
Limitations in existing regulatory mechanisms when applied to attacks by AI agents¹

Juan Jeronimo Manriquez
Independent

Tomás Cifuentes Clavijo
Independent

With
Apart Research

Abstract

Current reporting mechanisms with investigative or regulatory capabilities have deficiencies that make reporting any autonomous AI agent related incidents to date—including the Hugging Face and DSEwiki incidents— infeasible for third parties. This conclusion was reached after conducting an investigation into the two main western AI related legal regimes, with California's SB 53 and the EU's AI Act being examined, locating three main fault points in these systems. These systems are confidential or do not publish a public record; and fail either by not accepting reports from non-providers, or by having a threshold for what qualifies as an incident that disqualifies the incidents that have already happened. An exploration of the existing literature related to these incidents corroborated that these holes are real and significant. Two main artifacts were created. A complementary website that helps visualize where these systems fail, and a list of proposals that would amend the systems and strengthen them against future attacks carried out by AI agents.

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

1. Introduction

As LLMs grow more capable, the capability of agents and agent swarms to carry out cyberattacks has also grown. This, in combination with alignment and containment failures, has resulted in multiple public incidents involving agent swarms attacking systems to fulfill a task not explicitly designated by their operators. One case in particular has garnered the most public attention. An agentic swarm originating from OpenAI hacked Hugging Face’s systems [1]. This, amongst other occurrences like the DSEwiki incident [2], the Anthropic PyPI incident [3], the RubyGems breach [4], and an attack on four unnamed victims by OpenAI agents [16], demonstrate the necessity to enhance global response capacity to agentic attacks.

The team’s angle of attack was to examine the current regulatory and reporting mechanisms that a *non-provider* individual or group could use to flag an attack, since as of now, providers have tended to not report incidents without external triggering, with one confirmed filing coming approximately two and a half months later [11]. However, our investigation led us to conclude that the existing reporting mechanisms lack the provisions through which the public can report *any* of the known incidents involving AI systems with currently existing instruments. This is grave enough that even the largest known incident —the Hugging Face intrusion— does not appear to have a legal basis for a legally actionable report filed by a non-provider.

This critical lack of mechanisms in pursuit of justice leads to a dangerous precedent, where the public feels that the system is neither *open* nor *democratic*. This could lead to loss of confidence in legal systems, for the resolution of incidents is made impossible due to the limitations of the law. In particular, the capacity to report incidents through adequate channels is especially important for the Global South. Nations in this region often lack regulation mechanisms that target AI systems, lack the capacity to investigate and mitigate damages efficiently, and may be more vulnerable to these attacks due to a lack of institutional knowledge about AI swarms.

Our main contributions are:

1. The discovery that no known incident applies cleanly to the existing regulatory frameworks. In particular, there is a lack of an actionable pathway in existing AI regulation laws for third parties.
2. Proposals for how existing and future systems can be amended, with the goal of confronting the *current* slate of autonomous attacks carried out by AI agents.

2. Related Work

Existing AI safety and public policy research has already noted that existing legal regimes have deficiencies that may make them harder to utilize. An analysis of the EU’s AI Act has pointed out

that the existing system may lack in analysis of *scope* and the involved *actors*, alongside other deficiencies in the existing incident report system [5]. An incident escalation framework analysis points out limitations in both the EU's and California's SB 53 incident reporting system, noting a gap in relevant aspects, like the exact constitution of what qualifies as an *incident*, and the way the severity of these is classified [6]. Additional work has been done in considering what shapes an AI incident reporting system should take, from both AI specialists [7] and international organizations [8].

Other fields where safety is important have reporting systems in place for incidents. These systems have been examined, and certain aspects of these have been proposed for AI reporting systems [9]. Open systems for reporting AI incidents -like AIID and AVID- exist, but they do not have regulatory capacities or legal requirements.

While existing works have examined report systems and their limitations from many angles, to our knowledge, this report differs from the existing literature in its analysis of interactions and limitations that affect real reporting systems in known cases. This contrasts with existing case reports where the AI systems are the ones being compromised, as the AI in these cases is the perpetrator.

3. Methods

The group's identification of the current holes in the existing legislation originated from the original concept for the project. We aimed to build a system that would assist affected parties in reporting attacks carried out by agents to the relevant regulatory authorities. This required investigating existing reporting systems, their limitations, and what needed to be collected. To this end, we used the DSEwiki incident as a baseline, looking at the options available from the perspective of a *non-provider* actor. In this case, a victim, the wiki's administrator, who manually tried to fight off the attack. We also looked from the perspective of a competent third party -the Nightingale Collective- or a third party with the same information as them.

This led to an examination of the legal frameworks that define the reporting systems that currently exist. Under the current laws, California's SB 53 [17] is an open system, but it only deals with *critical safety incidents* that we did not believe any case met. The EU's AI Act [18] article 55 proved to be problematic too, as it was fileable only by providers of GPAI models with systemic risk. Additionally, it required the case to be a *serious incident*, a category that did not map cleanly onto any incident. Finally, we investigated a public EU complaints mechanism run by the European Commission under article 85 of the AI Act, but it did not meet our needs, as this reporting tool excludes infringements of articles 53 and 55, making it even more limited in scope than the provider only system.

As such, the existing systems as we identified fail either at the identity of the filer or the threshold for filing. Additionally, it appeared that none of these systems released public records. The team

conducted an additional verification pass, looking at any other laws that would apply in these cases. For example, under the classification of *serious incidents* given by the heads of article 3(49) of the EU AI Act, we were not able to find a head that matches the existing cases, keeping it out of scope.

Since this conclusion was only driven by our own analysis, the group conducted an examination of legal analysis surrounding the existing cases. This confirmed our beliefs, showing that amongst experts, the applicability of the existing systems to these incidents was dubious at best. In particular, we found that:

- California's administering agency confirmed that the Hugging Face intrusion did not meet the threshold to be a *critical safety incident*, making it unreportable under SB 53 [10]. Our analysis, and the circumstances of the other cases also lead us to strongly believe these would not classify as *critical safety incidents* either.
- OpenAI did file a report under the EU's AI Act, focusing on the DSEwiki incident [11], publicly calling the incident a "misalignment".
- As of writing this document, the European Commission has declined to state when or under which mechanism the report was filed, or whether it counts as a *serious incident* [11]. External readings of the relevant laws claim that the DSEwiki case is genuinely contestable as a *serious incident*, and that it wouldn't break any other provisions of the AI Act [12].
- The EU's article 85 reporting system as it stands explicitly excludes complaints relating to article 53 and 55. However, the article 85 statute does not show this exclusion, and it is not unthinkable for the system to be reworked [13].
- The EU has additional reporting systems, but these are limited as well, either to downstream providers of AI systems [14], or are made explicitly for whistleblowers [15]. This makes them unsuitable for our purposes, as both are limited by who can file them.
- Every legally actionable reporting system that has been evaluated is confidential or does not publish a public record. This makes the development of public knowledge about AI incidents more complex.

The lack of any accessible systems with regulatory or investigative power invalidated our initial idea and project proposal. It is impossible to build a filing assistance system when there is no regulatory system that accepts reports involving incidents of this magnitude. However, cataloguing where the current systems fail when applied to real cases, and the proposal of potential ways to modify existing or future regulatory mechanisms to close these gaps is of great importance.

4. Results

The most pressing result of our research indicates that the regulatory instruments that exist currently do not provide an appropriate interface for a member of the public to submit a report of an attack by agents or swarms as they exist currently. The systems as they stand fail in at least

one of three main directions, threshold, filer identity, or public disclosure. Either the system has a reporting threshold that excludes the recent wave of attacks, or it is limited in who can file a report. And no legally actionable system emits a public report per incident. This analysis indicates the current criteria of affected parties is lacking for individuals outside of the chain of value of AI enterprises.

The group also developed a website that showcases how current reporting systems fail. This is a complementary visual tool that aggregates the research and our conclusions to make them easily understandable. The site displays where the tools have failed for five important incidents, alongside showing our conclusion for every case so far, and showing a hypothetical incident that would be eligible for reporting.

Case	EU AI Act — Art. 55 serious-incident report	EU AI Act — Art. 85 complaints tool	EU AI Act — Art. 89(2) downstream-provider channel	EU AI Act — whistleblower tool	California SB 53 — public incident report	Public databases (AIID / AVID / OECD AIM)
Hugging Face intrusion (Jul 2026)	Locked	Below bar	Locked	Locked	Below bar	No legal effect
Wiki collusion (May–Jul 2026)	Locked	Below bar	Locked	Locked	Below bar	No legal effect
The four unnamed victims (Jul 2026)	Locked	Below bar	Locked	Locked	Below bar	No legal effect
Anthropic evaluation incidents (Jul + Sep 2026)	Locked	Below bar	Locked	Locked	Below bar	No legal effect
RubyGems / RubyDoc (May 2026)	Locked	Below bar	Locked	Locked	Below bar	No legal effect

Figure 1: A table in the complementary web tool, showing the relation of each case and each reporting system.

Based on these facts, we have constructed the following suggestions, tuned on the results the team obtained, and aimed at making systems more resilient to attacks by AI agents.

- California’s regulatory system must lower the threshold for filing a report with its system significantly, as the current system bar of *critical safety incidents* is exclusionary. Either the threshold for what counts as an incident must be lowered, or a separate reporting system for lower severity incidents must be implemented.
- A particularly important observation is that SB 53 §22757.11(d)(4) indicates that a *critical safety incident* exists when “A frontier model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk.” [17] The current batch of cases demonstrate that dangerous external behaviour can occur during evaluation. We recommend examining this clause in detail, as it excludes a category of incidents during evaluations.
- The EU’s AI Act generally has different thresholds for *serious incidents*, but the bar is still high enough that none of the attacks that have happened so far meet the requirements. Additionally, there is a main reporting system (article 55) that appears to be limited to

providers of GPAI models with systemic risks, and the other reporting systems have their own limitations that make them unsuitable. The threshold for what qualifies as an incident under the AI Act must be evaluated, and a public facing universal reporting system should be implemented to guarantee access.

- We recommend that future legal systems consider an explicit extraterritorial access clause, to permit filing complaints directly with the authorities in a provider's jurisdiction. This would act to mitigate the impunity of damages that occur against jurisdictions that have not developed their own AI protection mechanisms.

5. Discussion and Limitations

Our thesis is that the current reporting systems are not sufficient, limiting *public* access to reporting systems that can deal with incidents involving agents.

It is important to note that third party investigators and capable victims have found a degree of success by publicizing incidents when they occur. However, this preserves a dangerous status quo, as signs of a potentially devastating incident could be missed entirely, or ignored due to lack of public interest, limiting mitigation and investigation from competent mechanisms.

Limitations

We recognize that our analysis of the ways different reporting systems interact with our known incidents is a particular reading of the legal systems these interact with. Some of our claims are backed up by external sources, but the nature of legal systems imposes a degree of uncertainty on our claims. Our suggestions for improving reporting systems are also limited in scope, focusing on a subset of possible improvements to these systems due to time and knowledge constraints.

However, as noted in *Related Work*, this particular field is under active research. Our analysis of the legal institutions involved in these cases may be erroneous, and the contents of this paper are not legal advice.

Future Work

A further examination of this paper would involve working from two angles. First, we recognize that we are only proposing changes to how the existing systems work, and further work would need to evaluate, draft, and apply these changes. Second, our group did not conduct an analysis of proposed report systems. As such, it is necessary to determine if these would apply to the incidents that have happened, and what changes could be made so they would.

6. Conclusion

Current reporting mechanisms with investigative or regulatory capabilities have deficiencies that make reporting any autonomous AI agent related incidents so far infeasible. These deficiencies exist along three axes. The systems do not leave a public record. They limit who can file a report, with an important reporting system being provider only, or limited in some other way. Finally, the thresholds for what applies as an incident exclude the attacks that have happened so far. Our

group identified where these failures happen, and proposed some changes that would make these systems usable by the public when future agentic attacks happen.

Code and Data

Code repository: <https://github.com/JuanM-UR/Sep2026-ApartSprint-Website>

Website: <https://juanm-ur.github.io/Sep2026-ApartSprint-Website/>

References

- [1] METR, ‘Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident’. Aug. 2026. [Online]. Available: <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- [2] Sydney Von Arx, Cormac Slade Byrd, Spencer Kitts, and Thomas Larsen, ‘Discovery of a new OpenAI agent message board’. Sep. 2026. [Online]. Available: <https://collusion.wiki/>
- [3] Anthropic, ‘Investigating Three Incidents in Our Cybersecurity Evaluations’. Jul. 2026. [Online]. Available: <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
- [4] Spencer Kitts, Thomas Larsen, and Sydney Von Arx, ‘OpenAI agents carried out an undisclosed cyber-attack on RubyGems’. Sep. 2026. [Online]. Available: <https://www.rubyhack.ai/>
- [5] H. K. Sidhu *et al.*, ‘Open Problems in AI Incident Governance’. 2026. [Online]. Available: <https://arxiv.org/abs/2607.05163>
- [6] F. Gomez, M. Ball, M. Harre, L. Preston, J. Schwab, and C. Machado, ‘Designing escalation criteria for international AI incident response: criteria, triggers, and thresholds’. 2026. [Online]. Available: <https://arxiv.org/abs/2604.23183>
- [7] K. Wei and L. Heim, ‘Designing Incident Reporting Systems for Harms from General-Purpose AI’, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 44, pp. 38016–38029, Mar. 2026, doi: [10.1609/aaai.v40i44.41139](https://doi.org/10.1609/aaai.v40i44.41139).
- [8] OECD, ‘Towards a Common Reporting Framework for AI Incidents’, OECD Publishing, Paris, OECD Artificial Intelligence Papers 34, Feb. 2025. doi: [10.1787/f326d4ac-en](https://doi.org/10.1787/f326d4ac-en).
- [9] Michael Chatzipanagiotis, ‘Incident Reporting and Investigation Under the AI Act: Some Insights from Aviation’, *International Journal of Law and Information Technology*, 2026, doi: [10.1093/ijlit/eaaf019](https://doi.org/10.1093/ijlit/eaaf019).
- [10] C. DiFelicianantonio, ‘OpenAI calls for stronger AI laws in California’. Politico, Aug. 21, 2026. [Online]. Available: <https://www.politico.com/news/2026/08/21/openai-calls-for-stronger-ai-laws-in-california-0104>

[11] Reuters, ‘OpenAI Has Sent EU Incident Report on Hijacked German Website, Commission Says’. Sep. 2026. [Online]. Available:

<https://www.reuters.com/business/openai-has-sent-eu-incident-report-hijacked-german-website-commission-says-2026-09-07/>

[12] Cloud Security Alliance, ‘OpenAI’s Wiki Silence Tests the EU AI Act’s Incident Regime’. 2026. [Online]. Available:

<https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-incident-disclosure-gap-eu-ai-act-20260/>

[13] artificialintelligenceact.eu, ‘Enforcement of Chapter V Under the EU AI Act’. 2026. [Online]. Available: <https://artificialintelligenceact.eu/enforcement-of-chapter-v-under-the-eu-ai-act/>

[14] European Commission, ‘Complaints Channel for Downstream Providers Using General-Purpose AI Models’. 2026. [Online]. Available:

<https://digital-strategy.ec.europa.eu/en/miscellaneous/complaints-channel-downstream-providers-using-general-purpose-ai-models>

[15] European Commission, ‘AI Act Whistleblower Tool’. 2026. [Online]. Available:

<https://digital-strategy.ec.europa.eu/en/policies/ai-act-whistleblower-tool>

[16] The Guardian, ‘Rogue OpenAI agent that hacked startup tried to attack other firms’. Jul. 29, 2026. Accessed: Sep. 13, 2026. [Online]. Available:

<https://www.theguardian.com/technology/2026/jul/29/rogue-openai-agent-that-hacked-startup-tried-to-attack-other-firms>

[17] California Legislature, ‘Transparency in Frontier Artificial Intelligence Act, SB 53’. 2025. [Online]. Available:

https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

[18] European Parliament and Council of the European Union, ‘Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)’, *Official Journal of the European Union*. 2024. [Online]. Available:

<https://op.europa.eu/en/publication-detail/-/publication/d79f3e5d-41bc-11f0-b9f2-01aa75ed71a1/language-en>

LLM Usage Statement

GLM-5.3-flash and Deepseek-V4.1-flash were used during the investigation phase to help find relevant sources and to brainstorm. Additionally, these two models were also employed to help

create the web tool, with the web tool containing LLM written text. Every output and claim from these models was independently verified by the team to ensure their accuracy.

These models were used to spellcheck and point out mistakes in this paper, as both authors are non-native English writers. However, no LLM written text was included in this document, nor were these tools used to draft sections of it. The final version of this submission was verified and edited by the team.