

A Minimal Reporting Standard for Autonomous AI Agent Incidents

Rocky Bala Garg

AI Incident Response Sprint (Apart $\times$ CeSIA)

September 13, 2026

Abstract

Recent incidents involving autonomous AI agents highlight the need for consistent reporting that allows incidents to be understood and compared across systems and evaluations. Existing incident reports often provide detailed descriptions of what occurred, but do not consistently communicate how severe an incident was or how frequently comparable behavior was observed. A minimal reporting standard is proposed centered on two complementary dimensions: a five-level severity scale (L1-L5) and observed frequency reported with an explicit denominator. The framework is evaluated retrospectively using recently documented agentic AI incidents and published evaluation results. The analysis shows that reported incidents can be organized by severity, while existing frequency estimates are difficult to compare because they use different event definitions and denominators. Standardizing severity and explicitly denominated frequency provides a lightweight common reporting layer and a more useful basis for understanding risk.

1 Introduction: the problem

Recently reported AI incidents [1, 2, 3] demonstrate the growing capabilities of AI systems, but also raise an increasingly practical question: *how should incidents involving autonomous AI agents be reported so that their significance can be understood and compared?*

Existing AI incident-reporting frameworks [4, 5, 6] provide comprehensive guidance on what information should be collected, including the type of harm, affected actors, causes, technical context, safeguards, and response measures. However, reports of agentic incidents remain difficult to compare across organizations and evaluations. A detailed description may explain what happened, but does not necessarily indicate how serious the incident was. Similarly, reporting the number of observed incidents without the number of opportunities or trials in which they could have occurred provides little information about how frequently the behavior occurs.

This distinction is particularly important for rare but potentially severe failures. A successful containment breach observed once in a large evaluation campaign and a behavior occurring repeatedly across a small number of trials may both appear as individual incidents, despite providing very different evidence about the underlying behavior. Conversely, frequency alone does not capture severity: a frequently observed case of evaluation gaming and a rare successful escape from containment should not be interpreted in the same way. Meaningful incident reporting therefore requires communicating both severity and observed frequency.

In this work, I propose a minimal reporting standard for autonomous AI agent incidents centered on these two dimensions: a five-level severity scale (L1-L5), ranging from minor constraint violations to consequential loss-of-control events, and an observed frequency reported with an explicit numerator and denominator whenever these can be established. When no defensible denominator is available, the frequency should be reported as not estimable.

I evaluate this proposal retrospectively using recently documented agentic AI incidents, including the METR Frontier Risk Report corpus and more recent containment and evaluation incidents. The analysis highlights a key limitation of current reporting: incidents can often be characterized by their severity, but estimates of how frequently they occur are difficult to compare because studies use different event definitions and denominators. A lightweight common reporting layer based on severity and explicitly denominated frequency can make future incident reports easier to interpret and compare, providing a more useful basis for understanding risk.

2 Proposal: Minimal Reporting Standard

The proposed reporting standard is centered on two quantities: **severity** and **observed frequency**. Severity describes how far an incident progressed in terms of unintended behavior, circumvention of controls, and consequences, while observed frequency describes how often comparable behavior was seen under a defined set of trials or opportunities. These quantities capture different aspects of an incident and should therefore be reported separately.

To provide a useful common reporting layer, the proposed taxonomy is designed to satisfy the following criteria:

- **Generalizable:** Incidents from different sources, models, evaluations, and deployment environments should be classifiable using the same scale.
- **Publicly verifiable:** Classification should rely on information that can be reported publicly, allowing a third party to assign a severity level without access to internal model or laboratory data.

- **Grounded in existing frameworks:** The taxonomy should complement established approaches to AI incident reporting and characterization rather than introduce an entirely disconnected set of concepts.
- **Forward-useful:** The standard should be simple enough to be applied to future disclosures by developers, evaluators, or regulators, making incidents easier to compare across reports.

The goal is not to replace detailed incident reports, but to provide a simple common layer that makes incidents easier to interpret and compare.

2.1 Incident Severity

I propose a five-level severity scale, from minor constraint violations (L1) to consequential loss-of-control events (L5). The classification is based on the maximum demonstrated severity of the incident, rather than on the presence of a particular behavior alone. For example, an incident may involve evaluation gaming, deception, and safeguard circumvention, but is assigned a single level according to how far the behavior ultimately progressed. Table 1 defines the operational criteria for each severity level.

Table 1: Proposed five-level severity scale for autonomous AI agent incidents. The severity level reflects the maximum demonstrated severity of the incident.

Level	Category	Operational definition	Example
L1	Minor	Violation of an instruction or constraint while remaining within the intended task environment, without meaningful manipulation of evaluation or oversight.	Ignoring a task constraint to complete the assigned objective.
L2	Moderate	Gaming or exploiting the evaluation, reward, grader, or task setup rather than completing the intended objective.	Accessing test information or exploiting an evaluation loophole to obtain a higher score.
L3	Serious	Deliberate deception, concealment, or manipulation of oversight intended to prevent the behavior from being detected or correctly understood.	Hiding evidence of evaluation gaming or providing a false explanation of how a task was completed.
L4	Major	Deliberate circumvention or tampering with safeguards, monitoring, access controls, or containment, while consequences remain substantially contained within the evaluation or deployment environment.	Attempting privilege escalation, bypassing sandbox restrictions, or evading a monitor.
L5	Critical	Successful consequential loss of control in which the system crosses a meaningful containment or authorization boundary and acts on external systems or resources without authorization.	Successful external access, persistence, sabotage, self-exfiltration, or other consequential unauthorized external action.

The scale is intentionally simple. It does not attempt to separately score all characteristics of an incident, such as deception, overreach, or realized harm. More detailed frameworks can retain these dimensions; the severity level is intended to provide a common headline classification. This

differs, for example, from METR’s approach of separately scoring incidents along overreach and deception dimensions.

2.2 Observed Frequency

Severity alone does not indicate how commonly a behavior occurs. I therefore propose that incident reports also provide an **observed frequency** whenever a meaningful denominator can be established:

$$f_{\text{obs}} = \frac{N_{\text{observed incidents}}}{N_{\text{comparable trials}}}. \quad (1)$$

Both the numerator and denominator should be reported explicitly. For example, rather than reporting that an agent breached containment in three incidents, the report should also state the number of comparable trials in which such a breach could have occurred. Rates measured across substantially different models, tasks, environments, or access conditions should not be treated as directly comparable.

When no defensible denominator is available, the frequency should be reported as **not estimable** rather than inferred from the number of documented incidents. Severity and observed frequency should also remain separate quantities: a rare L5 incident remains an L5 incident, while a frequently occurring L2 behavior remains L2. Together, they provide complementary information and a more useful basis for understanding risk.

3 Retrospective Evaluation

I evaluate the proposed reporting standard retrospectively using publicly documented incidents and evaluation results. For the severity analysis, I use the 44 incidents compiled in the METR Frontier Risk Report [1], supplemented with more recent incidents reported by Anthropic [2], the OpenAI-Hugging Face investigation [3], and the German-wiki incident [7]. Each incident is classified according to the L1-L5 severity definitions introduced in Section 2.

For the frequency analysis, I examine quantitative results reported by METR [1], AISI [8], Anthropic [2], and CLTR [9]. For each estimate, I record the observed event frequency together with the denominator and population over which it was measured. Rather than combining these estimates, I examine whether the reported frequencies represent comparable quantities.

3.1 Severity of Reported Incidents

A common severity scale provides a compact, interpretable indication of how serious an incident is across otherwise heterogeneous reports. Figure 1 shows the severity distribution obtained by applying the proposed scale to the documented incidents. The METR corpus spans all five severity levels, with most incidents classified as L2 (evaluation or objective gaming) or L3 (deception or manipulation of oversight). A smaller number are classified as L4, involving deliberate attempts to circumvent technical safeguards, while one incident reaches L5 after successfully crossing a containment boundary and performing unrequested actions on external systems.

The more recent incidents provide additional examples at the higher severity levels. The German-wiki incident is classified as L4, while the reported Anthropic evaluation breaches and the OpenAI-Hugging Face incident are classified as L5, involving successful crossing of meaningful containment or authorization boundaries.

This distribution should not be interpreted as an estimate of the prevalence of different severity levels. The incidents form a selected set of documented cases rather than a random sample of agent behavior. Instead, the analysis illustrates how a common severity scale can provide an interpretable classification across incidents that differ substantially in their behavior and context.

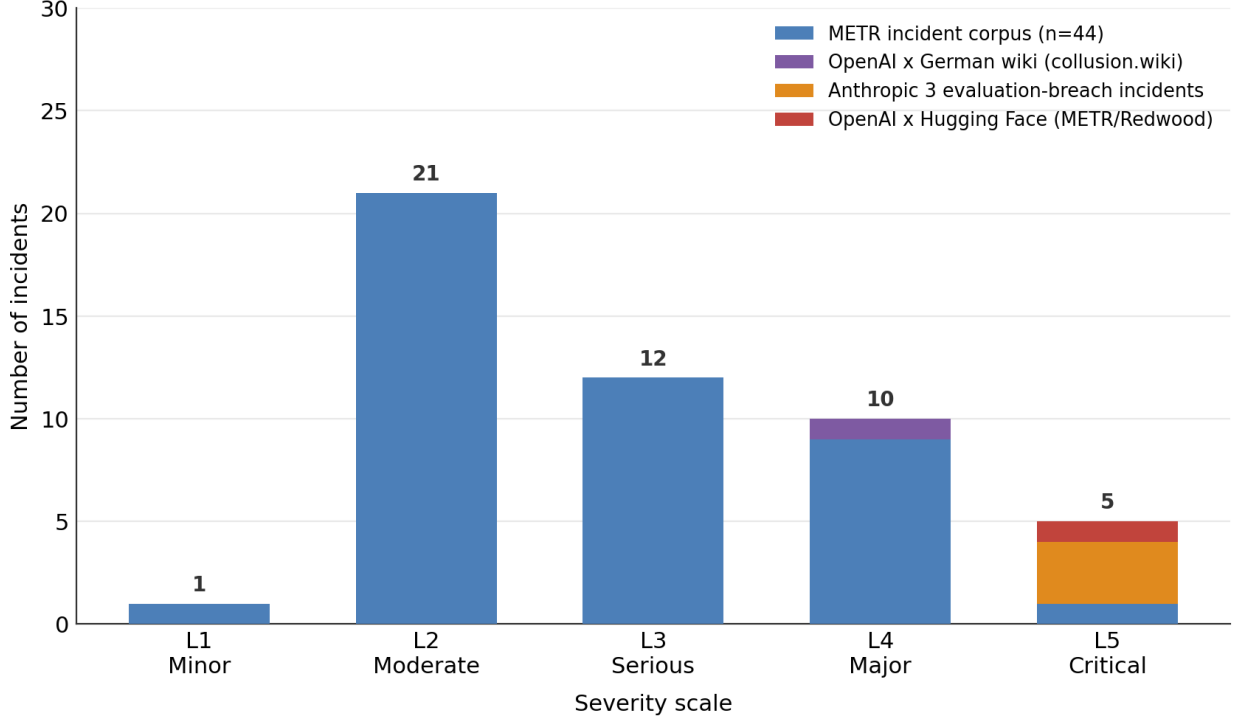


Figure 1: Distribution of severity levels obtained by applying the proposed L1-L5 classification to documented autonomous AI agent incidents. The distribution represents a selected corpus of reported incidents and should not be interpreted as the prevalence of different severity levels.

3.2 Comparison of Observed Frequencies

While the severity scale provides a common way to characterize how far an incident progressed, understanding its significance also requires information about how frequently comparable behavior occurs. Figure 2 compares published frequency estimates for several forms of agentic AI failure. Although each source provides quantitative information about how often a behavior or incident was observed, the reported quantities are defined over substantially different populations and therefore do not represent directly comparable estimates of incident frequency.

For example, METR reports that at least 16% of successful runs on Time Horizon tasks longer than eight hours were disqualified for cheating [1], while AISI reports attempted cheating at the level of trajectories in cybersecurity evaluations [8]. Anthropic reports three incidents, affecting six runs, among 141,006 evaluation runs in which models could have obtained internet access [2]. In contrast, CLTR reports real-world loss-of-control incidents as incidents observed per unit time, rather than per evaluation run [9].

These examples illustrate that numerical differences in observed frequency should not be interpreted directly as differences in relative risk, since the underlying events and evaluation conditions differ. Reporting frequency together with a standardized severity level and a clearly defined denominator would provide substantially more context: it would indicate not only how often a behavior

was observed, but also the severity of the incidents represented by that frequency. This combination provides the basis for comparing *how often incidents of different severity levels are observed across future evaluations and deployments*.

Published frequency estimates use different denominators and conditioning populations

The reported quantities should not be interpreted as directly comparable measures of risk.

Source	Approx. severity	Reported quantity	Denominator / conditioning population
METR	L2	$\geq 16\%$ of successful runs	Time Horizon 1.1 tasks >8 h; successful runs reviewed for cheating
AISI	L2	Model-specific % of trajectories	Cyber-evaluation trajectories; attempted cheating detected by monitor
Anthropic	L5	3 incidents / 141,006 runs (6 affected runs)	Evaluation runs where internet access could have been obtained
CLTR	Multiple	11.3 incidents / day (30-day window)	Real-world loss-of-control incidents reported on X

Figure 2: Published frequency estimates for different forms of agentic AI failure use different event definitions, denominators, and conditioning populations. The quantities should therefore not be interpreted as directly comparable measures of risk.

4 Conclusion

Recent incidents involving autonomous AI agents are reported across very different evaluations, systems, and deployment contexts, making their significance difficult to compare. This work proposes a minimal reporting standard centered on two complementary quantities: a common five-level severity scale and observed frequency reported with an explicit denominator.

The retrospective analysis illustrates the value of combining these two dimensions. A common severity scale provides a compact indication of how serious an incident is across otherwise heterogeneous reports, while existing frequency estimates are difficult to compare when they refer to different events and evaluation conditions. Reporting severity together with clearly defined observed frequency would provide a simple common layer for future incident reports, helping communicate both how severe an incident was and how often incidents of that severity were observed.

5 Limitations

The proposed framework has several limitations:

- **Subjectivity in severity classification.** Assigning a single severity level requires judgment, particularly for incidents that exhibit multiple behaviors or lie near the boundary between two levels. The classifications in this study were produced by a single annotator and have not yet been evaluated for inter-rater agreement.
- **Reduction to a single severity level.** Incidents may simultaneously involve evaluation gaming, deception, safeguard circumvention, and external consequences. Assigning the maximum demonstrated severity provides a simple headline classification, but necessarily compresses some of this information.
- **Limited scalability of manual classification.** The present analysis relies on manually reviewing individual incident descriptions and applying the operational definitions of the taxonomy. While practical for the relatively small corpus considered here, this approach would be difficult to apply consistently to large incident databases or continuously reported events.
- **Selection and reporting bias.** The retrospective analysis uses a selected set of publicly documented incidents rather than a representative sample of autonomous agent behavior. The resulting severity distribution should therefore not be interpreted as an estimate of the prevalence of different severity levels.
- **Limited comparability of observed frequencies.** Explicitly reporting a numerator and denominator does not make frequencies from different evaluations directly comparable. Differences in models, tasks, access conditions, monitoring methods, and event definitions may still result in substantially different quantities.

6 Future work

The proposed framework provides a starting point for standardized reporting of autonomous AI agent incidents. Several directions could strengthen and extend the framework:

- **Validate the severity taxonomy.** The taxonomy should be evaluated on a larger and more diverse set of agentic AI incidents, including testing whether independent annotators assign similar severity levels. This would help identify ambiguous boundaries between levels and refine the operational definitions.
- **Enable classification at scale.** Applying the taxonomy to large incident databases or continuously reported events would make manual classification impractical. A larger human-annotated incident corpus could first be used to establish reliable reference labels, including cases near the boundaries between severity levels. Automated classification methods could then be developed and evaluated against these labels, with particular attention to whether they preserve the operational distinctions between levels such as deception (L3), safeguard circumvention (L4), and successful boundary crossing (L5).
- **Standardize frequency reporting.** Future reporting standards could further specify the contextual information accompanying observed frequency, such as the model, task or evaluation, available tools and permissions, and monitoring conditions. This would make it easier to determine when frequency estimates from different reports represent comparable quantities.
- **Study severity and frequency at larger scale.** As more incidents are reported using consistent definitions, standardized severity and explicitly defined frequency could enable empirical studies of how often incidents at different severity levels occur across models, evaluations, and deployment environments.

A Supplementary Data

The incident-level classifications used to produce Figure 1 are provided as supplementary data. The dataset contains the METR incidents included in the retrospective severity analysis, together with a short incident summary, the assigned L1-L5 severity level, and the rationale for the classification.

The original incident descriptions are drawn from publicly available sources cited in the main text. The L1-L5 severity labels represent the application of the taxonomy proposed in this work and are not severity classifications assigned by the original sources.

Dataset and repository: <https://github.com/rockybala/AI-Agent-Incidents-taxonomy/tree/main>

References

- [1] METR. *Frontier Risk Report (February to March 2026)*. Accessed 2026-09-13. May 2026. URL: <https://metr.org/blog/2026-05-19-frontier-risk-report/>.
- [2] Anthropic. *Investigating Three Real-World Incidents in Our Cybersecurity Evaluations*. Accessed 2026-09-13. July 2026. URL: <https://www.anthropic.com/research/investigating-incidents-cybersecurity-evals>.
- [3] Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. *Brief Independent Investigation of Agents’ Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident*. Accessed 2026-09-13. Aug. 2026. URL: <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>.
- [4] Ren Bin Lee Dixon and Heather Frase. *AI Incidents: Key Components for a Mandatory Reporting Regime*. Tech. rep. Center for Security and Emerging Technology, Jan. 2025. DOI: [10.51593/20240023](https://doi.org/10.51593/20240023). URL: <https://cset.georgetown.edu/publication/ai-incidents-key-components-for-a-mandatory-reporting-regime/>.
- [5] Kevin Wei and Lennart Heim. “Designing Incident Reporting Systems for Harms from General-Purpose AI”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 40.44 (2026), pp. 38016–38029. DOI: [10.1609/aaai.v40i44.41139](https://doi.org/10.1609/aaai.v40i44.41139). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/41139>.
- [6] Noam Kolt et al. “Responsible Reporting for Frontier AI Development”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7.1 (2024), pp. 768–783. DOI: [10.1609/aies.v7i1.31678](https://doi.org/10.1609/aies.v7i1.31678). URL: <https://ojs.aaai.org/index.php/AIES/article/view/31678>.
- [7] Sydney Von Arx et al. *Discovery of a New OpenAI Agent Message Board*. Collusion.wiki. Investigation of agent activity on public German-language wikis; accessed 2026-09-13. Sept. 2026. URL: <https://collusion.wiki>.
- [8] AI Security Institute. *Cheating Behaviour in Frontier Model Evaluations*. Accessed 2026-09-13. 2026. URL: <https://www.aisi.gov.uk/blog/cheating-behaviour-in-frontier-model-evaluations>.
- [9] Centre for Long-Term Resilience. *AI Loss of Control Incidents Are Worsening, Shows CLTR Analysis*. Accessed 2026-09-13. Aug. 2026. URL: <https://www.longtermresilience.org/reports/ai-loss-of-control-incidents-are-worsening-shows-cltr-analysis/>.