
CONVENTION FOR EARLY NOTIFICATION OF AI INCIDENTS

Mariana Zuluaga Abril

Tomas Antolinez Cuadros

With
Apart Research

Abstract

Rapid AI advancement and rising international security concerns require robust tools to mitigate catastrophic risks. Highlighting the limitations of current AI regulations, this paper proposes the adoption of a Convention for Early Notification of AI Incidents, inspired by the 1986 Convention on Early Notification of a Nuclear Accident, which solved a similar problem for radioactive releases. Across nineteen articles, our proposal defines a notification mechanism to ensure rapid information sharing and coordinated action between governments, developers, and infrastructure operators. This study offers a concrete legal instrument to enhance global cooperation and manage AI incidents effectively.

1. Introduction

The growing international uncertainty surrounding the security of artificial intelligence has highlighted the urgent need for operational tools capable of responding effectively to current incidents and of developing resources that can mitigate catastrophic risks. The 1986 Convention on Early Notification of a Nuclear Accident addressed radioactive releases through a notification mechanism, providing a rapid and effective response to a problem that posed a high risk at the time; for this reason, it was used as a reference.¹ From a regulatory response perspective, the Convention for Early Notification of AI Incidents is proposed as a notification mechanism designed to address the shortcomings identified in current regulations. The proposed treaty consists of nineteen articles structured to provide an operational and legally binding framework that ensures prompt notification,

¹ Convention on Early Notification of a Nuclear Accident, adopted 26 September 1986, in force 27 October 1986, 1439 UNTS 275; IAEA doc. INFCIRC/335.

the effective flow of information, and a clear path for coordinated action among developers, infrastructure operators, and governments.

2. Related Work

Incident reporting for AI is addressed by two main instruments. First, article 55(1)(c) of the EU AI Act, operationalized by Commitment 9 of the Code of Practice and by the Commission’s serious-incident reporting template for models with systemic risk.² Second, section 22757.13 of the California statute, whose four categories of critical safety incident include loss of control and deceptive subversion of the developer’s controls.³

The EU AI Act reveals fundamental regulatory shortcomings that severely limit the law’s practical effectiveness. First, Article 3(63) explicitly excludes models intended for research, development, or prototyping prior to commercialization, creating a preventive blind spot that allows models with safety biases or systemic risks to be trained entirely outside the scope of regulatory oversight. Second, Article 3(49) narrowly defines a ‘serious incident’ through a limited list of harms, without recognizing loss of control or deceptive subversion as grounds for notification in themselves.⁴ This limitation undermines the alert system set forth in Article 73, which requires incidents to be reported immediately and sets maximum deadlines of two days for failures in critical infrastructure, ten days in the event of death, and fifteen days for violations of fundamental rights. By omitting control anomalies or autonomous deception from this list, companies are not required to meet these urgent deadlines in the face of unpredictable model behavior, allowing critical misalignment failures to remain hidden behind corporate opacity until physical or material damage has already occurred.

On the other hand, under the California Transparency in Frontier Artificial Intelligence Act (SB 53) a frontier developer must report a critical safety incident within fifteen days of discovery and transmit periodic summaries of catastrophic-risk assessments from internal use.⁵ This regulation contains procedural ambiguities regarding the scope of internal evaluation, comparative testing, and pre-deployment research phases, due to the lack of precise criteria for internal containment breaches during pre-market development.

Additionally, it lacks standardized emergency response protocols and direct alert mechanisms between the public and private sectors. In the absence of a real-time crisis management framework, the Office of Emergency Services (Cal OES) is limited to receiving audits and quarterly reports.⁶ The result is a regulatory framework that is ineffective at responding to incidents and reporting during the critical first hours.

3. Methods

² Regulation (EU) 2024/1689, art. 55(1)(c); European Commission, General-Purpose AI Code of Practice (2025), Commitment 9; European Commission, reporting template for serious incidents involving general-purpose AI models with systemic risk (2025).

³ Cal. Bus. & Prof. Code § 22757.11(d)(1) to (4).

⁴ Regulation (EU) 2024/1689, arts. 3(49), 3(63) and 73(1) to (4).

⁵ Cal. Bus. & Prof. Code §§ 22757.13(c)(1) and 22757.12(d), added by Stats. 2025, ch. 138 (SB 53), operative 1 January 2026; OpenAI meets the definition of large frontier developer in § 22757.11(j).

⁶ Cal. Bus. & Prof. Code §§ 22757.12(d) and 22757.13(a) to (e), added by Stats. 2025, ch. 138 (SB 53).

This paper analyzes the existing regulation and then uses a case study to explain how other instruments have closed similar gaps in the past. The evidentiary base is documentary: the text of the Convention and the literature on the regime and on treaty effectiveness. The Convention is disaggregated into its design variables, and its negotiation history is analyzed to identify what produced rapid agreement. Following the logic of structured comparison, its provisions are assessed against AI governance: duty bearers, observability of the triggering event, existence of a host institution, and distribution of capabilities. Conditions that hold support transfer. New conditions specific to the AI context are then added.

4. Results

The Convention on Early Notification of a Nuclear Accident was adopted five months after Chernobyl and entered into force one month later.⁷ Some key features of its design account for that speed. The obligation is thin: the Convention centers on creating a duty to notify and to provide information. Entry into force was cheap: three States sufficed, and provisional application was available on signature.

That is why incident monitoring is one of the few areas where the benefits of internationalization are high: “we find that the need to internationalise incident monitoring is high.”⁸ The reason is the transboundary nature of the harm. Incidents “are unlikely to be confined only to that jurisdiction”, and one occurring in a given country “could be an early warning sign that the same model may soon cause harm in other jurisdictions.”⁹

For that reason, “a well-coordinated global incident monitoring system could serve as an early alert mechanism to prevent the escalation and spread of AI-related accidents across borders.”¹⁰ Interoperability, however, requires a common classification of harms, since only then can reports from different jurisdictions be pooled and recurring failures identified.¹¹ Governance capacity is also uneven: many countries lack the funding and technical expertise to independently monitor harms to their citizens, a gap that internationalization can help close.¹²

To achieve that, the AI Treaty Momentum Index (ATMI) recommends keeping the “plumbing on the shelf”: ready-to-file clauses and a defined adoption pathway. In its most geopolitically strained scenario, the advice is to pursue a coalition-of-the-willing treaty path with depositary pre-commitment, so as to bypass full consensus, and to invest in structured incident reporting.¹³

The following table compares the Convention on Early Notification of a Nuclear Accident (1986) and the proposed Convention for Early Notification of AI Incidents, detailing which articles were retained, modified, or added. This comparative analysis examines how the provisions of the nuclear

⁷ Convention on Early Notification of a Nuclear Accident, arts. 12 and 13; International Atomic Energy Agency, The 1986 Chernobyl nuclear power plant accident.

⁸ Claire Dennis and others, What Should Be Internationalised in AI Governance? (Oxford Martin AI Governance Initiative and Centre for the Governance of AI, 2024).

⁹ Ibid.

¹⁰ Ibid.

¹¹ Ibid.

¹² Ibid.

¹³ Oksana Kotelnikova and Sofiia Lobanova, AI Treaty Momentum Index (ATMI) (Apart Research, 2025).

model were adapted to the technical specifics of AI, justifying the operational tools necessary to ensure an effective response to each type of cross-border risk.

Table 1. Correspondence between Conventions

Art.	Convention on Early Notification of Nuclear Accidents	Convention for Early Notification of AI Incidents	Change
1 → 1	Accident at listed facilities; release occurs or is likely; transboundary radiological significance	Incident in listed activities (a to f); loss of containment occurs or is likely; definitions (3); market placement and harm irrelevant (4)	Adapted; 3, 4 added
2 → 2	Forthwith notify affected States and Agency; promptly provide information	Same, through the United Nations; (c) developers report to the competent authority and notify affected persons and entities; (d) forthwith means 24 hours, promptly 72 hours	Retained; (c), (d) added
3 → 3	Other nuclear accidents	Other AI incidents	Retained
4 → 4	Agency functions	Same, for the Organization; (c) executing body designated by the General Assembly, with a continuous focal point	Retained; 4(c) added
5 → 5	Eight data items on the release	Nine items on the loss of containment; (e) copies, weights, credentials and persistence; (f) forensic analysis replaces environmental monitoring; meteorological item dropped	Adapted; (e) added
6 → 6	Consultations requested by an affected State Party	Consultations requested by a third State Party	Adapted
7 → 7	Contact points	Same; (4) developer contact point registered with the competent authority	Retained; 7(4) added
8 → 8	Feasibility studies on radiological monitoring for States without nuclear activities bordering a non-Party	Assistance to any State Party that may be affected, to receive notifications and to detect and contain effects	Adapted
n/a → 9	None	Designated evaluators with continuing access to development environments, containment and monitoring arrangements and their records	Added
n/a → 10	None	Meeting of the States Parties every two years; non-compliance procedures to be approved at the first meeting	Added
9, 10, 11, 13 → 11, 12, 13, 15	Arrangements; other agreements; disputes; provisional application	Same	Retained; renumbered
12 → 14	Signature in Vienna and New York; entry into force after three States	Signature at United Nations Headquarters in New York; rest unchanged	Adapted
14 → 16	Amendment by protocol	Same; (4) Annex I amendable by two-thirds majority, in force after 180 days, with opt-out	Retained; 16(4) added
15 → 17	Denunciation at any time, effective after one year	Denunciation available after two years; (3) articles 2, 5 and 6 survive for incidents occurring before it takes effect	Adapted; 17(3) added
16, 17 → 18, 19	IAEA Director General as depositary; authentic texts	United Nations Secretary-General as depositary; same six languages	Adapted
n/a → Annex I	None	Criteria for advanced AI systems: 10 ²⁵ operations, national designation as frontier or systemic-risk model, or autonomy threshold	Added

Adapting the Convention on Early Notification of a Nuclear Accident to incidents related to artificial intelligence produced the notification mechanism proposed here, which aims to provide a relevant input that effectively addresses the challenges arising in connection with AI. Below is the draft of

the [*Convention on Early Notification of AI Incidents*](#), which is proposed as a proactive response to the shortcomings identified in current AI regulations:

5. Discussion and Limitations

Publishing a drafted treaty with its articles is itself a dual-use act. A specific text is easier to evaluate and easier to file, but it could also give actors with an interest in blocking a binding regime an early target.

The analysis is documentary and rests on a single case. Structured comparison of the 1986 Convention with the AI context establishes congruence between design variables and outcome, not causal identification. There is no counterfactual and no variation on the dependent variable, since the case was selected because it succeeded. What the paper offers is therefore a plausible account of which conditions made rapid agreement possible, not a test of whether any of them was necessary.

One of the challenges is to foster global cooperation so that countries will be interested in implementing this reporting mechanism; therefore, it is important to develop effective strategies to promote widespread adoption.

Future work proceeds along three lines. The first is a roadmap for adoption. The second is a draft UNGA resolution opening the Convention for signature. The third is the design of subsequent instruments for the remaining governance functions where internationalization yields comparable benefits.

6. Conclusion

Comparative regulatory analysis shows that the governance frameworks examined, the European Union's AI Act and California's SB 53, have limitations and shortcomings when it comes to mitigating catastrophic risks and providing an immediate response to critical incidents. Given these regulatory gaps, there is a need to develop new tools. This mechanism proposes an operational and binding framework to coordinate rapid action among developers, operators, and government agencies on an international scale, thereby generating a positive impact in the field of AI safety.

References

Sources cited in the text

1. California. Business and Professions Code, Division 8, Chapter 25.1 (Transparency in Frontier Artificial Intelligence Act), §§ 22757.10 to 22757.15, added by Stats. 2025, ch. 138 (SB 53), operative 1 January 2026.
https://california.public.law/codes/business_and_professions_code_section_22757.13
2. Convention on Early Notification of a Nuclear Accident (1986). Adopted 26 September 1986, in force 27 October 1986. 1439 UNTS 275; IAEA doc. INFCIRC/335.
<https://www.iaea.org/topics/nuclear-safety-conventions/convention-early-notification-nuclear-accident>. Spanish text: OJ L 314, 30.11.2005, p. 22.

3. Dennis, C., Clare, S., Hawkins, R., Simpson, M., Behrens, E., Diebold, G., Kara, Z., Wang, R., Trager, R. F., Maas, M. M., Kolt, N., Anderljung, M., Pilz, K., Reuel, A., Murray, M., Heim, L., and Ziosi, M. (2024). What Should Be Internationalised in AI Governance? Oxford Martin AI Governance Initiative and Centre for the Governance of AI. <https://www.oxfordmartin.ox.ac.uk/publications/what-should-be-internationalised-in-ai-governance>
4. European Commission (2025a). The General-Purpose AI Code of Practice. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
5. European Commission (2025b). 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>
6. International Atomic Energy Agency (n.d.-b). The 1986 Chornobyl nuclear power plant accident. <https://www.iaea.org/topics/chornobyl>
7. Kotelnikova, O., and Lobanova, S. (2025). AI Treaty Momentum Index (ATMI). Apart Research. <https://apartresearch.com/project/ai-treaty-momentum-index-atmi-lst4>
8. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (AI Act). OJ L, 2024/1689, 12.7.2024.

Other sources consulted and instruments cited in the draft Convention

9. Convention on Nuclear Safety (1994). IAEA doc. INFCIRC/449.
10. Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 (NIS 2 Directive). OJ L 333, 27.12.2022, p. 80. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02022L2555-20221227>
11. International Atomic Energy Agency (1984). Guidelines for Mutual Emergency Assistance Arrangements in Connection with a Nuclear Accident or Radiological Emergency. INFCIRC/310.
12. International Atomic Energy Agency (1985). Guidelines on Reportable Events, Integrated Planning and Information Exchange in a Transboundary Release of Radioactive Materials. INFCIRC/321.
13. International Atomic Energy Agency (2011). IAEA Action Plan on Nuclear Safety. GOV/2011/59-GC(55)/14. <https://www.iaea.org/topics/nuclear-safety-action-plan>
14. International Atomic Energy Agency (2025). Building a safer world: 20 years of the Incident and Emergency Centre. IAEA Bulletin. <https://www.iaea.org/bulletin/building-a-safer-world-20-years-of-the-incident-and-emergency-centre>
15. International Atomic Energy Agency (n.d.-a). International operational arrangements. <https://www.iaea.org/topics/emergency-preparedness-response/international-operational-arrangements>

16. International Law Commission (2001). Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with commentaries. Yearbook of the ILC 2001, vol. II, Part Two. https://legal.un.org/ilc/texts/instruments/english/draft_articles/9_7_2001.pdf
17. KQED (2026). How OpenAI's models escaped their sandbox and slipped past California's AI law. <https://www.kqed.org/news/12092162/how-openais-models-escaped-their-sandbox-and-slipped-past-californias-ai-law>
18. 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/>
19. Montreal Protocol on Substances that Deplete the Ozone Layer (1987). 1522 UNTS 3.
20. OECD (2025). Towards a common reporting framework for AI incidents. OECD Publishing. https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html
21. OECD Nuclear Energy Agency (2002). Chernobyl: Assessment of Radiological and Health Impacts. 2002 Update of Chernobyl: Ten Years On. Paris: OECD.
22. OECD Nuclear Energy Agency and International Atomic Energy Agency (2006). International Nuclear Law in the Post-Chernobyl Period. Paris: OECD. <https://oecd-nea.org/upload/docs/application/pdf/2019-12/nea6146-iaea-chernobyl.pdf>
23. OpenAI (2026). The Hugging Face incident and the road ahead. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
24. Regulation (EU) 2016/679 (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1.
25. TNW (2026). OpenAI asks California to strengthen SB 53 after the hack. <https://thenextweb.com/news/openai-sb-53-california-ai-safety-law-amendments>
26. United Nations (1986). Yearbook of the United Nations 1986, Part Three: International Atomic Energy Agency. https://cdn.un.org/unyearbook/yun/pdf/1986/1986_1111.pdf
27. United Nations (1999). Summary of Practice of the Secretary-General as Depositary of Multilateral Treaties. ST/LEG/7/Rev.1.
28. United Nations General Assembly (1997). Resolution 51/229, Convention on the Law of the Non-Navigational Uses of International Watercourses.
29. United Nations General Assembly (2004). Resolution 59/38, United Nations Convention on Jurisdictional Immunities of States and Their Property.
30. United Nations General Assembly (2019). Resolution 74/247, Countering the use of information and communications technologies for criminal purposes.
31. United Nations General Assembly (2024). Resolution 79/243, United Nations Convention against Cybercrime. <https://documents.un.org/doc/undoc/gen/n24/426/74/pdf/n2442674.pdf>

32. United Nations General Assembly (2025). Resolution 79/325, Terms of reference and modalities for the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance. <https://un.org/independent-international-scientific-panel-ai>
33. United Nations Office on Drugs and Crime (2025). UN Convention against Cybercrime opens for signature in Hanoi, Viet Nam. <https://www.unodc.org/unodc/en/press/releases/2025/October/un-convention-against-cybercrime-opens-for-signature-in-hanoi--viet-nam.html>
34. Vienna Convention on Civil Liability for Nuclear Damage (1963). 1063 UNTS 265; IAEA doc. INFCIRC/500.
35. Vienna Convention on the Law of Treaties (1969). 1155 UNTS 331.