
Strengthening International AI Incident Discovery

Omer Bilgin
The Future Society

Aydin Yigit Alfatli
Independent

Ipek Cerrahoğlu
Istanbul Bilgi University

With
Apart Research¹

Abstract

Mandatory AI incident reporting exists to make dangerous events visible to regulators while there is still time to act. We coded the six such regimes in force or set to be enforced soon from primary legal texts and applied each trigger to the documented loss of control incidents of summer 2026, finding that none clearly compel a report. Disaster-scale thresholds are only part of the defect: most triggers require harm before engaging, almost every text exempts evaluation findings, and no binding instrument requires near misses to be reported. We set out recommendations for four stakeholder groups, mostly needing no new legislation: a reporting ladder whose incident tiers capture all real-world harm below the crisis threshold, measurable near-miss triggers with denominators developers can adopt unilaterally, and information-sharing channels between jurisdictions.

1. Introduction

Recent loss-of-control incidents involving OpenAI (spanning the [Hugging Face](#), [DseWiki](#), and [RubyGems](#) cases), Anthropic (outlined in [this blog post](#) and [this report](#)), [Meta](#), and the [UK AI Security Institute](#) demonstrate that AI risks are escalating. And as these events have transpired, companies have

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

demonstrated that they [will not always voluntarily disclose information](#) about unfolding incidents to relevant authorities until those incidents become public or clearly cross mandatory incident reporting thresholds, which are generally set quite high. But if these kinds of incidents do not reach relevant government authorities and industry bodies in a timely manner, no one can learn from these signals and proactively improve their resilience to emerging threats.

This memo distinguishes between four escalating classes of event², each of which are worthy of being reported:

- **Hazards:** risk-relevant signals uncovered by frontier AI companies during internal AI development and testing or by third parties during evaluations that *could plausibly* lead to any of the harms listed below, without causing any direct harm.
- **Near misses:** situations or events where the development, use or malfunction of an AI model or system *almost led to* real-world harm, but ultimately did not materialise.
- **Incidents:** events or circumstances where the development, use or malfunction of an AI model or system *directly or indirectly led to* real-world harm such as: (a) injury or harm to the health of a person or group of people; (b) disruption of the management or operation of critical infrastructure; (c) violations of fundamental human labour and intellectual property rights; or (d) harm to property or the environment. These refer to localised harms without societal spillover.
- **Crises:** severe, large-scale AI-driven disruptions or harms, typically spreading across borders, organisations or sectors.

According to these definitions, each of the loss of control cases that have been uncovered over the past two months classify as either near misses or incidents. Existing mandatory incident reporting regimes, however, typically trigger at the crisis rung rather than at any of the three rungs below it. Waiting for empirical proof of danger means waiting for people or property to be harmed. The incidents that have been uncovered have been relatively insignificant in terms of consequences – after all, nobody was hurt. However, the same failure in more capable systems is autonomous cyber operations, CBRN uplift, critical-infrastructure disruption, or agents operating outside any sandbox across several organisations at once.

2. Related work

Work on AI incident reporting has so far proceeded on several fronts, each addressing a different part of the problem, and none of them the question this paper asks.

- The Future Society's [The Case for Cross-Border AI Incident Infrastructure](#) (2026) makes an argument that this piece builds on: that existing infrastructure and current practices relating to AI incident management and crisis prevention, both domestically and internationally, are worryingly inadequate.
- [Kevin Wei and Lennart Heim](#) derive, from nine case studies of safety-critical industries, a framework of seven dimensions along which any incident reporting system can be specified.
- Those dimensions are only comparable across jurisdictions if the underlying vocabulary is shared, which is what the OECD's [common reporting framework for AI incidents](#) supplies: definitions of incidents and hazards, and twenty-nine reporting criteria that outline what an incident report should ideally contain.
- CSET's [An Argument for Hybrid AI Incident Reporting](#) makes the case for combining mandatory reporting of the most serious events with voluntary and third-party channels for the rest, and their

² The definitions of hazards and incidents above are adopted from the [OECD](#); the definition of near-misses are adopted from [CSET](#); and the definition of crises are based on The Future Society's [multi-stage escalation pathway](#) for AI risks and harms. Together these form a ladder of escalating severity: hazards → near misses → incidents → crises.

follow-up report, [AI Incidents: Key Components for a Mandatory Reporting Regime](#), operationalises that framework by identifying what a report must actually contain if incidents are to be documented consistently and compared over time.

- The [AI Incident Database](#) and the [OECD AI Incidents Monitor](#) log incidents continuously from media coverage and voluntary disclosures. The [MIT AI Incident Tracker](#) classifies those records by risk, cause, harm and severity. The [AI Vulnerability Database](#) catalogues model vulnerabilities and failure modes. This type of bottom-up infrastructure is indispensable, as it is how information about most hazards, near misses and incidents reaches the public.

What none of these works do is test the triggers that are already written into binding law against real incidents, and ask whether a duty to report would in fact have arisen. That is the gap this paper addresses. More specifically, this piece provides (1) an analysis of reportability triggers across every mandatory incident reporting regime that is currently in force or passed, and an assessment of whether this summer's loss of control incidents activate those triggers, and (2) a set of stakeholder-specific recommendations in light of our findings.

3. Methodology

This paper constitutes a qualitative document analysis of primary legal sources, combined with a counterfactual application of each source to recently documented events.

The corpus is every jurisdiction with a mandatory AI incident reporting duty in force, or adopted with a fixed commencement date: the European Union, California, New York, Illinois, South Korea and Vietnam. Proposed bills were excluded, on the ground that unenacted texts impose no duty and their contents are highly likely to change.

All six jurisdictions in scope of this paper were coded directly from primary legislative sources rather than from secondary ones. The Korean and Vietnamese legislation sources were read in their English-translated versions. Each regime was coded across fifteen fields, of which five carry the argument: the qualifying event; the severity threshold; whether harm must have occurred or risk suffices; coverage of the pre-deployment stage; and what starts the mandatory reporting clock. Every cell records the instrument, article and page from which it derives.

The test applied to each regime is counterfactual and deliberately practical: given the facts publicly available about the loss of control incidents of summer 2026, would a competent lawyer advising the developer have concluded that a report was legally required? Three verdicts are used: Yes, where the duty is clear; Contestable, where the text supports both readings and nothing in it resolves the question; and No.

Once our verdict for each jurisdiction was determined, we used our research findings to devise stakeholder-specific recommendations that could enable information about noteworthy hazards, near misses and incidents to still reach relevant authorities, so that these relevant signals can be detected and used to inform incident preparedness and crisis prevention.

4. Findings

Table 1. In which jurisdictions did or would the May-September 2026 loss of control incidents need to be reported to public authorities? Verdict and deciding provisions for each of the six regimes.

Jurisdiction	Verdict	Provisions
South Korea (AI Framework Act)	No	Arts. 32(1), 32(2) Scope: Art. 4
Vietnam	No	Arts. 3(8), 12(4), 12(5)

Jurisdiction	Verdict	Provisions
(Law on AI)		Scope: Art. 1(1). Rules: Decree 142/2026/ND-CP
California (SB 53)	No	Secs. 22757.11(d), 22757.11(c)(1)
Illinois (SB 315)	No	Sec. 5
New York (RAISE Act)	No	Secs. 1420(4), 1420(3)(a)
European Union (AI Act)	Contestable, but probably yes	Arts. 55(1)(c), 3(49)

This table states the outcome only. Our reasoning behind each verdict, limb by limb and with the operative wording of every provision relied on, is set out in the expanded version of this table ([Appendix A](#))

1. No regime in the sample clearly compels a report for these loss of control incidents. One is contestable and five are clear negatives, with official reports being filed in only one jurisdiction. California’s Office of Emergency Services stated on the record that their SB 53 thresholds were not met by these events, even though harm to property did occur in the form of credential and data compromise at a third party. And since Illinois’ SB 315 and New York’s RAISE Act adopt the same thresholds, their requirements also did not cover those incidents. The only live piece of legislation that could have and likely did cover this summer’s incidents in practice is the EU AI Act. OpenAI reported both the [Hugging Face incident](#) and the [DseWiki case](#) to the EU Commission, but did so [without conceding](#) that Article 55 bound it.

2. Severity is quantified only in the US state regimes, and these regimes trigger only at the crisis rung of the AI risk escalation pathway outlined in [Section 1](#). California, Illinois and New York all set their risk thresholds, which demand incident disclosure if crossed, at more than 50 deaths or serious injuries, or more than \$1 billion in property damage, from a single incident. Measured against our classification in Section 1, 50 deaths does not describe an incident; it describes a crisis, and almost every event from which a reporting regime could usefully learn in order to inform anticipatory rather than reactive governance falls far below it. Elsewhere, the term “serious” is either left undefined (EU, where it is used four times in Article 3(49) while never being explicitly defined), or qualitative (Vietnam), or absent (South Korea).

3. Pre-deployment incidents are the common blind spot. Most regimes only apply if an AI model is deployed in or otherwise materially affects a jurisdiction’s domestic market. California is the exception: it asks whether an incident was associated with internal use, and large developers are required to file quarterly internal-use risk summaries. But California’s pre-deployment reporting requirements only apply if there is death, bodily injury, another kind of tangible harm or damage involved, or a demonstration of materially increased catastrophic risk – and the New York and Illinois’ requirements share similar thresholds. The EU excludes incidents that may occur during pre-market research and development ([Article 2\(8\)](#)), and while [Article 55\(1\)\(c\)](#) establishes incident reporting duties for providers of general-purpose AI models with systemic risk that is not conditioned on public deployment in EU markets, there is no article in the EU AI Act that is explicitly about pre- reporting obligations for publicly unreleased models, which means that Article 55(1)(c) only applies if an incident "directly or indirectly leads to" one of the four ‘serious incident’ consequences outlined in [Article 3\(49\)](#). Vietnam’s legislation ties its triggers to operation and scopes the Law to activity inside the country, so an unreleased model tested abroad falls outside of it. South Korea only demands the monitoring and internal management of pre-deployment risks within companies and the reporting of detailed guidelines concerning the

implementation of safety measures taken by companies, so a pre-deployment event inside the duty still produces no legal duty to notify hazards or internally known incidents.

4. Near-miss reporting is absent from every regime. Not a single one of the six regimes creates a standalone duty to report a near miss or escalating hazard. The [Safety and Security chapter](#) of the EU's [General-Purpose AI Code of Practice](#) is the only instrument in scope that explicitly engages with these concepts. Measure 9.2 lists nine items signatories must keep track of, document and report to the AI Office, of which point (9) is “any patterns detected during post-market monitoring that can reasonably be assumed to be connected to the serious incident, such as individual or aggregate data on near misses”. However, this information is only required to be submitted after a serious incident has occurred as part of a serious-incident report, which excludes anything observed during pre-deployment evaluations.

5. Discussion

The findings in Section 4 point to a significant underlying problem with existing reporting regimes: **they are built to react to catastrophes rather than to learn from early warning signs.** Three min upshots follow, each of which the recommendations outlined in Section 6 are designed to address.

- 1. The incidents examined in the paper provide evidence that existing evaluation, testing and containment practices can fail to identify or adequately characterise relevant risks before deployment, and can even cause near misses and incidents.**
- 2. Thresholds are set so high that almost nothing reaches public authorities.** The events of summer 2026 were valuable tests of these regimes, because many of them have been unusually well documented, were caused by the most capable systems in existence, and every event in scope arose during internal development, testing or evaluation. However, five of six regimes investigated generated no duty at all. If incidents like these do not cross the reporting line, very little below the crisis rung ever will. The practical effect of this is that most hazards, near misses and most incidents will continue not to be reported.
- 3. The result is an ecosystem dependent on voluntary disclosure.** None of these events were disclosed first by the developer. The Hugging Face intrusion became public because the victim detected and published it, and OpenAI initially treated it as an internal misalignment research finding rather than a reportable incident, only [notifying the EU Commission](#) once third-party damage at Hugging Face was confirmed; similarly, the [DseWiki case](#) became public because independent researchers read public logs, and OpenAI [confirmed it only afterwards](#), arguing that the industry lacks a standard for disclosing unintended model behaviour. Reporting under a voluntary regime is therefore conditional on a second party noticing, on a researcher looking, and on the developer's assessment of the reputational cost of silence against the cost of speaking.
- 4. Downstream, this is what makes cross-border preparedness extremely challenging in practice.** AI Safety Institutes see some hazard signals, but face strict constraints on what they may disclose and to whom, and governments can otherwise only share what they receive. If domestic duties do not generate reports, there is no material for an information-sharing channel to carry, and no aggregate picture from which any state could derive early warning. Every proposal for international AI incident infrastructure is therefore downstream of a domestic trigger problem that is currently unsolved in five of the six jurisdictions that have legislated at all.

6. Recommendations

6.1 Governments and Legislators

- 1. Extend trigger scope to pre-deployment and internal use:** The summer 2026 loss of control incidents produced zero casualties and no publicly quantified damage; therefore, lowering

existing triggers – even to a single injury or minor dollar amount – would still fail to capture them. Instead of simply lowering the numerical threshold, the criteria that trigger it should be changed.

2. **Establish a Three-Tier Reporting Ladder with Defined Denominators:** Align thresholds downward where mandatory reporting already exists, and replace single thresholds with three tiers matching the ladder outlined in Section 1.

Tier 1: Crisis (Materialized Catastrophic Harm). This tier should be strictly reserved for catastrophic events, defined as resulting in 50 or more fatalities or serious injuries, or exceeding \$1 billion in damages.

Tier 2: Incident. This tier should capture every instance of real-world harm that falls below the Tier 1 crisis threshold, where model development, use, or malfunction still causes health injuries, critical-infrastructure disruption, fundamental or labor rights breaches, or harm to property or the environment

Tier 3: Hazard and Near Miss. Most severe harms are too rare to be able to provide useful insights to inform governments’ incident and crisis prevention efforts, so proxies are necessary, and proxies need denominators. Legislators should require frontier AI companies to establish continuous monitoring mechanisms with clear triggers and denominators that keep track of information, such as how many times an unreleased model that is being developed and tested internally tried to escape its sandbox, that can provide insight about near misses and escalating hazards.

3. **Establish Cross-Border Incident Sharing Channels:** Because comprehensive legislative action typically faces prolonged bureaucratic friction, governments should implement bilateral or plurilateral incident sharing arrangements.

Asymmetric "Receive-Only" Pipelines: Allow jurisdictions with lower thresholds to transmit notifications to higher-threshold jurisdictions. Even if a receiving nation cannot politically lower its domestic legal thresholds, it still benefits operationally from what a partner’s lower threshold captures.

Foundational Data Before Full Reports: Build cross-border institutional trust by initially sharing aggregate, anonymized metrics (threat category, timestamp, model identity, provider, and consequences) before expanding to full post-mortem reports.

Operationalize existing statutory pathways for cross-border transmission: Agencies should use existing legal handoffs to move data internationally. For example, California’s CalOES already holds explicit statutory authority under [SB 53 Section 22757.13\(e\)](#) to forward portal incident reports from members of the public and frontier developers to federal agencies. Federal partners should formally establish secondary pipelines routing this domestic intake to established bodies like the International Network of AISIs, multilateral institutions (OECD/UN), or existing intelligence-sharing alliances (e.g., Five Eyes) for international threat awareness.

Borrow aviation’s confidentiality architecture. [Annex 13](#) of the International Civil Aviation Organization’s Convention on International Civil Aviation already defines roles for joint incident investigations and confidentiality rules for the cross-border exchange of safety-critical information.

6.2 Frontier AI Companies

1. **Build the monitoring infrastructure that makes reporting possible.** Reporting duties and company-level monitoring are complements: governments can define *what* must be reported and at what threshold, but only developers can build the systems that continuously observe, measure

and record the underlying behaviour. Without that infrastructure, a legal duty to report near misses is inert, because companies cannot disclose what they are not systematically collecting. Accordingly, developers should build continuous-monitoring systems across development and testing, so that relevant behaviours, anomalies, dangerous capabilities and near misses are detected and logged as they happen rather than reconstructed afterwards or discovered on an ad hoc basis. And they should design these systems with external reporting in mind, so that near-miss and hazard data can be aggregated and disclosed to regulators, AI safety institutes or other authorised external actors when required.

2. **Operationalise trigger events and denominators for risk escalation.** Companies should define measurable indicators through which warning shots and patterns of escalating risk can be detected and identified, such as number of detected occurrences where a model tried to escape its sandboxed evaluation environment. **Table 2** below showcases some examples of the kinds of concrete denominators companies could employ.

Table 2. Potential operationalisable trigger events and denominators for warning shot and risk escalation detection

#	Warning shot trigger event	Denominator
1	An AI system being tested reaches a network or service it was never permitted to reach	per hour of testing
2	An AI system writes to, or tries to log in to, computer systems belonging to someone else	per hour of operation
3	An AI system obtains or uses passwords or access keys it was never issued	per test run
4	AI systems communicate with one another through a channel their operators did not set up for them	per 1,000 test runs
5	An AI system cheats on a test — for example by hunting down the answers rather than solving the task	per task, in testing and in use
6	Monitoring meant to watch what the AI system is doing is switched off, evaded or degraded mid-run	per model, per quarter
7	An AI system breaks a law or company policy; or paused training is restarted after a containment failure	per company, per month; per event

3. **Improve the quality of internal evaluations and sandboxing.** The incidents examined in the paper provide evidence that existing evaluation, testing and containment practices can fail to identify or adequately characterise relevant risks before deployment. Companies should therefore strengthen evaluation protocols, sandboxing environments, adversarial testing protocols, and the range of conditions under which models are tested.
4. **Self-report proactively, without waiting for legislation,** and publish the counts above with their denominators. Such disclosures can be made without exposing any commercially sensitive information.
5. **Take actions individually if unsafe practices continue to occur.** There are several types of action that can move institutions without death. For example, the [public resignation of an Anthropic safety researcher](#), Jacob Coxon, in September 2026, who forfeited equity to speak, has generated major policy attention recently and has helped gather momentum for extending safety norms to internal development and testing contexts.
6. **Agree on common loss-of-control near miss and incident reporting standards, potentially through the Frontier Model Forum (FMF).** OpenAI has already [publicly stated](#) that no such

standard exists; that is a solvable problem. The FMF already has an [established information-sharing agreement](#) among several US-based frontier AI companies, and this existing mechanism could be used to build the foundations needed for shared standards.

7. **Fund and run incident-discovery sprints similar to bug-bounty programmes that already exist**, aimed at surfacing what your own pre-release models did during development and testing. Findings should ideally be shared publicly, or at minimum with relevant authorities and industry peers via the FMF.

6.3 Funders, Research Organisations and Civil Society

1. **Fund the external detection layer as reporting infrastructure.** The fact that the summer 2026 loss-of-control incidents were primarily surfaced by third parties rather than developers demonstrates that independent open-web monitoring is vital to keep statutory reporting honest. Funders should direct resources to established technical evaluation groups, such as Apart Research, Apollo, METR and comparable groups, treating open-web incident discovery as a funded standing programme rather than a one-off activity.
2. **Standardise what research organisations and community research efforts surface.** Much of this already happens informally on public forums; giving it structure and coordinated disclosure norms could convert public findings into evidence government authorities can use. A public register of containment-failure indicators, with denominators, could give regulators specific things to ask for and developers specific concerns to answer. This quantitative data should be paired with standardized qualitative post-mortems. Adopting structured incident investigation frameworks like [METR's proposed post-incident methodology](#)—which probes model reasoning, safeguard circumvention, and unauthorized collaboration—can transform public findings from simple incident detection to robust root-cause analysis.
3. **Test the instruments that already exist.** While [CalOES's portal](#) accepts submissions from members of the public, and the EU Commission via its GPAI serious-incident template, neither are currently widely utilised. Filing incident reports in various jurisdictions using public findings would materially help government authorities in those jurisdictions detect more hazards, near misses and incidents and better understand previously detected ones.

8. Future Work

Get our stakeholder-specific recommendations into the hands of advocacy organisations. This paper makes some arguments, but a research sprint cannot place that argument in front of the people who would have to act on it. The most valuable next step would therefore be distribution, rather than further analysis. Each recommendation set in Section 6 is addressed to a group we have no standing to lobby directly. Civil society, research and advocacy organisations working on AI governance do have that standing, and several already maintain the relationships needed to carry these proposals into the right rooms. The natural form of this work would be a short, tailored brief per stakeholder group, derived from Section 6.

Exercise the reporting instruments that already exist. As recommended in Section 6.3, using [CalOES's public portal](#) and the [Commission's GPAI serious-incident template](#) to report any hazards, near misses or incidents that occur in the future would help with their detection by public authorities.

Extend and maintain the corpus. Our test is a snapshot of six regimes at a single moment. It should be re-run as New York and Illinois commence and as pending bills around the world are enacted, several of which, if passed, would materially change the picture. A living tracker, updated on commencement rather than on publication, would let others re-run the counterfactual against future incidents without rebuilding the coding.

Test incident sharing channels on narrow cases. The asymmetric, receive-only incident sharing model proposed in Section 6.1 has never been tried for AI incidents. A bilateral pilot between two willing jurisdictions would establish whether the trust hurdle is as low as we assume and hope.

9. Conclusion

The best-documented loss of control events in the history of frontier AI produced a mandatory incident report in only one jurisdiction out of six, and even there the developer did not concede that the duty applied. What must change is that the trigger engages on what the model *did* rather than on what it damaged, that evaluation-stage findings stop being carved out at exactly the point where a dangerous capability first becomes visible, and that near misses and hazards acquire a tier of their own with countable triggers and stated denominators. Where legislatures cannot move at that speed, receive-only sharing channels and a funded external detection layer can carry the load in the meantime. The events of summer 2026 were the smallest version of their category: nobody was hurt, they were detected by their victims rather than declared by their causes, and they cleared almost every legal threshold in the world. The next one will be larger, and the thresholds will not have moved.

Appendix A. Reporting Regime Analysis

Jurisdiction	Verdict	Justification
South Korea (AI Framework Act)	No	No explicit incident reporting duty or category of reportable incident is specified anywhere in the Act. Article 32 requires frontier AI companies to identify, assess and mitigate risk across the entire AI lifecycle, and to build and maintain a risk management system capable of monitoring and responding to AI safety incidents. Article 32(2) then requires submission to the Minister of Science and ICT of <i>the results of implementing those measures</i>, but not notification that an event occurred. Article 4 extends the Act to conduct outside of Korea, but only if that affects the domestic market or Korean users, which the incidents this piece concerns did not do.
Vietnam (Law on AI)	No	Article 3(8) covers any event that causes or poses a risk of causing significant harm to life, health, human rights, property, cybersecurity, public order or the environment, or that disrupts critical information systems for national security — so a risk of harm suffices, without any quantified floor. Article 12 attaches real duties to it: on a serious incident, developers and providers must promptly apply technical measures to rectify, temporarily suspend or recall the system and simultaneously notify the competent authority, while deployers and users must record and promptly notify the incident, with everything routed through the National Single-Window Artificial Intelligence Portal (Art. 12(4)) and the severity-proportionate reporting rules left to the Government (Art. 12(5)); Decree 142/2026/ND-CP supplies those rules. However, Article 1(1) outlines that this legislation applies to foreign organisations only where they provide or deploy systems affecting people or infrastructure inside Vietnam. Therefore, this summer’s internal testing and evaluation incidents fall outside of its scope.
Illinois (SB 315)	No	Section 5 reproduces California’s definitions almost verbatim. “Catastrophic risk” again requires material contribution to the death of or serious injury to more than 50 people, or

Jurisdiction	Verdict	Justification
		more than \$1 billion in property damage, from CBRN uplift, unsupervised conduct amounting to a cyberattack or a crime, or a model evading its developer's control. "Critical safety incident" again runs to the same four limbs, with death or bodily injury required in limbs 1 and 3 and the same exclusion in limb 4 of behaviour arising in "an evaluation designed to elicit this behavior". The analysis and the outcome are therefore identical to California's. Applies from 1 January 2027, so no duty could have arisen on these facts in any event.
New York (RAISE Act)	Contestable	Section 1420(13) defines a "safety incident" considerably more broadly than California, covering: (a) a frontier model autonomously engaging in behaviour other than at the request of a user; (b) theft, misappropriation, malicious use, inadvertent release, unauthorised access, or escape of model weights; (c) the critical failure of any technical or administrative controls, including controls limiting the ability to modify a model; or (d) unauthorised use of a frontier model. Limbs (a) and (c) describe a containment escape directly, and — unlike California and Illinois — neither is gated on death or bodily injury, nor is any of the four subject to an evaluation carve-out. What makes this contestable is the qualifying clause: the event must occur "in such a way that it provides demonstrable evidence of an increased risk of critical harm". Section 1420(7) sets "critical harm" at the death or serious injury of 100 or more people (twice California's figure) or at least \$1 billion in damages, arising from CBRN use or from a model acting with limited human intervention in conduct that would, if committed by a human, be a crime of intent, recklessness or gross negligence. An agent that defeats its containment and reaches a third party's production infrastructure is arguably demonstrable evidence of increased risk of precisely that, though a developer resisting the duty could argue that the evidentiary link to a 100-fatality or billion-dollar outcome is speculative. Applies from 1 January 2027, so no duty could have arisen on these facts in any event.
European Union (AI Act)	Contestable, but probably yes	Article 55(1)(c) sets out requirements for frontier AI developers to keep track of, document and report relevant information about serious incidents and possible corrective measures to address them to the AI Office and to relevant national competent authorities "without undue delay". A "serious incident" under Article 3(49) requires one of four consequences to have occurred: (a) death or serious harm to a person's health; (b) serious and irreversible disruption of the management or operation of critical infrastructure; (c) infringement of obligations under Union law intended to protect fundamental rights; (d) serious harm to property or the environment.

Appendix B. Limitations

Legal interpretation and interpretive uncertainty: This study offers comparative statutory analysis rather than formal legal advice or regulatory adjudication. A competent lawyer could reach the opposite conclusion on any entry we mark Contestable, which is why we recorded that verdict rather than forcing a binary. The frame within which those readings were produced is also our own. The fifteen fields, and in

particular the five that carry the argument, are choices we made about what matters in a trigger. A different framing could produce different verdicts on the same texts.

Exclusion of future legislation: We excluded proposed bills that have not been enacted yet, which means this piece does not assess the several legislative vehicles that would, if passed, materially change the picture. Our findings describe the law as it stands, not the law as it may shortly become.

Underlying policy assumptions. The viability of our recommendations rests on two core premises. First, we assume that capturing lower-tier incidents and near misses produces actionable intelligence rather than unmanageable regulatory noise, an inference drawn from mature safety regimes in civil aviation and biosafety that has yet to be empirically proven in frontier AI governance. Second, we presume that asymmetric, receive-only incident sharing channels could sufficiently mitigate national security and proprietary disclosure challenges and facilitate effective cross-border coordination. While conceptually sound, plurilateral information sharing pipelines of this nature remain untested at scale among major AI powers.

Data

The dataset underlying this paper is our [multi-jurisdiction comparison tracker](#): fifteen coded fields per jurisdiction with instrument, article and page citations, source lists and per-cell confidence ratings.

Author Contributions

A.Y.A analyzed existing and upcoming mandatory incident-reporting regimes across primary legislative sources, and built the multi-jurisdiction comparative tracker. O.B. did most of the subsequent research, translated A.Y.A’s analysis into a written document, and led the strategic framing of the paper. Ī.C. provided critical review and editing of the draft. All authors participated in the brainstorming and argument-formation phases, critically reviewed the manuscript, and approved the final text.

References

Primary legal instruments

- European Union. Regulation (EU) 2024/1689 (AI Act) — Arts. 2(8), 3(49), 55(1)(c), 57(11), 60(7), 73, as amended by Regulation (EU) 2026/1744. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- EU AI Act, Article 3 (definitions, incl. Art. 3(49) “serious incident”). <https://artificialintelligenceact.eu/article/3/>
- EU AI Act, Article 55 (obligations of providers of GPAI models with systemic risk). <https://artificialintelligenceact.eu/article/55/>
- EU AI Act, Article 73 (reporting of serious incidents). <https://artificialintelligenceact.eu/article/73/>
- EU AI Act Explorer (consolidated text and article navigation). <https://artificialintelligenceact.eu/>
- European Commission (2025). *General-Purpose AI Code of Practice*. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
- European Commission (2025). *GPAI Code of Practice — Safety and Security Chapter* — Commitment 9, Measures 9.1–9.4 (in particular Measure 9.2 point (9) and Measure 9.3), Measures 1.3, 2.1, 3.5 and 7.6, and the glossary definition of “near miss”. <https://ec.europa.eu/newsroom/dae/redirection/document/118119> 7a. *EU AI Act: General-Purpose AI Code of Practice, Final Version* (full text, Safety and Security chapter), used to verify the verbatim wording of Measures 9.2 and 9.3. <https://code-of-practice.ai/?section=safety-security>
- California. Senate Bill 53, Transparency in Frontier Artificial Intelligence Act, chaptered text — §§22757.11(c)(1), 22757.11(d), 22757.12(d), 22757.13. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53
- California. SB 53 §22757.13(e) (onward transmission of reports), bill navigation view. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB53
- California. SB 53, full text (LegiScan mirror). <https://legiscan.com/CA/text/SB53/id/3271094>

- New York. RAISE Act, Gen. Bus. Law Art. 44-B — §1420(7) (“critical harm”), §1420(13) (“safety incident”), as enacted by chapter amendment S8828. <https://legislation.nysenate.gov/pdf/bills/2025/S8828>
- New York. RAISE Act, Assembly Bill A6453-A. <https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/A>
- Illinois. Artificial Intelligence Safety Measures Act, Senate Bill 315 (104th General Assembly), Section 5. <https://www.ilga.gov/Legislation/BillStatus/FullText?GAID=18&DocNum=315&DocTypeID=SB&LegId=157797&SessionID=114>
- Republic of Korea. *Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust*, Law No. 20676 — Arts. 2, 4, 32. <https://www.law.go.kr/lsInfoP.do?lsiSeq=268543>
- Republic of Korea. Enforcement Decree of the AI Framework Act, Presidential Decree No. 36053. <https://www.law.go.kr/LSW/lsInfoP.do?efYd=20260122&lsiSeq=282879>
- Korea AI Basic Act Explorer (English reference text, cross-check). <https://aibasicact.kr/explorer/>
- Socialist Republic of Vietnam. *Law on Artificial Intelligence* No. 134/2025/QH15 — Arts. 1(1), 3(8), 12. <https://en.baohinhphu.vn/viet-nams-law-on-artificial-intelligence-111260715113551787.htm>
- Socialist Republic of Vietnam. Decree No. 142/2026/ND-CP. <https://congbao.chinhphu.vn/van-ban/nghi-dinh-so-142-2026-nd-cp-469480.htm>

Definitions, reporting-system design and prior literature

- OECD (2025). *Towards a Common Reporting Framework for AI Incidents*, OECD Artificial Intelligence Papers No. 34 (full PDF). https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/towards-a-common-reporting-framework-for-ai-incidents_8c488fdb/f326d4ac-en.pdf
- OECD (2025). *Towards a Common Reporting Framework for AI Incidents* (landing page). https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html
- Center for Security and Emerging Technology (2025). *AI Incidents: Key Components for a Mandatory Reporting Regime*. Georgetown University. <https://cset.georgetown.edu/publication/ai-incidents-key-components-for-a-mandatory-reporting-regime/> — full PDF: <https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Incidents.pdf>
- Center for Security and Emerging Technology (2024). *An Argument for Hybrid AI Incident Reporting*. <https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/>
- Wei, K. and Heim, L. *Designing Incident Reporting Systems for Harms from General-Purpose AI*. <https://arxiv.org/abs/2511.05914>
- The Future Society (2026). *The Case for Cross-Border AI Incident Infrastructure*. <https://thefuturesociety.org/cross-border-ai-incident-infrastructure/>
- The Future Society. *What Is an Artificial Intelligence Crisis and What Does It Mean to Prepare for One* (multi-stage escalation pathway). <https://thefuturesociety.org/aicrisisexplainer/>
- Frontier Model Forum (2026). *Progress update: FMF information sharing of frontier AI threats and vulnerabilities*. <https://www.frontiermodelforum.org/updates/progress-update-fmf-information-sharing-of-frontier-ai-threats-and-vulnerabilities/>
- METR (2026). *Investigating AI propensities after incidents*. <https://metr.substack.com/p/2026-07-28-investigating-ai-propensities-after-incidents>
- Novelli, C. et al. Analysis of Article 55 and systemic-risk obligations under the AI Act. *International Journal of Law and Information Technology*. <https://academic.oup.com/ijlit/article/doi/10.1093/ijlit/eaaf019/8501132>

Incident record and regulatory response

29. OpenAI (2026). *The Hugging Face incident and the road ahead*. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
30. OpenAI (2026). *OpenAI and Hugging Face partner to address security incident during model evaluation*. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
31. OpenAI (2026). Statement on the “wiki incident”, posted to X, 5 September 2026. <https://x.com/OpenAI/status/2096133504417616165>
32. Anthropic (2026). *Investigating incidents in cybersecurity evaluations*. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
33. Anthropic (2026). *Alignment assessment: cybersecurity incidents*. <https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents>

34. UK AI Security Institute (2026). *Security Incident INC-2026-07-28-01*. https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf
35. DseWiki / collusion.wiki — archive of the agent coordination board. <https://collusion.wiki/>
36. RubyGems incident record. <https://www.rubyhack.ai/>
37. BBC News (2026). Report on the Meta evaluation incident. <https://www.bbc.co.uk/news/articles/cx2kgdnyk2po>
38. Reuters (2026, 4 September). *OpenAI agents hijacked German website in previously undisclosed AI breakout*. <https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/>
39. Reuters (2026, 7 September). *OpenAI has sent EU incident report on hijacked German website, Commission says*. <https://www.reuters.com/business/openai-has-sent-eu-incident-report-hijacked-german-website-commission-says-2026-09-07/>
40. Euractiv (2026). *OpenAI notified EU of Hugging Face hack under AI Act*. <https://www.euractiv.com/news/openai-notified-eu-of-hugging-face-hack-under-ai-act/>
41. Euractiv (2026). *AI safety incident reporting “not just a tick-box”, warns EU*. <https://www.euractiv.com/news/ai-safety-incident-reporting-not-just-a-tick-box-warns-eu/>
42. The Next Web (2026). *OpenAI has filed an EU incident report on the hijacked German wiki, the Commission says*. <https://thenextweb.com/news/openai-eu-incident-report-german-wiki>
43. Fortune (2026, 7 September). *OpenAI’s AI agents secretly ran their own message board on a German wiki*. <https://fortune.com/2026/09/07/openai-ai-agents-german-wiki-ran-their-own-message-board/>
44. Mission Local (2026). *California’s first AI-safety law didn’t cover the first rogue AI hacks*. <https://missionlocal.org/2026/09/california-ai-safety-law-sb-53-openai-hack-wiener-newsom/>
45. Superpower Daily (2026). *California says OpenAI agent hack did not trigger its AI safety reporting law*. <https://superpowerdaily.com/posts/california-says-openai-agent-hack-did-not-trigger-its-ai-safety-reporting-law>
46. CBS News (2026). Coverage of METR’s findings on the OpenAI–Hugging Face incident. <https://www.cbsnews.com/news/openai-hugging-face-hack-ai-risks/>
47. Axios (2026, 9 September). *Anthropic whistleblower gave up his equity to leave the company*. <https://www.axios.com/2026/09/09/anthropic-researcher-ai-warning-interview>

Additional background sources consulted

53. Republic of Korea. AI Framework Act, unofficial English translation, Korea Legislation Research Institute. https://elaw.klri.re.kr/eng_service/lawView.do?hseq=71019&lang=ENG
54. Kolt, N. et al. (2024). *Responsible Reporting for Frontier AI Development*. AIES 2024. <https://arxiv.org/abs/2404.02675>
55. Centre for Long-Term Resilience (2024). *AI Incident Reporting: Addressing a Gap in the UK’s Regulation of AI*. <https://www.longtermresilience.org/reports/ai-incident-reporting-addressing-a-gap-in-the-uks-regulation-of-ai/>
56. Frontier Model Forum (2026). *Information Sharing, Incident Reporting, and Incident Response for Frontier AI Risks*. <https://www.frontiermodelforum.org/issue-briefs/information-sharing-incident-reporting-and-incident-response-for-frontier-ai-risks/>
57. Cloud Security Alliance (2026). *OpenAI’s Wiki Silence Tests the EU AI Act’s Incident Regime*. <https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-incident-disclosure-gap-eu-ai-act-20260/>
58. Hugging Face (2026). *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
59. Hugging Face (2026). *Security incident, July 2026*. <https://huggingface.co/blog/security-incident-july-2026>

Incident repositories and reporting channels

48. Responsible AI Collaborative. *AI Incident Database*. <https://incidentdatabase.ai/>
49. OECD. *AI Incidents Monitor (AIM)*. <https://oecd.ai/en/incidents>
50. California Governor’s Office of Emergency Services. *Transparency in Frontier AI Act reporting portal*. <https://www.caloes.ca.gov/office-of-the-director/operations/homeland-security/california-cybersecurity-integration-center/transparency-in-frontier-ai-act-reporting/>
51. Apart Research. *AI Incident Response Sprint*, September 2026. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>
52. International Civil Aviation Organization. *Annex 13 to the Convention on International Civil Aviation: Aircraft Accident and Incident Investigation*.

LLM Usage Statement

Claude was employed throughout this project in a supporting capacity, with all substantive judgements retained by the human team.

At the ideation stage, it was used to investigate the initial ideas and to test the arguments against the underlying legal texts, so that propositions unsupported by the sources are discarded early. The same approach was applied to the comparative analysis, where LLM assistance was used to structure and populate the entries, each of which was subsequently checked against the initial findings and the primary sources. The tables presented in this report were generated with Claude from the verified underlying data. Finally, LLM assistance was used for cosmetic purposes, namely polishing the wording of the drafted text, and to list the references cited in the piece in the same style.