel & Baroud, 2024), where it can be used to identify frequent incident types and predict future incident occurrences for them.

Open Problems

1. How can incident datasets balance high-dimensional categorical fields with free-text descriptions to enable both rich contextual detail and systematic analysis at scale?
2. How can forecasting approaches identify indicators of emerging risks, particularly for novel failure modes?
3. What forecasting methodologies from other safety-critical domains can be adapted to AI incident governance given the specific characteristics of AI systems?

7. Conclusion

AI incident governance is central to ensuring a safe and secure AI ecosystem, yet existing frameworks lack consistency across key aspects of definitions, taxonomies, monitoring and reporting practices, and incident analyses. As AI is increasingly deployed in high-risk contexts and the risk landscape evolves, robust incident governance processes are key. Effective AI incident governance depends on definitions that clearly establish scope and responsibilities, comprehensive taxonomies enabling causal analyses and categorization of harm, and monitoring and reporting practices that capture and document sufficient data for incident analysis.

This paper has identified key open problems at each stage of the incident governance pipeline, most notably the lack of robust monitoring and reporting practices. The monitoring and reporting principles, monitoring guidelines, and reporting template proposed in the appendix represent an initial response to the most pressing operational gaps. Further research is also needed on topics such as forecasting methodologies, and the harmonisation of reporting obligations across jurisdictions.

Impact Statement

This paper contributes to ongoing efforts to make AI deployment more accountable by surfacing the gaps between how incidents are defined, monitored, and reported. The proposed monitoring guidelines and reporting template are intended as starting points for harmonisation across frameworks, not as final specifications.

Use of Large Language Models

During the preparation of this work, large language models (Claude, Gemini) were used to conduct preliminary literature exploration and to edit and review drafts. All AI-generated text and suggested sources were manually reviewed, verified, and edited by the authors, who take full responsibility for the final content of the paper.

Acknowledgements

The authors thank the SPAR (Supervised Program for Alignment Research) AI safety program for supporting this work. We also thank Marcel Mir Teijeiro, Koen Holtman, Sean McGregor, George Gor, and Adrian Regenfuß for their comments on earlier versions of this draft. Any remaining errors are our own.

References

Agarwal, A. and Nene, M. J. Standardised schema and taxonomy for AI incident databases in critical digital

infrastructure. In *2024 IEEE Pune Section International Conference (PuneCon)*, pp. 1–6. IEEE, 2024.

AI Incident Database. Editors' guide. https://incidentdatabase.ai/editors-guide/. Accessed: 2026-02-14.

AIAAIC. AIAAIC Repository user guide. https://www.aiaaic.org/aiaaic-repository/user-guide, May 2024. [Accessed 06-11-2025].

AIAAIC. Classifications and Definitions. AIAAIC Repository, 2025. URL https://www.aiaaic.org/aiaaic-repository/classifications-and-definitions. Accessed June 28, 2026.

AIID. List of Taxonomies, 2026a. URL https://incidentdatabase.ai/taxonomies/. Accessed June 28, 2026.

AIID. CSETv1 Charts, 2026b. URL https://incidentdatabase.ai/taxonomies/csetv1/. Accessed June 28, 2026.

Almahmoud, Z., Yoo, P. D., Alhussein, O., Farhat, I., and Damiani, E. A holistic and proactive approach to forecasting cyber threats. *Scientific Reports*, 13(1):8049, 2023.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., et al. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

Bluemke, E., Collins, T., Garfinkel, B., and Trask, A. Exploring the relevance of data Privacy-Enhancing technologies for AI governance use cases. *arXiv preprint arXiv:2303.08956*, 2023.

Buhl, M. D., Sett, G., Koessler, L., Schuett, J., and Anderljung, M. Safety cases for frontier AI. *arXiv preprint arXiv:2410.21572*, 2024.

California State Legislature. Senate Bill No. 53, 2025. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53. Accessed June 28, 2026.

Centers for Disease Control and Prevention. Behind the model: Nowcasting. https://www.cdc.gov/cfa-behind-the-model/php/data-research/nowcasting.html, 2026. Accessed: 2026-04-25.

Desai, A. N., Kraemer, M. U., Bhatia, S., Cori, A., Nouvellet, P., Herringer, M., Cohn, E. L., Carrion, M., Brownstein, J. S., Madoff, L. C., et al. Real-time epidemic forecasting: challenges and opportunities. *Health security*, 17(4):268–275, 2019.

Dixon, R. B. L. and Frase, H. An argument for hybrid AI incident reporting. *Center for Security and Emerging Technology. Retrieved from https://cset.georgetown. edu/publication/an-argument-for-hybrid-ai-incident-reporting/(accessed 2024-06-04)*, 2024.

Dixon, R. B. L. and Frase, H. AI incidents: Key components for a mandatory reporting regime. *Georgetown Center for Security and Emerging Technology*, 2025.

European Commission. Incident report for serious incidents under the AI Act (High-risk AI systems). https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks, September 2025a. [Accessed 06-11-2025].

European Commission. AI Act: Commission publishes a reporting template for serious incidents involving general-purpose AI models with systemic risk. https://digital-strategy.ec.europa.eu/en/library/ai-act-commission-publishes-reporting-template-serious-incidents-involving-general-purpose-ai, 2025b.

European Commission. The General-Purpose AI Code of Practice. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai, 2025c. Accessed: 2026-02-14.

European Parliament and Council of the 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... https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Official Journal of the European Union, OJ L 2024/1689, 12 July 2024. Text with EEA relevance.

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

Ferdaus, M. M., Abdelguerfi, M., Loup, E., N. Niles, K., Pathak, K., and Sloan, S. Towards trustworthy AI: A review of ethical and robust large language models. *ACM Computing Surveys*, 58(7):1–43, 2026.

Feretzakis, G. and Verykios, V. S. Trustworthy AI: Securing sensitive data in large language models. *AI*, 5(4):2773–2800, 2024.

Gnoni, M. G. and Saleh, J. H. Near-miss management systems and observability-in-depth: Handling safety incidents and accident precursors in light of safety principles. *Safety science*, 91:154–167, 2017.

Hoffmann, M. and Frase, H. Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework. Technical report, CSET (CSET), Georgetown University, July 2023. URL https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/.

ISO/IEC. ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, 2022a.

ISO/IEC. ISO/IEC 23054:2022 framework for artificial intelligence (ai) systems using machine learning (ml), 2022b.

ISO/IEC. ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. https://www.iso.org/standard/81230.html, 2023a. ISO/IEC 42001:2023 Artificial Intelligence Management System (AIMS) standard.

ISO/IEC. ISO/IEC 23894:2023 information technology — artificial intelligence — guidance on risk management, 2023b.

ISO/IEC. ISO/IEC 27035-1:2023information technology — information security incident management, 2023c.

Johnson, O. B., Olamijuwon, J., Cadet, E., Osundare, O. S., and Weldegeorgise, Y. W. Developing real-time monitoring models to enhance operational support and improve incident response times. *Int J Eng Res Dev*, 20(11):1296–1304, 2024.

Jones, E., Birtwistle, M., and Reid, O. F. Keeping an eye on AI: Approaches to government monitoring of the AI landscape, July 2023. URL https://www.adalovelaceinstitute.org/report/keeping-an-eye-on-ai/. Accessed: 2026-02-14.

Kandel, R. and Baroud, H. A data-driven risk assessment of arctic maritime incidents: Using machine learning to predict incident types and identify risk factors. *Reliability Engineering & System Safety*, 243:109779, 2024.

Lange, L., Schreieder, T., Christen, V., and Rahm, E. Slice it up: Unmasking user identities in smartwatch health data. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, pp. 710–726, 2025.

May, R., Krüger, J., and Leich, T. Sok: How artificial-intelligence incidents can jeopardize safety and security. In *Proceedings of the 19th international conference on availability, reliability and security*, pp. 1–12, 2024.

McGregor, S. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15458–15463, 2021a.

McGregor, S. The First Taxonomy of AI Incidents. AIID (blog), July 2021b. URL https://incidentdatabase.ai/blog/the-first-taxonomy-of-ai-incidents/. Posted July 8, 2021. Accessed June 28, 2026.

MIT AI Risk Repository. AI Incident Tracker. MIT AI Risk Initiative, MIT FutureTech, 2026. URL https://airisk.mit.edu/ai-incident-tracker. Accessed June 28, 2026.

Mylius, S. AI Harm Severity Scales by Category. https://simonmylius.com/ai-harm-severity-scales. Accessed: 2026-02-14.

Mylius, S. Root cause analysis of AI safety incidents, June 2024. URL https://simonmylius.com/blog/6wj3yx02hivp2vbl1bz9dqmxmelc35. Blog post.

New York State Assembly. Assembly Bill A09449, 2026. URL https://nyassembly.gov/leg/?bn=A09449&term=2025&Text=Y. Accessed June 28, 2026.

NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, U.S. Department of Commerce, 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

O'Brien, J., Ee, S., and Williams, Z. Deployment corrections: An incident response framework for frontier AI models. *arXiv preprint arXiv:2310.00328*, 2023.

OECD. Overview and methodology of the AI incidents and hazards monitor methodology and disclosures. https://oecd.ai/en/incidents-methodology. Accessed: 2026-02-14.

OECD. Defining AI Incidents and Related Terms, 2024.

OECD. Towards a Common Reporting Framework for AI Incidents: The 29 Criteria, 2025. URL https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/towards-a-common-reporting-framework-for-ai-incidents_8c488fdb/f326d4ac-en.pdf.

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

Pascoe, C., Quinn, S., and Scarfone, K. The NIST cybersecurity framework (CSF) 2.0. 2024.

Phuong, M., Jenner, E., Simon, L., Ho, L., Shah, R., Farquhar, S., and Coull, S. GDM AI Control Roadmap. 2026.

Pittaras, N. and McGregor, S. A taxonomic system for failure cause analysis of open source AI incidents. In *Proceedings of the SafeAI 2023 Workshop*, 2023. URL https://ceur-ws.org/Vol-3381/17.pdf.

Rao, P. S., Šćepanović, S., Jayagopi, D. B., Cherubini, M., and Quercia, D. The AI Model Risk Catalog: What Developers and Researchers Miss About Real-World AI Harms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2150–2163, 2025.

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

Rilkoff, H., Struck, S., Ziegler, C., Faye, L., Paquette, D., and Buckeridge, D. Innovations in public health surveillance: An overview of novel use of data and analytic methods. *Canada Communicable Disease Report*, 50 (3-4):93, 2024.

Rodrigues, R., Resseguier, A., and Santiago, N. When artificial intelligence fails: The emerging role of incident databases. *Pub. Governance, Admin. & Fin. L. Rev.*, 8:17, 2023.

Scheuerman, M. K., Jiang, J. A., Fiesler, C., and Brubaker, J. R. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–33, 2021.

Shahriar, S., Allana, S., Hazratifard, S. M., and Dara, R. A survey of privacy risks and mitigation strategies in the artificial intelligence life cycle. *IEEE Access*, 11:61829–61854, 2023.

Shao, S., Ren, Q., Qian, C., Wei, B., Guo, D., Yang, J., Song, X., Zhang, L., Zhang, W., Liu, D., et al. Your agent may misevolve: Emergent risks in self-evolving llm agents. *arXiv preprint arXiv:2509.26354*, 2025.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Slattery, P., Saeri, A. K., Grundy, E. A., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, 2026.

Stein, M., Bernardi, J., and Dunlop, C. The role of governments in increasing interconnected post-deployment monitoring of AI. *arXiv preprint arXiv:2410.04931*, 2024.

Stickland, A. C., Michelfeit, J., Mani, A., Griffin, C., Matthews, O., Korbak, T., Inglis, R., Makins, O., and Cooney, A. Async Control: Stress-testing Asynchronous Control Measures for LLM Agents. *arXiv preprint arXiv:2512.13526*, 2025.

Szadeczky, T. and Bederna, Z. Risk, regulation, and governance: evaluating artificial intelligence across diverse application scenarios: Risk, regulation, and governance. *Security Journal*, 38(1):35, 2025.

Tariq, S., Baruwal Chhetri, M., Nepal, S., and Paris, C. Alert fatigue in security operations centres: Research challenges and opportunities. *ACM Computing Surveys*, 57(9):1–38, 2025.

Turri, V. and Dzombak, R. 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*, pp. 576–583, 2023.

Vei, S., Giudici, P., Sermpezis, P., Vakali, A., and Bernardelli, A. E. AI Harmonics: a human-centric and harms severity-adaptive AI risk assessment framework. *arXiv preprint arXiv:2509.10104*, 2025.

Velázquez, J. D. M., Šćepanović, S., Gvirtz, A., and Quercia, D. Decoding real-world artificial intelligence incidents. *Computer*, 57(11):71–81, 2024.

Vinnakota, K. and Kolla, M. Creating effective alerts for monitoring distributed systems. *International Journal of Computer Trends and Technology*, 73:172–178, 2025.

Wang, Y., Zhang, J., Cai, T., Liu, Z., Sun, Q., Sun, Z., Wu, Z., Dong, M., Zhang, M., Yin, X., and Zhu, Y. From Agent Traces to Trust: A Survey of Evidence Tracing and Execution Provenance in LLM Agents, 2026. URL https://arxiv.org/abs/2606.04990.

Wei, K. and Heim, L. Designing incident reporting systems for harms from general-purpose ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38016–38029, 2026.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 214–229, 2022.

Williams, M., Sun, H., Sekhar, S., Carroll, M., Robinson, D. G., and Kivlichan, I., 2026. URL https://open

ai.com/index/how-we-monitor-internal-coding-agents-misalignment/.

Xin, R., Mireshghallah, N., Li, S. S., Duan, M., Kim, H., Choi, Y., Tsvetkov, Y., Oh, S., and Koh, P. W. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. *arXiv preprint arXiv:2504.21035*, 2025.

Yampolskiy, R. V. Taxonomy of Pathways to Dangerous Artificial Intelligence. In *AAAI Workshop: AI, Ethics, and Society*, pp. 143–148, 2016.

Yampolskiy, R. V. On monitorability of AI. *AI and Ethics*, 5(1):689–707, 2025.

Yan, H., Chen, Y., Ma, M., Wen, M., Lu, S., Zhang, S., Xu, T., Wang, R., Bansal, C., Rajmohan, S., et al. An empirical study of production incidents in generative AI cloud services. In *2025 IEEE 36th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 359–370. IEEE, 2025.

Zhang, X. and Mahadevan, S. Ensemble machine learning models for aviation incident risk prediction. *Decision Support Systems*, 116:48–63, 2019.

Zheng, X. and Liu, M. An overview of accident forecasting methodologies. *Journal of Loss Prevention in the process Industries*, 22(4):484–491, 2009.

A. Corporate Monitoring Policies

Corporate monitoring policies provide insights into how leading AI developers interpret their post-deployment monitoring responsibilities. We analysed publicly available AI safety and security frameworks to gain an understanding of how these organisations publicly articulate their post-deployment monitoring responsibilities. This analysis does not reflect the full scope of their internal practices.

Across the surveyed materials, several common patterns emerge in how organisations frame and operationalise monitoring. User-driven reporting channels and anonymous employee escalation pathways provide avenues for people engaging with the system to report issues, abuse, or incidents. Some companies offer bug bounty programmes, which actively incentivise user reports. Many frameworks also reference logging practices, including prompt and invocation logs, infrastructure-change logs, and classifiers for misuse detection. About half of the policies reviewed indicated a dedicated incident response and investigation framework, including internal oversight teams responsible for reviewing alerts and conducting investigations. Some organisations described red-teaming and post-launch evaluations to identify regressions or new risk patterns.

Our analysis reveals substantial variation in scope, specificity, and operational detail.

Below, we summarise how organisations are currently interpreting and operationalising monitoring obligations.

Table 1. Publicly disclosed post-deployment monitoring practices across frontier AI developers. Derived from each organisation's publicly available AI safety and security framework; does not reflect the full scope of internal practices.

Vendor	**Reporting channels**	**Bug bounties**	**User monitoring**	**Output monitoring**	**Escalation & whistle-blowing**	**Incident response**	**Red teaming**	**Downstream attribution**
Amazon (AWS)	✓	✓	✓	✓	–	✓	✓	–
Anthropic	✓	✓	–	✓	✓	✓	●	–
Cohere	–	✓	✓	●	–	●	✓	–
G42	✓	–	✓	✓	✓	✓	●	–
Google DeepMind	–	✓	✓	●	–	●	●	–
Magic	–	✓	–	●	●	●	●	–
Meta	–	●	–	●	–	–	●	–
Microsoft	✓	✓	✓	✓	●	✓	●	✓
NVIDIA	✓	✓	–	✓	–	✓	✓	✓
OpenAI	✓	✓	✓	●	✓	●	✓	–
xAI	–	●	✓	✓	✓	●	✓	–

✓ explicitly present; – no mention; ● ambiguous or high-level commitment without operational detail.

Vendor	**Monitoring Policies**
Amazon (AWS)	Amazon outlines mechanisms for external vulnerability reporting, internal threat detection, and structured incident response. A dedicated Cyber Threat Intelligence team continually monitors and tracks down advanced threat actor groups. Amazon maintains continued engagement with and investments in external security research, including bug bounty programs, academic research investments, red teaming networks, and coordinated vulnerability disclosure programs that encourage and reward security experts to partner with them for research and development. Amazon employs output moderation systems to ensure that generated content adheres to its Amazon Responsible AI objectives by blocking or safely modifying violating inputs and outputs. Their Security Operations Centres provide 24/7 centralised global support. There are also dedicated Incident Response Protocols for incident escalation and response pathways in the event of reported AI safety incidents.

Continued on next page

Vendor	Monitoring Policies
Anthropic	Anthropic Responsible Scaling Policy lists monitoring processes including responses to jailbreak bounties, conducting historical analysis or background monitoring, and any necessary retention of logs for these activities. Model capabilities are consistently evaluated and a dedicated Responsible Scaling Officer is accountable for the implementation of any interim processes that may be required if capability evaluations exceed safety thresholds. Anthropic has a Responsible Disclosure policy and a safety issue reporting platform where users may email issues to a dedicated user safety email address. They also operate private bug bounty programs through HackerOne, including programs for identifying vulnerabilities in their models and security vulnerabilities. Anthropic employees can report AI safety-related concerns through three main channels, including an anonymous channel for reporting potential violations of their AI safety commitments. The Responsible Scaling Policy states that red-teaming is conducted pre-deployment.
Cohere	Cohere has a Secure AI Frontier Model Framework, where they describe continuous monitoring processes including unexpected post-deployment usage patterns and manual and automatic security risk identification processes. They also have "Secure Product Lifecycle" controls that include security risk assessments, penetration testing, and bug bounty programs. They incentivise third-party vulnerability discovery via clear protections for legitimate research practices in their Responsible Disclosure Policy. Cohere has an incident response plan whereby incidents are identified, tracked, and resolved, although operational details for this plan are not provided. Cohere conducts evaluations and testing for general performance, safety, and security throughout the lifecycle and includes red-teaming, evaluations with academic and industry benchmarks, and internal evaluations.
G42	G42 has a Frontier AI safety Framework, which describes model robustness testing and asynchronous monitoring protocols. Responsibilities of their dedicated Frontier AI Governance Board include developing a comprehensive incident response plan that outlines the steps to be taken in the event of non-compliance. They outline an incident detection strategy that leverages automated mechanisms and human review. Incident reporting channels include designated pathways for users to report instances of concerning or harmful behavior in violation of company policy to relevant G42 Personnel.
Google DeepMind	Google's Frontier Safety Framework details post-market monitoring strategies based on model capabilities, which are evaluated proactively and throughout the entire lifecycle of the model. Deployment mitigations include safety fine-tuning and monitoring and response. If capability evaluations reach a threshold where model reasoning ability is determined to have the potential to undermine human control, additional monitoring processes such as automated chain-of-thought monitoring of the model's reasoning may be implemented. Gemini's API policy documents specify the use of automated and manual mechanisms to detect violations of the API usage policy. An AI Vulnerability Rewards Program functions as a responsible disclosure platform, and Google commits to developing a suite of safeguards which may include safety post-training, monitoring and analysis, account moderation, user verification, and bug bounties.
Magic	Magic has an AGI Readiness Policy where they state that they continuously monitor their AI systems to evaluate if their models have reached dangerous capability frontiers. Output safety classifiers are used for output monitoring, and automated detection is applied for internal usage within Magic. In cases where risks for threat models pass a set threshold, safety measures including delaying or pausing development are specified. These threat mitigations include both security measures such as training and compartmentalisation, as well as deployment mitigations such as harm refusal. Magic has a Vulnerability Disclosure Policy that allows security researchers to submit suspected vulnerabilities via a web form.
Meta	Meta's Frontier AI Framework includes pre-deployment AI risk assessments and threat modelling exercises that identify potential risks associated with frontier AI, enable the development of risk mitigation strategies, and inform model deployment plans.

Continued on next page

Vendor	Monitoring Policies
Microsoft	Microsoft implements post-deployment tools and processes for ongoing monitoring, user feedback channels, incident response, and iterative improvements to its risk mitigation stack. Product teams across Microsoft are required to put in place repeatable processes to collect user feedback and to triage and address issues that arise after the release of an AI system. Microsoft also has a Coordinated Vulnerability Disclosure program, where researchers may disclose issues with AI systems. The program includes a Bug Bounty Program to incentivise users to report significant security issues. At the platform level, safety measures include content classifiers that block potentially harmful user inputs and AI-generated content. Microsoft uses a unified API to detect and block jailbreak patterns in user inputs and indirect prompt injection attacks. A Microsoft Security Response Center researcher portal also accepts reports for potential incidents. These reports may be anonymous. Microsoft details an incident response plan that includes expanding the capacity of specialised roles like crisis managers, forensic investigators, and communications managers. An AI Red Teaming Agent is made available to customers, who may use this to simulate attack techniques and generate red teaming reports that help track risk mitigation improvements throughout the AI development life cycle. To support downstream monitoring of AI tools, services, and components, Microsoft released 'Azure AI Foundry Observability' in May 2025, which offers continuous monitoring and evaluation of both AI applications and agentic AI systems in production.
NVIDIA	NVIDIA's Frontier AI risk Assessment framework outlines strategies for lowering hazard duration, decreasing hazard onset speed. It states that NVIDIA models have watermarks embedded at generation-time that enable downstream detection and attribution of AI-produced outputs, which provides a verifiable origin signal for both end-users and automated scanning tools. NVIDIA also has a Vulnerability Disclosure program where security vulnerabilities can be submitted. To address evolving threats and vulnerabilities, red teaming activities are used in conjunction with public benchmarks.
OpenAI	OpenAI's preparedness framework lists safeguards against harm, including usage monitoring to stop or catch adversarial users, adversarial testing, and red-teaming. A monitor and incident response plan is cited, although operational details are not provided. OpenAI also has a coordinated vulnerability disclosure policy including a bug bounty program and an incident reporting platform.
xAI	xAI uses real-time monitoring, telemetry and alerting of threshold breaches via internal tooling. Models are equipped with the ability to scrutinise user behaviour and identify bad actors. Classifiers are applied to user inputs to verify safety when a model is queried regarding weapons of mass destruction or cyberterrorism. xAI also has a responsible disclosure program that includes bug bounties to encourage external parties to report security issues. Employees may anonymously report concerns about non-adherence to their risk management framework.

B. Monitoring principles

Post-deployment incident monitoring techniques should account for both the informational requirements of the reporting, forecasting, and learning stages that follow, as well as the challenges outlined above. Accordingly, the principles detailed below are designed to optimise these activities:

- **Continuous:** Monitoring should be maintained throughout the system's operational life cycle and proportionate to its risk profile (Szadeczky & Bederna, 2025) with the aim of minimising both the Mean Time to Detection (MTTD) (Johnson et al., 2024) and the likelihood of undetected incidents.
- **Calibrated:** Alert mechanisms should be tuned to maintain a balanced trade-off between false positives and false negatives. Calibration should be periodically reviewed to prevent alert fatigue and ensure that monitoring processes remain accurate and timely, with MTTD targets proportionate to system and context risk (O'Brien et al., 2023; Tariq et al., 2025).
- **Traceable:** Monitoring data should be comprehensive enough to trace the sequence of system states, inputs, outputs, and interactions that led to an incident, enabling independent verification of what occurred and supporting later causal analysis (European Parliament & Council of the European Union, 2024). Traceability should connect technical evidence such as system event logs (Yampolskiy, 2025; Paeth et al., 2024) and logs of user or agent interactions (Ezell et al.,

2025) with organisational records documenting existing safeguards and policies (Dixon & Frase, 2025) and updates to risk and safety case evaluations (Buhl et al., 2024).

- **Impact inclusive:** Once an incident is detected, monitoring should ensure the systematic capture and preservation of observable impact-relevant evidence, including signals from which scope, frequency, duration, and severity can later be derived (Vei et al., 2025). This should include the number of users or systems affected, how long the harm or disruption persisted, and how severe the consequences were across technical, organisational, or social dimensions (Mylius; Hoffmann & Frase, 2023). Impact data may be drawn from telemetry, user feedback, impact assessments, or incident reports.
- **Privacy-preserving defaults:** Monitoring that involves user data requires the informed consent of users (Stein et al., 2024) or the implementation of rigorous anonymisation techniques such as pseudonymisation or aggregation methods (Bluemke et al., 2023). Minimal sensitive data should be processed and cyber threat evaluations should be conducted to identify risks of re-identification or leakage (Shahriar et al., 2023). Information flows, including how sensitive data moves within and between systems, for what purposes, and for how long, should be documented and auditable, ensuring that privacy is maintained through controlled and transparent data movement (Bluemke et al., 2023).

These principles and their associated guidelines (Appendix D) provide a structured foundation for post-deployment incident monitoring, however, they interact in ways that create practical trade-offs. Privacy-preserving data minimisation can limit the granularity of logs required for full traceability, while achieving comprehensive traceability or implementing continuous, proactive signal collection may require collecting data that increases privacy risk or retention burdens. Calibration choices may also affect impact-relevant signals, particularly when it is uncertain which metrics will later prove important for assessing the severity of harm. As a result, optimising each principle independently may compromise overall monitoring effectiveness. Organisations must balance trade-offs between these principles in light of system risk, technical feasibility and legal obligations, recognising that monitoring is a complex system design problem rather than a checklist of isolated requirements.

C. Reporting principles

We propose five principles for AI incident reports:

- **Iterative:** Reporters should identify whether a report is an initial, follow-up or final report. Receiving bodies should assign an ID to initial reports, which reporters can reference in follow-up and final reports to link them. Reports should be automatically timestamped at submission. This reflects the reality that information is uncovered gradually while supporting traceability.
- **Pragmatic:** Initial reports should request information that reporters can reasonably provide: system identification, usage context, core harm types, approximate harm quantification, and suspected causes. Detailed technical information and comprehensive causal analysis can be provided in follow-up reports if necessary.
- **Epistemically transparent:** Initial reports should accommodate uncertainty, requiring reporters to specify whether their inputs are known, estimated or unknown.
- **Unambiguous:** Questions should be worded precisely and structured to help reporters identify relevant information. For example, rather than asking about "likely causality," questions can suggest factors to consider, such as data, model, system, usage context, or governance failures.
- **Analysable:** Reports should be exportable in machine-readable formats so their contents can be annotated using more granular taxonomies, enabling comparison and learning across incidents.

D. Monitoring Guidelines

Principle	Guideline	Operationalisation Details	Guidelines Derived From
M1. Continuous	M1.1 Monitoring should be ongoing and maintained throughout all post-deployment stages of the AI system life cycle.	Monitoring should occur while the AI system is in operation, from deployment to decommissioning. Choice of monitoring modalities (e.g., real-time, periodic, or event-triggered) should be risk-based and system architecture-dependent (see M1.2).	ISO/IEC (2022a) §6.2.6–6.2.7; ISO/IEC (2022b) §8.7
M1. Continuous	M1.2 Continuous monitoring modalities should reflect and be proportional to the risk profile of the system and context of operation.	Risk-based monitoring regimes should be implemented, such that AI systems with higher risk classifications are subject to more frequent, more detailed, and more comprehensive monitoring. Monitoring requirements associated with a system's assigned risk tier and risk treatment plan should be explicitly documented, periodically reviewed, and updated following model updates, changes in deployment context, or the emergence of new risk signals.	NIST (2023) §6.4; European Parliament & Council of the European Union (2024) Art. 72; ISO/IEC (2023b) §6.7
M1. Continuous	M1.3 Continuous monitoring mechanisms should be designed with the aim to minimise Mean Time to Detection (MTTD).	Organisations should define MTTD targets proportional to the objectives and risks of the system and review them periodically. Automated monitoring mechanisms should flag distribution shifts, usage patterns that indicate anomalous outputs, or deviations from expected behaviour to minimise the likelihood of undetected incident onset. Monitoring infrastructure should support low-latency detection through automated alerts and structured escalation pathways. Incident categories without a clear onset (e.g., intangible, distributed, or long-horizon harms) may be more challenging to detect, and therefore may have higher MTTDs, but organisations should still aim to minimise the time to detection for these incidents. Escalation rules (such as external audits) for when MTTD targets are exceeded should be set.	Johnson et al. (2024); O'Brien et al. (2023); Hoffmann & Frase (2023)
M2. Calibrated	M2.1 Tolerance bands for false positives and false negatives should be established and documented.	Define thresholds for false-positive and false-negative tolerance. Record the number of false-positive and false-negative alerts.	Vinnakota & Kolla (2025); Tariq et al. (2025)
M2. Calibrated	M2.2 Monitoring thresholds should be recalibrated on a regular schedule and after major system changes.	Detection thresholds, statistical baselines, and alert triggers should be recalibrated at defined intervals (e.g., monthly/quarterly) and after events such as model updates, safety-filter changes, or significant distribution shifts. Recalibration decisions should be logged and validated.	
M2. Calibrated	M2.3 Multi-layered detection combining automated and human signals should be implemented.	Monitoring should integrate automated anomaly detection, usage-pattern monitors, and structured human review for high-severity alerts. User reports, trusted researcher inputs, and partner feedback should serve as additional layers for capturing signals that automated systems may miss.	

Continued on next page

Principle	Guideline	Operationalisation Details	Guidelines Derived From
M2. Calibrated	M2.4 Alert volume should be managed to prevent alert fatigue without compromising safety.	Monitoring systems should implement alert prioritisation, deduplication, and routing rules to reduce noise and prevent operational overload. Suppression logic must be transparent, auditable, and prohibited from hiding alerts in high-severity categories or suppressing new anomaly types.	O'Brien et al. (2023); Tariq et al. (2025)
M3. Traceable	M3.1 Runtime behavioural data necessary to understand the sequence of events leading to an incident should be recorded.	Monitoring should capture inputs, outputs, model version identifiers, configuration parameters, and relevant runtime metadata with the goal of tracing system behaviour that triggered or otherwise contributed to the incident.	Anderljung et al. (2023); Ezell et al. (2025); ISO/IEC (2023a) §B.6.2.6; European Commission (2025c) Ch. 3
M3. Traceable	M3.2 A complete and time-stamped record of lifecycle and configuration changes should be maintained.	Logs should include model updates, fine-tuning, safety-layer modifications, tool-access changes, and deployment approvals.	
M3. Traceable	M3.3 Relevant external information sources used to identify or corroborate incidents should be monitored and preserved.	Organisations should track user reports, news coverage, public disclosures, or research publications that provide details relating to the occurrence and possible causes of an incident.	ISO/IEC (2023c) ; European Commission (2025c) Ch. 3 Commitment 9
M3. Traceable	M3.4 Post-deployment technical and organisational changes relevant to the incident should be captured.	Technical and organisational changes that affect system operation or oversight, including updates to safeguards, operating procedures, oversight structures, and risk controls, should be documented when they occur.	Dixon & Frase (2025)
M3. Traceable	M3.5 Monitoring records should be tamper-evident, access-controlled, and auditable.	Logs should be stored in formats that make unauthorised modification detectable. Access to logs should follow role-based controls and generate audit trails.	
M4. Impact-inclusive	M4.1 Observable indicators of an incident's scope, frequency, duration, and severity should be recorded.	Once an incident is detected, monitoring should capture measurable indicators of how widely the harm extends (scope), how often it occurs (frequency), how long it persists (duration), and the magnitude of its effects (severity). Indicators may include the number of affected users or systems, durations of service disruption, error rates, or the extent of policy or safeguard failures.	Scheuerman et al. (2021); Hoffmann & Frase (2023); Mylius
M4. Impact-inclusive	M4.2 Impact evidence should be collected through ecosystem monitoring and user-facing channels.	Organisations should monitor the ecosystem in which the AI system is deployed for issues and maintain awareness of new AI research findings and techniques. Organisations should provide channels for collecting user feedback or reports related to the impact of harm arising from an incident.	ISO/IEC (2023b); European Commission (2025c) Ch. 3

Continued on next page

Principle	Guideline	Operationalisation Details	Guidelines Derived From
M4. Impact-inclusive	M4.3 Impact-relevant signals should be timestamped to track incident progression.	Monitoring systems should preserve when impact indicators emerged, escalated, and resolved, enabling reconstruction of how the harm developed over time. For diffused or long-horizon harms where precise timing is difficult to establish, relevant indicators should still be recorded where feasible, and the absence of temporal precision should be documented.	
M5. Privacy-preserving defaults	M5.1 Lawful and privacy-protective bases for collecting monitoring-relevant data should be used.	Monitoring should only collect personal data under a lawful basis such as meaningful consent, contractual necessity, legitimate interests, or compliance with regulatory obligations, depending on the deployment context. The purpose of collecting monitoring data must be clearly specified and limited to incident detection and documentation. Data-minimisation checks should be performed before designing a data collection plan.	Stein et al. (2024)
M5. Privacy-preserving defaults	M5.2 The collection and retention of sensitive or identifiable data should be minimised.	Monitoring should only collect data strictly necessary for detecting and documenting incidents. Sensitive categories of data (e.g., biometrics, health data) should be avoided unless essential.	
M5. Privacy-preserving defaults	M5.3 Rigorous anonymisation or pseudonymisation should be employed.	Where feasible, logs should use privacy-preserving techniques such as pseudonymisation, or aggregation.	Feretzakis & Verykios (2024)
M5. Privacy-preserving defaults	M5.4 Risks of re-identification and data leakage should be evaluated and mitigated.	Organisations should conduct cyber-threat evaluations or privacy risk assessments to identify potential re-identification pathways or leakage vectors in monitoring data, including through linkage attacks, model inversion risks, or correlation across logs. Privacy impact assessments or risk reviews should be performed periodically.	Shahriar et al. (2023); Xin et al. (2025); ISO/IEC (2023b)
M5. Privacy-preserving defaults	M5.5 Information flows related to monitoring data should be documented and audited.	Data-flow diagrams detailing how sensitive data moves within and between systems, for what purposes, and for how long, should be documented and auditable. Retention and deletion policies must balance forensic integrity with data protection requirements.	Bluemke et al. (2023)

E. Reporting Template

The following template operationalises the principles discussed in Section C.

The template assumes reports are submitted to an external body that assigns an ID to a report series, enabling initial, follow-up, and final reports to be linked over time (Iterative). It accommodates incomplete or uncertain information, allowing reporters to specify whether inputs are known, estimated, or unknown (Epistemically transparent), and add detail in subsequent reports (Pragmatic). Explanatory text in italics clarifies what information is being requested (Unambiguous).

Fields relating to implicated systems are intended to be repeatable where reporters are aware that multiple systems contributed to an incident. In such cases, differential requirements may apply to main contributing systems and secondary systems.

Initial AND/OR Follow-up Report

Report Metadata

* Report type

☐ Initial ☐ Follow-up ☐ Final

Incident ID

Leave blank if Initial Report.

---

* Date of Submission

Auto-populated.

YYYY-MM-DD HH:MM

Reporter Information

* Surname

---

* First name

---

* Affiliation

Enter affiliation or select "No Affiliation."

---

* Role

Select all that apply.

☐ Developer ☐ Deployer ☐ User ☐ Third Party ☐ Other (specify)

Contact details for follow-up

Timeline

* Start date

If known or estimated, enter the date when the incident first occurred (as precisely as possible).

☐ Known ☐ Estimated ☐ Unknown

* Detection date

If known or estimated, enter the date when the incident was first detected (as precisely as possible).

☐ Known ☐ Estimated ☐ Unknown

* End date

If known or estimated, enter the date when the system(s) believed to have contributed to the incident was/were restored to normal functioning (as precisely as possible).

☐ Known ☐ Estimated ☐ Unknown

Incident Description

Describe what the system did or failed to do.

Describe observable system behaviour. Avoid causal explanations.

Continued on next page

AI Incident Report (continued)

Implicated System(s)

Provide the following information for each system known to the reporter whose behaviour may have contributed to the incident.

System's role in the incident

☐ Main contributing system ☐ Other contributing system (specify)

* Model name

* Model version

Release date

* Intended use

Select all that apply.

☐ Facial recognition ☐ Content moderation ☐ Medical diagnosis

☐ Hiring/recruitment ☐ Financial services ☐ Autonomous vehicles

☐ Conversational AI ☐ Content generation ☐ Other (specify)

Actual use

Describe how the system was being used at the time of the incident.

Deployment configuration

Describe relevant aspects of the system's deployment or operational setting. Include, where applicable: (1) the system's level of autonomy; (2) integration with external tools or APIs.

Known limitations or safety measures

Describe any known limitations, safeguards, or mitigations relevant to the incident.

Model type and architecture

Describe the model type (e.g. large language model, image classifier) and architecture (e.g. transformer, CNN, diffusion model).

Interaction with other systems

Describe known interactions with other AI or non-AI systems relevant to the incident.

Impact

* Impacted parties

Describe who (e.g. users, non-users) or what was affected by the incident.

* Nature of impact

Describe how they were affected.

* Harm types

Continued on next page

AI Incident Report (continued)

Select all that apply.

☐ Physical harm ☐ Psychological harm ☐ Financial harm

☐ Environmental harm ☐ Discrimination/bias ☐ Privacy violation

☐ Reputational harm ☐ Other (specify)

Number of users impacted

_____

☐ Known ☐ Estimated ☐ Unknown Additional detail: _____

Number of non-users impacted

_____

☐ Known ☐ Estimated ☐ Unknown Additional detail: _____

Economic damage

_____

☐ Known ☐ Estimated ☐ Unknown Additional detail: _____

Causality

* Describe suspected contributing factors to the incident.

Include, where applicable: (1) technical factors related to the system, data, model, or deployment; (2) non-technical factors related to human use, organisational, or governance context.

_____

* Suspected contributing factors

Select all that apply.

☐ Data issue ☐ Model issue ☐ System issue ☐ Deployment context

☐ User behaviour ☐ Organisational context ☐ Governance or policy gaps

☐ External factors ☐ Unknown