

Project title: Proposals for Improving AI Incident Reporting Templates

Description: I reviewed the AI incident reporting templates *TFAIA Critical Safety Incident Report* and *Report for Serious Incidents involving General-Purpose AI Models with Systemic Risk*, identified their weaknesses, and sent letters to the relevant authorities proposing improvements (copies of the sent emails are attached below).

Letter to the European Commission

Hello,

I am an AI safety enthusiast, and I would like to suggest several small improvements to the *Report for Serious Incidents involving General-Purpose AI Models with Systemic Risk* template. Here is my list of suggestions:

1. Add the date when the incident was discovered.

The current template asks for the start and end dates of the incident, but not when the provider or independent researchers first became aware of it. Since Article 55 requires serious incidents to be reported “without undue delay,” this distinction could help differentiate late detection from late reporting.

For example, although the attack on RubyGems occurred on May 11, information about the attack was only publicly disclosed on September 11 by independent researchers.

Source: <https://www.rubyhack.ai/>

2. Add information about the AI system in addition to the model involved.

In the case of autonomous agents, behavior is determined not only by the underlying model, but also by the agent harness, available tools, permissions, and other parts of the system. I suggest keeping the existing *Model involved* field while adding a separate section describing the broader AI system.

3. Add information about reproducibility.

Is the incident reproducible? If so, has reproduction already been attempted? How many attempts were made, and in what proportion of runs did the AI system exhibit the same behavior?

4. Add a reporting basis / incident classification field.

The form could include checkboxes corresponding to the relevant AI Act reporting criteria, for example:

- Article 3(49)(a): death or serious harm to health;
- Article 3(49)(b): serious and irreversible disruption of critical infrastructure;
- Article 3(49)(c): infringement of obligations under Union law intended to protect fundamental rights;
- Article 3(49)(d): serious harm to property or the environment;
- Serious cybersecurity breach related to the GPAI model or its infrastructure;
- Uncertain / other.

I hope these suggestions may be useful. I would be grateful if you could consider them in future updates to the reporting template.

Kind regards,
Marko Prakhov-Donets

Letter to the California Cybersecurity Integration Center

Hello,

This email is intended as feedback on the structure of the TFAIA Critical Safety Incident reporting mechanism, not as a report of a specific incident.

I am an AI safety enthusiast, and I would like to suggest several small improvements to the reporting form available on your website. Here is my list of suggestions:

1. Add the date when the incident was discovered by the developer.

Section 22757.13 requires a developer to report a critical safety incident within 15 days of discovering it. However, the reporting mechanism asks for the date of the incident itself, rather than the date of discovery. In some cases, incidents may be discovered substantially later than when they actually occurred.

2. Add a section for basic information about the model and AI system involved in the incident, the developer, and affected parties.

For reports submitted by members of the public, these fields could allow an "Unknown" response where the information is not available.

3. Add information about reproducibility.

Is the incident reproducible? If so, has reproduction already been attempted? How many attempts were made, and in what proportion of runs did the AI system exhibit the same behavior?

4. Structure the existing question about why the event qualifies as a Critical Safety Incident.

In addition to a free-text explanation, the form could provide checkboxes for the categories in §22757.11(d)(1)–(4), together with an "Uncertain / unable to determine" option.

This could make the reporting process easier for members of the public who do not have legal expertise.

I hope these suggestions may be useful. I would be grateful if you could consider them in future updates to the reporting mechanism.

Kind regards,
Marko Prakhov-Donets

Copies of Sent Emails

Original PDF copies of the emails sent to the European AI Office and the California Cybersecurity Integration Center are included below.



Marko <markoprakhovdonets@gmail.com>

Suggested improvements to the GPAI serious-incident reporting template

1 message

Marko <markoprakhovdonets@gmail.com>
To: CNECT-AIOFFICE@ec.europa.eu

Sun, Sep 13, 2026 at 3:12 PM

Hello,

I am an AI safety enthusiast, and I would like to suggest several small improvements to the "*Report for Serious Incidents involving General-Purpose AI Models with Systemic Risk*" template. Here is my list of suggestions:

1. Add the date when the incident was discovered.

The current template asks for the start and end dates of the incident, but not when the provider or independent researchers first became aware of it. Since Article 55 requires serious incidents to be reported "without undue delay," this distinction could help differentiate late detection from late reporting.

For example, although the attack on RubyGems occurred on May 11, information about the attack was only publicly disclosed on September 11 by independent researchers.

Source: <https://www.rubyhack.ai/>

2. Add information about the AI system in addition to the model involved.

In the case of autonomous agents, behavior is determined not only by the underlying model, but also by the agent harness, available tools, permissions, and other parts of the system. I suggest keeping the existing *Model involved* field while adding a separate section describing the broader AI system.

3. Add information about reproducibility.

Is the incident reproducible? If so, has reproduction already been attempted? How many attempts were made, and in what proportion of runs did the AI system exhibit the same behavior?

4. Add a reporting basis / incident classification field.

The form could include checkboxes corresponding to the relevant AI Act reporting criteria, for example:

- Article 3(49)(a): death or serious harm to health;
- Article 3(49)(b): serious and irreversible disruption of critical infrastructure;
- Article 3(49)(c): infringement of obligations under Union law intended to protect fundamental rights;
- Article 3(49)(d): serious harm to property or the environment;
- Serious cybersecurity breach related to the GPAI model or its infrastructure;
- Uncertain / other.

I hope these suggestions may be useful. I would be grateful if you could consider them in future updates to the reporting template.

Kind regards,
Marko Prakhov-Donets



Marko <markoprakhovdonets@gmail.com>

Suggested improvements to the TFAIA Critical Safety Incident reporting mechanism

1 message

Marko <markoprakhovdonets@gmail.com>
To: calcsic@caloes.ca.gov

Sun, Sep 13, 2026 at 3:13 PM

Hello,

This email is intended as feedback on the structure of the TFAIA Critical Safety Incident reporting mechanism, not as a report of a specific incident.

I am an AI safety enthusiast, and I would like to suggest several small improvements to the reporting form available on your website. Here is my list of suggestions:

1. Add the date when the incident was discovered by the developer.

Section 22757.13 requires a developer to report a critical safety incident within 15 days of discovering it. However, the reporting mechanism asks for the date of the incident itself, rather than the date of discovery. In some cases, incidents may be discovered substantially later than when they actually occurred.

2. Add a section for basic information about the model and AI system involved in the incident, the developer, and affected parties.

For reports submitted by members of the public, these fields could allow an "Unknown" response where the information is not available.

3. Add information about reproducibility.

Is the incident reproducible? If so, has reproduction already been attempted? How many attempts were made, and in what proportion of runs did the AI system exhibit the same behavior?

4. Structure the existing question about why the event qualifies as a Critical Safety Incident.

In addition to a free-text explanation, the form could provide checkboxes for the categories in §22757.11(d)(1)–(4), together with an "Uncertain / unable to determine" option.

This could make the reporting process easier for members of the public who do not have legal expertise.

I hope these suggestions may be useful. I would be grateful if you could consider them in future updates to the reporting mechanism.

Kind regards,
Marko Prakhov-Donets