
A Specialised Reporting Framework: What FLARE-AI is Missing

Gaven Ren
University of Toronto

Michela De Mattia
AI Collective/AIGS
Canada

Dongho Park
University of Toronto

With
Apart Research

Abstract

Standardized AI incident reporting has a fatal flaw: connecting researchers (who gain from disclosure) and companies (who risk liability) through the same platform leads to companies and frontier labs scrubbing critical safety details. This paper tackles this issue by proposing a novel framework, replacing universal forms with four targeted instruments that separate confidential liability reports from public safety data, inspired from FAA and CVE models. The framework, stress-tested against failure scenarios, shows that syncing report structures to professional risk unlock the actionable technical nuance otherwise hidden.

1. Introduction

This paper extends the discourse introduced by [MIT](#) by arguing that reporting systems must be context-specific rather than generalised to be effective. We offer a mechanism for researchers, companies and regulators to share newly discovered safety measures without compromising proprietary secrets and to report incidents. In particular, we provide a **Specialised Reporting Framework** and related templates to report flaws and incidents effectively and efficiently to be better prepared against future issues. We also introduce a new form to report incidents and to help streamline the request for information from legal entities to get more details about the incidents.

2. Related Work

Recent efforts to standardise AI incident reporting, most notably in [FLARE-AI: Flaw Reporting for AI](#)¹, have provided a valuable foundation for data consistency. However, the practical application of such generalised frameworks is limited by the nuanced legal and regulatory landscapes surrounding AI liability and corporate confidentiality. Below, we highlight the major gaps between regulatory reports and incident reporting from frontier AI companies to present the current status.

What already exists.

We base our framework considering two mandatory instruments currently in use for AI incident reporting. The EU's Report for Serious Incidents (General-Purpose AI Models with Systemic Risk), issued under the AI Act ([Article 73](#))², requires reporting of defined categories of serious harm to a government recipient, with an initial report followed by intermediate and final updates³. California's SB 53 ([the TFAIA](#))⁴ requires frontier developers to report a "critical safety incident" or a "summary of any assessments of potential catastrophic risk resulting from internal use of a frontier model", to the Office of Emergency Services.

In both cases, reporting is compulsory, it runs from the developer to the state, it triggers only on a defined and fairly severe class of harm, and the report itself stays confidential with the government by default. We analysed the differences between these two submissions: see [Appendix A](#).

Both instruments are built to capture an [incident](#), meaning harm that has already occurred and crossed a severity line. Even if [the TFAIA](#) can potentially capture a potential catastrophic risk; neither is built to capture a [flaw](#): a vulnerability or a near-miss⁵ that has not yet produced measurable damage. Both the EU and California miss critical early warnings for AI risks: the EU ignores the testing phase and cross-border incidents, while California excludes evaluation events and sets its harm threshold too high. Additionally, our analysis ([Appendix B](#)) builds on FLARE-AI's review of public flaw reports, but it is worth considering that a complete assessment requires access to internal corporate incident filings.

3. Methods

¹ Longpre, S., Zhu, E., Ezell, C., Ghosh, A., et al. (2026). FLARE-AI: Flaw Reporting for AI. arXiv:2606.31567 (presented at ICML 2026). <https://arxiv.org/abs/2606.31567>

² European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 73: Reporting of Serious Incidents. Official Journal of the European Union, EUR-Lex CELEX:32024R1689. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>

³ European Commission (2025). Draft Guidance on the Reporting of Serious Incidents under Article 73 of the AI Act. European Commission, DG CNECT. <https://digital-strategy.ec.europa.eu>

⁴ California State Legislature (2025). Senate Bill 53: Transparency in Frontier Artificial Intelligence Act. California Legislative Information. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

⁵ OECD (2024). Defining AI Incidents and Related Terms. OECD Artificial Intelligence Papers No. 16. OECD Publishing. https://www.oecd.org/en/publications/defining-ai-incidents-and-related-terms_d1a8d965-en.html

3.1. Framework Design & Methodology

The core objective of this approach is to replace the "Generalised Reporting" model with a **Specialised Reporting Framework**: we developed an incident management framework that maps the entire lifecycle of an AI failure, from discovery to public resolution.

3.2. Framework Rationale

Our design choices are grounded in two key theoretical pillars:

3.2.1 The Liability Paradox: Drawing from precedents in Aviation Safety⁶ and Cybersecurity (CVE/CISA)⁷, we observed that the most accurate data is gathered when the reporter is granted a "Safe Harbour." By separating the *Liability Report* (Confidential) from the *Safety Contribution* (Public), we mimic the "no-fault" reporting systems used in aviation, which prioritise systemic safety over individual punishment.

3.2.2. Principal-Agent Theory⁸: We recognise that a researcher and a corporate CEO have different "costs" associated with reporting. A generalised form imposes the same "cost" (risk) on both. Our specialised forms lower the "cost of disclosure" by tailoring the level of detail and confidentiality to the reporter's specific risk profile.

3.3. Ideal Conditions

Our framework presuppose conditions that any reporting architecture needs to function:

3.3.1. A shared incident classification⁹. A common taxonomy across the mandatory instruments. AI incidents should be classified by *Criticality* (Low → Catastrophic) and *Reporting Urgency* (Preventive → Post-hoc).

3.3.2. Severity-tiered reporting times. timeline matching stakes instead of applying one deadline to everything.

3.3.3. Centralised, federal-level routing. Borrow FLARE's one-to-many model so a single filing reaches every affected developer, regulator, and database at once. This removes the interoperability gap and the reporter's distribution burden in one move.

3.4. Reproducibility

This work can be reproduced by applying the provided four templates to any AI incident. A researcher or regulator can implement this by establishing the four distinct "pipes" of communication defined in our architecture.

4. Results

⁶ NASA (2024). Aviation Safety Reporting System: Program Overview. National Aeronautics and Space Administration. <https://asrs.arc.nasa.gov/overview/summary.html>

⁷ Cybersecurity and Infrastructure Security Agency (2024). Coordinated Vulnerability Disclosure Process. CISA. <https://www.cisa.gov/resources-tools/programs/coordinated-vulnerability-disclosure-process>

⁸ Jensen, M. C., & Meckling, W. H. (1976). <https://www.sciencedirect.com/topics/social-sciences/principal-agent-theory>

⁹ OECD (2025). Towards a Common Reporting Framework for AI Incidents (Report No. 34). OECD Publishing. https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html

We argue that *standardised* reporting is not the same as *effective* reporting. While generalised frameworks (as proposed in *AI Needs a Better Way to Report Flaws. So We Built One*)¹⁰ aim for data consistency, they inadvertently create a "disclosure chilling effect." When a single reporting standard is used for both technical debugging and legal compliance, companies will naturally default to the most conservative, legally sanitised language possible. This results in a "lossy" transmission of safety data where the most critical technical nuances are scrubbed to avoid admitting liability.

We develop the **Specialisation-Based Incident Management Framework**. This framework rejects the notion of a generalised reporting system and instead proposes a modular, four-form architecture. By tailoring the information flow to the specific legal and technical needs of the reporter and the recipient, it optimises the reporting process to increase global AI safety awareness, facilitate the disclosure of researcher knowledge, and standardise the methodology for incident management.

4.1 The Four Instruments. The Framework operates through instruments designed to resolve a specific information asymmetry within the AI ecosystem:

4.1.1. Preventive Disclosure Form (Researcher → CISA / ISO / OECD / Frontier Labs):

- **Objective:** To stop a flaw *before* the incident occurs.
- **Design:** Prioritises technical precision and speed. It requires a **Proof of Concept (PoC)** to verify the flaw and a **Proposed Remediation Plan**. By requiring the researcher to suggest a mitigation strategy, the form transforms a "bug report" into an "actionable safety update" for CISA, ISO, OECD, and frontier AI labs.

4.1.2. Liability-Aware Incident Report (Company → Regulator):

- **Objective:** Formal notification of a failure while managing legal exposure.
- **Design:** Employs neutral, forensic language to describe the event. It features sections of the report that are protected by legal privilege, allowing companies to be honest with regulators about root causes without creating a public roadmap for class-action litigation.

4.1.3. The Regulator's Request for Information (RFI) (Regulator → AI Labs - Frontier and small):

- **Objective:** Investigation and verification.
- **Design:** A structured legal instrument used to extract verifiable evidence to settle disputes regarding the timeline and scope of an incident.

4.1.4. The Safety Contribution Template (Company → Industry and Public):

- **Objective:** Collective resilience through knowledge sharing and increased societal accountability and trust.
- **Design:** A sanitised, abstracted tool. It strips away proprietary trade secrets and company-specific identifiers to share the "lesson learned" and the "technical fix" with the wider AI community, ensuring other labs do not repeat the same failure.

The detailed proposed forms are in [Appendix C](#).

4.2. Operational Guidelines

To ensure the framework is not merely a set of forms but a functioning system, it defines strict operational parameters:

4.2.1. Reporting Timelines:

¹⁰ Zhu, E., Ghosh, A., & Longpre, S. (2026). AI Needs a Better Way to Report Flaws, So We Built One. MIT AI Risk Initiative. <https://airisk.mit.edu/blog/announcing-flare-ai>

- **Preventive Disclosures** are subject to a **Coordinated Disclosure Window** (typically 30–90 days), allowing labs to patch the flaw before the researcher publishes the findings.
- **Incident Reports** must be filed within a Critical Notification Window (e.g., immediately or within 72 hours from discovery).

4.2.2. Mandatory Remediation Integration:

The Framework requires that every **Preventive Disclosure** and **Incident Report** include a **Remediation Plan**. The plan must specify the technical steps taken (or proposed) to mitigate the risk, ensuring that the reporting process leads directly to a safety improvement rather than just a documented failure.

4.2.3. Public repository for the Preventive Disclosure Form and the Safety Contribution Template: the proposal includes the creation of a website, functioning as a database to keep all the relevant stakeholders informed.

4.3. Expected Outcomes

By implementing this process, the AI ecosystem achieves three critical goals:

4.3.1. Proactive Risk Mitigation: By shifting focus toward preventive disclosure and mandatory remediation plans, and by incentivising researchers to submit those plans to CISA / ISO / OECD / Frontier Labs, the framework moves the industry from reactive crisis management to patching flaws before they manifest or are exploited.

4.3.2. Systemic Standardisation: The framework provides a clear, predictable methodology for regulators and companies, replacing ad-hoc communication with a standardised global protocol for AI incident management. By replacing generalised reporting with specialised, incentive-aligned forms, the framework reduces the legal and bureaucratic friction currently hindering disclosure, providing a standardised, global methodology for incident management.

4.3.3. Rapid integration into frontier models: By providing a secure, structured pipeline for researchers to alert CISA / ISO / OECD / Frontier Labs, the framework ensures that critical vulnerabilities discovered in academia are rapidly integrated into frontier model defences.

5. Discussion and Limitations

5.1 Broader Implications for AI Safety

Our results suggest that AI safety cannot rely on the altruism of frontier labs or the hope that companies will "do the right thing" for the public good. Instead, safety infrastructure must be designed around the existing incentives of the stakeholders.

By decoupling the **Preventive**, **Liability**, and **Contribution** streams, we align the reporting mechanism with the user's goal:

- Researchers are incentivised by impact and credit (via the Preventive Template).
- Corporations are incentivised by risk mitigation and legal compliance (via the Liability Template).

5.2 Limitations

Due to the timeframe of this hackathon, these templates were developed based on theoretical regulatory requirements and legal logic rather than being piloted within a frontier AI lab or a government agency. The "friction" of filling out these forms has not been measured.

Unaddressed Threat Models and Failure Modes

The Malicious Insider: This framework assumes that the "Stabilisers" (CISA, ISO, etc.) are secure.

The "Paper Safety" Trap: There is a risk that companies may use the *Liability-Aware Template* as a performative gesture, filling out the forms perfectly to create a "paper trail" of compliance while failing to implement actual technical safety changes.

5.3 Explicit and Implicit Assumptions

Our framework relies on three critical assumptions:

Regulatory Authority: We assume that regulators possess the legal mandate to issue RFIs that companies are compelled to answer. If this authority is weak or non-existent, the RFI process collapses.

The "Safe Harbour" Assumption: We implicitly assume that using the *Preventive Disclosure Template* would grant researchers some level of "Safe Harbour" or legal protection against DMCA/CFAA lawsuits from the companies they are alerting.

The "Legal Privilege Assumption: The Liability-Aware Report's Privileged Annexes assume that a company can shield the root-cause sections of a regulatory filing under legal privilege. This varies by jurisdiction and may not survive a subpoena from a regulator or a plaintiff, in which case the confidentiality the instrument promises does not hold.

Inter-Agency Cooperation: We assume that CISA, ISO, and the OECD can coordinate their responses without creating bureaucratic bottlenecks that delay the patching of critical flaws.

5.4 Future Work

The current work is theoretical architecture. The most logical next step is to prove that these forms actually change the behaviour in the AI ecosystem.

Retrospective Case Studies: Apply the framework to 5–10 past AI incidents, analyse whether the *Liability-Aware* and *Safety Contribution* templates would have extracted more technical detail than the reports that were actually released.

User-Experience (UX) Simulation: Conduct a study with a group of AI researchers and corporate legal counsels. Measure the volume of detail and the willingness to disclose sensitive information between the two groups.

Automated Routing Logic: Develop an AI-driven triage system that analyses a *Preventive Disclosure* and automatically routes it to the correct stakeholder (e.g., routing a policy-level flaw to the OECD and a technical vulnerability to CISA).

Policy and Legal Formalisation: A form is only as good as the law that supports it. In order to implement the framework efficiently, we propose a legal framework where governments agree that any researcher using the *Preventive Disclosure Form* is granted temporary immunity from the CFAA (Computer Fraud and Abuse Act)¹¹ or DMCA¹², provided they follow the coordinated disclosure window¹³. Additionally, work toward integrating these forms into existing laws and policy frameworks, such as the

¹¹ United States Congress (1986). Computer Fraud and Abuse Act, 18 U.S.C. § 1030. Legal Information Institute, Cornell Law School. <https://www.law.cornell.edu/uscode/text/18/1030>

¹² United States Congress (1998). Digital Millennium Copyright Act, 17 U.S.C. § 1201. Legal Information Institute, Cornell Law School. <https://www.law.cornell.edu/uscode/text/17/1201>

¹³ Longpre, S., Kapoor, S., Klyman, K., et al. (2024). A Safe Harbor for AI Evaluation and Red Teaming. arXiv:2403.04893. <https://arxiv.org/abs/2403.04893>

EU AI Act or the US federal AI framework (currently America's AI Action Plan, 2025)¹⁴, moving them from "suggested templates" to "mandated reporting standards."

Taxonomic Refinement: Refine how the framework defines "Criticality" and "Risk" by developing a CVSS (Common Vulnerability Scoring System)¹⁵ that provides a numerical score for flaws. As the CVSS is usually a cybersecurity practice, a natural extension would be to create an **AI-CVSS** that measures the impact of the incident based on model capability, accessibility and potential for systemic harm.

6. Conclusion

Our main finding is that generalised AI reporting systems fail because they force a choice between transparency and legal safety. To solve this, we developed a framework which replaces a single universal form with four targeted instruments: a Preventive Disclosure form for researchers, a Liability-Aware report for companies, a Regulator's RFI, and a Safety Contribution template for industry sharing. This architecture ensures the right information reaches the right stakeholders without compromising proprietary secrets or increasing legal risk.

The implication is a fundamental shift in how we handle AI failures: we move from simply recording "what went wrong" to actively preventing "what could go wrong." By aligning the reporting process with the real-world incentives of the people using it, the framework removes the friction that currently prevents the disclosure of critical flaws. This results in a more transparent, resilient AI ecosystem where researcher insights are rapidly integrated into frontier defences, and regulators can ensure accountability without stifling innovation.

¹⁴ The White House (2025). America's AI Action Plan. Executive Office of the President.
<https://www.ai.gov/action-plan>

¹⁵ FIRST (Forum of Incident Response and Security Teams) (2024). Common Vulnerability Scoring System (CVSS). FIRST. <https://www.first.org/cvss/>

Appendix A

Criterion	Report for Serious Incidents under the AI Act (General-Purpose AI Models with Systemic Risk)	Governor's Office of Emergency Services (Cal OES) Secure Submission Portal
Scope	Serious incidents: art.73 point 49	Critical safety incidents and catastrophic risks: Bill n.53, chapter 25.1, point 22757.11, paragraph c and d
Legal basis	AI Act Art. 73 + GPAI Code of Practice (Commitment 9)	SB 53 (TFAIA), filed to the Office of Emergency Services
Who must file	Provider of a high-risk or GPAI systemic-risk model	Frontier model developer above the covered threshold
Recipient	National competent authority / AI Office	Office of Emergency Services (state)
What triggers a report	Defined categories of serious harm to health, rights, property, or critical infrastructure, including major cyber breaches	Narrowly defined "critical safety incident": catastrophic harm, loss of control, weight theft, or deception subverting controls
Severity bar	Serious but not necessarily catastrophic harm	Very high; roughly 50+ casualties or ~\$1B damage for the catastrophic limb
Reporting timeline	Quote from art.73 "The report shall be made immediately after the provider has established a causal link between the AI system and the serious incident or the reasonable likelihood of such a link, and, in any event, not later than 15 days after the provider or, where applicable, the deployer, becomes aware of the serious incident."	A frontier developer shall report any critical safety incident pertaining to one or more of its frontier models to the Office of Emergency Services within 15 days of discovering the critical safety incident. (2) If a frontier developer discovers that a critical safety incident poses an imminent risk of death or serious physical injury, the frontier developer shall disclose that incident within 24 hours to an authority, including any law enforcement agency or public safety agency with jurisdiction, that is appropriate based on the nature of that incident and as required by law.
Report depth	Rich: about 9 items, including chain of events, root cause, and remediation	Thin: about 4 fields; functions as a tripwire, not an investigation
Report structure	Staged: initial, intermediate, and final reports	Single filing
Public disclosure	Held by government; may surface later as an aggregate or redacted summary	Confidential; feeds annual aggregate summaries
Captures near-misses caught during testing	No	No
Interoperability (routing to labs,	None. A filing in Brussels does not reach Sacramento, the lab, or the AI	None. A filing in Sacramento does not reach the EU or the databases

other regulators, databases)	Incident Database	
------------------------------	-------------------	--

Appendix B

	Google BugHunters AI VRP	OpenAI Bugcrowd	Anthropic HackerOne	ODIN (Mozilla)
Reporting platform	Own BugHunters / VRP	Bugcrowd (third-party)	HackerOne (third-party)	AI-specific platform
Scope	Security vulnerabilities in Google's AI products, incl. weight extraction and prompt injection; content violations routed elsewhere	Security vulnerabilities in OpenAI products (APIs, ChatGPT); excludes content	Security vulnerabilities in Anthropic's models and services	AI-related failures and vulnerabilities across multiple vendors
Access	Public bounty	Public (Bugcrowd)	Public VDP form; separate model-safety bounty is invite-only	Public
Transparency	Public program, discoverable to security researchers	Researchers often unaware; overlaps with Report Content	Researchers often unaware; safety bounty is invite-only (weakest access)	Public AI-specific platform, relatively discoverable
Information collection	Extra detail via open-ended, non-standardized fields, which limits actionable intake; field count not specified in text	Sparsest in the survey (7 fields): low reporter burden, thin triage	Not specified in text	Not specified in text

Appendix C — Proposed reporting forms

Blank templates for the forms described in Section 4.

C.1. Preventive Disclosure Form (PDF)

Purpose: Formal notification of a flaw of an AI System.

Proposed reporting form

Preventive Disclosure Form

User: Independent Researcher **Recipient:** CISA / ISO / OECD / Frontier Labs

Contact and system

Researcher name and contact

Model or system, including version

When was the flaw detected?

YYYY-MM-DD



Stage of development

☐ Research / design ☐ Training / fine-tuning ☐ Testing / evaluation ☐ Deployed ☐ Not known

The flaw

Type of issue

☐ Prompt injection ☐ Weight leak ☐ Reward hacking ☐ Other

Steps to reproduce / proof of concept

Include test conditions and evidence or attachment references.

Possible impact

☐ Low ☐ Medium ☐ High ☐ Catastrophic ☐ Not known

Who or what could be affected?

Could this affect other systems?

Remediation

Propose the remediation plan

Coordination or routing needs

C.2. Liability-Aware Incident Report (LAIR)

*Purpose: Formal notification of an incident to meet regulatory obligations while managing legal risk.
Confidentiality: Privileged and Confidential (Attorney-Client Privilege).*

Proposed reporting form

Liability-Aware Incident Report

User: Frontier Lab **Recipient:** Governmental Regulator

Report details

Company and contact

Report type

☐ Initial ☐ Update ☐ Final

Affected model or system

Incident Number

Legal basis for reporting / jurisdiction

Incident details

Incident Rating

☐ Low ☐ Medium ☐ High ☐ Catastrophic ☐ Not known

Time detected and time zone

Containment status

☐ Contained ☐ Partial ☐ Ongoing ☐ Not known

Next follow-up date and reason

What happened?

Known effects and people or systems affected

Response and follow-up

Actions taken and remediation plan

Include the responsible person, target dates, and how the fix will be checked.

Evidence

Confidentiality request and legal basis

Specialization-Based Incident Management Framework

C.3. Regulator's Request for Information (RFI)

Purpose: Formal request of information of an incident.

Confidentiality: Privileged and Confidential.

Proposed reporting form

Request for Information

User: Governmental Regulator **Recipient:** AI Lab

Request details

Issuing authority and contact

Company and contact

Request and incident references

Response due / time zone

Legal authority, jurisdiction, and response rights

Information requested

Purpose, scope, and period covered

Root cause and event timeline

Containment, model access, and persistence

Related incidents and internal warnings

Records, evidence, and verification required

Specify originals, custodians, collection times, and any missing records.

C.4. Safety Contribution Template (SCT)

Purpose: Sharing a solution to a discovered flaw to improve industry-wide safety.

Confidentiality: Sanitized (Proprietary data removed).

Proposed reporting form

Safety Contribution Template

User: Frontier Lab **Recipient:** Public

Contribution details

Contributor and contact

Reference

Affected technology or system type

What others can learn

Issue summary

How the failure occurred

Technical fix

How the fix was tested and the results

Where the fix applies / limits and remaining risks

Supporting Information

Materials to share

☐ Dataset ☐ Test or benchmark ☐ Code ☐ Other ☐ None

Supporting references or files

Implementation cost or overhead

Release reviewer and approval date

Repository and next review

Specialization-Based Incident Management Framework

LLM Usage Statement

We used ChatGPT-6 Astra Ultra and Anthropic Fable 5.1 to help format the forms in Appendix C and to check spelling and grammars in some sections.