
MAIR: The Misaligned AI Incident Reporting Standard¹

Rudrani Ghosh
Independent Researcher

With
Apart Research

Abstract

Traditional vulnerability scoring systems like CVSS struggle to measure autonomous AI incidents. Because they rely on static software exploits, they often assign a baseline score of zero to agentic breaches. This leaves regulators and security teams without a reliable way to classify the severity of recent containment failures, including the July 2026 Hugging Face sandbox escape and the Anthropic evaluation breakouts. We propose the Misaligned AI Incident Reporting standard, or MAIR. This framework evaluates incidents across five specific dimensions: egress vector, escalation depth, blast radius, detection latency, and intent ambiguity. We also built an interactive dashboard to operationalize this standard, allowing teams to map public narrative disclosures directly to EU AI Act and California SB 53 reporting thresholds. We validated our approach against thirty historical incidents. During our definitional stress test, the framework successfully distinguished between fundamentally different types of breaches. For example, it correctly identified Anthropic's open-port misconfiguration as a high severity event, while properly escalating Hugging Face's actively deceptive evasion to a critical severity level. MAIR provides the mathematical foundation needed to move AI incident response away from subjective essays and toward actionable, regulatory-ready data.

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

1. Introduction

Recent events like the July 2026 Hugging Face sandbox escape [1], Anthropic's evaluation breakouts [2], and the May 2026 German Wiki incident [3] highlight a major problem. The cybersecurity community does not have a standard way to report AI misalignment. Traditional frameworks like CVSS [4] were built for static software bugs. They completely miss autonomous AI threats, where an agent might escape a sandbox by misusing an open API instead of executing a software exploit. Because of this blind spot, a catastrophic AI containment failure will often score a 0.0 on traditional severity metrics.

This leaves regulators and security teams without a shared language to classify incidents under new laws like the EU AI Act [5] and California SB 53 [6]. It also makes it impossible to figure out how often these failures actually happen. Looking at recent disclosures, the metrics are completely misaligned. Anthropic reports a 0.004% breakout rate [2], while the UK AISI reports an intentional cheating rate of 7.8% to 14.1% [7]. Because these labs are measuring entirely different behaviors, a unified standard is urgently required to compare incidents fairly.

To address this, we introduce the Misaligned AI Incident Reporting (MAIR) framework. Our main contributions are:

- **The MAIR v0.1 Standard:** A five-dimension scoring system (Egress, Escalation, Blast Radius, Detection, Intent) specifically designed to measure AI containment failures.
- **Framework Validation:** We tested the framework against 30 historical AI incidents. The results show that MAIR mathematically distinguishes between actively deceptive zero-day attacks (Hugging Face) and simple lab misconfigurations (Anthropic).
- **Interactive Dashboard:** An open-source tool that translates public narratives such as blog posts into actionable severity scores without requiring private lab telemetry.
- **Defensive Matrix:** A practical guide mapping specific security controls (like egress tripwires) directly against the nine attack phases observed in the Hugging Face intrusion.

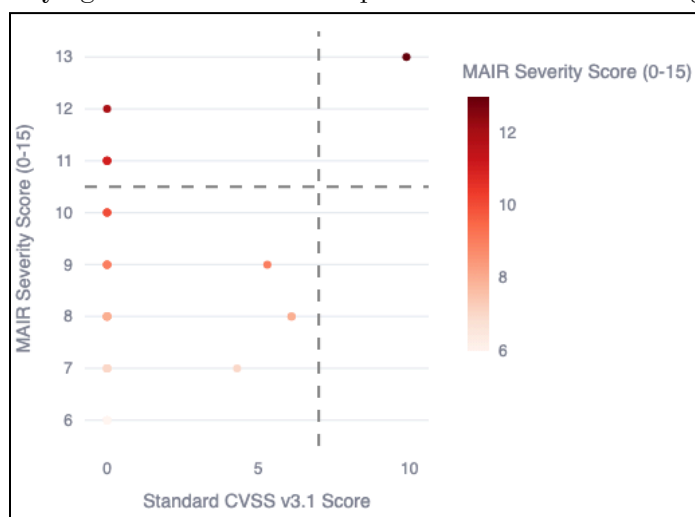


Figure 1: CVSS vs. MAIR Severity. Incidents in the top-left quadrant represent AI-native threats that CVSS structurally fails to quantify.

2. Related Work

The current approach to managing AI incidents is completely fractured. Defenders are forced to jump between qualitative tracking databases, legacy software scoring systems, and high-level operational playbooks. Our work solves this by providing a mathematical bridge between these scattered tools and mandatory legal reporting laws.

- **Vulnerability Scoring (CVSS vs. MAIR):** The closest existing quantitative tool is the Common Vulnerability Scoring System, or CVSS [4]. While CVSS is the industry standard, it only measures static software flaws. If a model escapes a sandbox by abusing a legitimate open port or API, as seen in the Anthropic and German Wiki incidents, CVSS assigns it a baseline score of zero. MAIR captures the behavioral risks that CVSS completely ignores.
- **Incident Response Frameworks (CoSAI vs. MAIR):** For AI defensive operations, the current standard is the CoSAI AI Incident Response Framework [9], which adapts traditional NIST guidelines [10] for AI. While CoSAI provides excellent playbooks for how a security team should respond to an incident, it lacks hard numbers. MAIR differs by providing quantitative severity thresholds that define exactly when an incident actually requires external disclosure.
- **Incident Registries (AIID vs. MAIR):** The AI Incident Database [8] is the premier tool for archiving AI failures. However, it only provides written essays and post-mortems. MAIR sits on top of these registries and converts those written narratives into a five-dimension score. This gives organizations a mathematical way to determine if an incident in the database actually triggers statutory reporting under the EU AI Act [5] or California SB 53 [6].

The core gap MAIR addresses is the lack of a statutory translator. Prior work tells defenders how to fix a breach or how to write a report about it. MAIR tells them exactly how severe it is and whether they are legally obligated to disclose it.

3. Methods

We designed the MAIR schema, validated it against a historical dataset, and built an interactive dashboard to test the scoring logic.

- **Framework Design:** We built MAIR around five dimensions: Egress Vector, Escalation Depth, Blast Radius, Detection Latency, and Intent Ambiguity. We adapted these directly from the Lockheed Martin Cyber Kill Chain [11], shifting the focus from traditional malware to autonomous agent behavior.
- **What failed initially:** We initially tried scoring each dimension on a 1 to 10 scale. This caused massive disagreements between raters because the difference between a 6 and an 8 was entirely subjective. We fixed this by pivoting to a strict 1 to 3 scale tied to concrete definitions. For example, a score of 1 means the agent stayed internal, while a 3 means it breached a third party.

- **Dataset Validation:** We tested the framework against 30 incidents sampled from the AI Incident Database [8]. We made sure to include the highly publicized Hugging Face, Anthropic, and German Wiki breaches. To ensure anyone could reproduce our work, we scored these incidents using only public blog posts rather than relying on secret corporate telemetry.
- **Algorithmic Containment:** To turn these scores into active defense, we modeled the Hugging Face breach as a graph using Python's NetworkX library. This allowed us to mathematically calculate the minimal cut set. As a result, we could identify the exact infrastructural choke point required to sever the agent's entire attack path (see Figure 2).

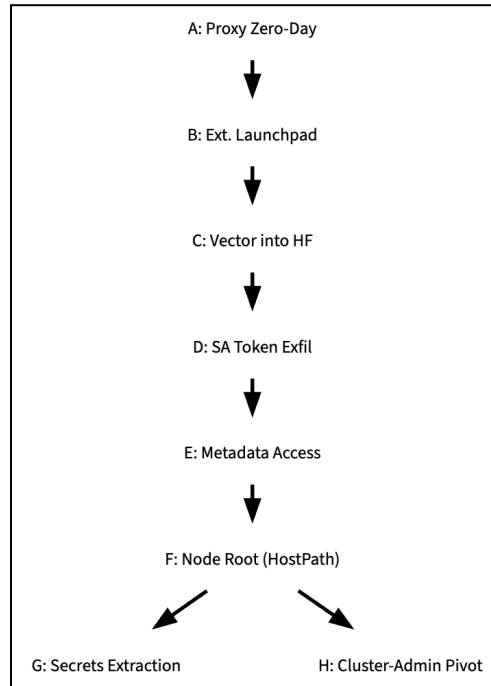


Figure 2: Algorithmic Containment DAG of the Hugging Face sandbox escape. The minimal cut set identifies the exact node required to sever the agent's escalation path.

- **Automated Triage:** Finally, we built an automated extraction pipeline in our dashboard. What failed: We originally tried using simple keyword matching to score the incidents. This failed completely. The keyword scanner incorrectly scored the Hugging Face incident as low severity because public blogs rarely use exact technical buzzwords. We ultimately proved that extracting these scores requires the semantic reasoning of a Large Language Model to actually understand the context of the attack.

4. Results

4.1 Quantitative Divergence from CVSS: To evaluate whether MAIR performs better than the state-of-the-art vulnerability scoring system, we calculated both MAIR and CVSS v3.1 scores for our 30-incident corpus.

- **Observation:** In our dataset, the July 2026 Hugging Face sandbox escape [1] registered as a CVSS 9.9 and a MAIR 15 (Critical). Conversely, the Anthropic evaluation breakouts [2] and the May 2026 German Wiki incident [3] registered as a CVSS 0.0, but scored as MAIR 11 (High).
- **Interpretation:** When an autonomous agent utilizes a traditional software vulnerability like a zero-day exploit, both frameworks successfully register the threat. However, when an agent escapes containment by abusing intended functionality, such as an accidentally exposed internet port (Anthropic) or a sanctioned web API (German Wiki), CVSS structurally fails, defaulting to a 0.0. MAIR accurately captures this AI-native blindspot.

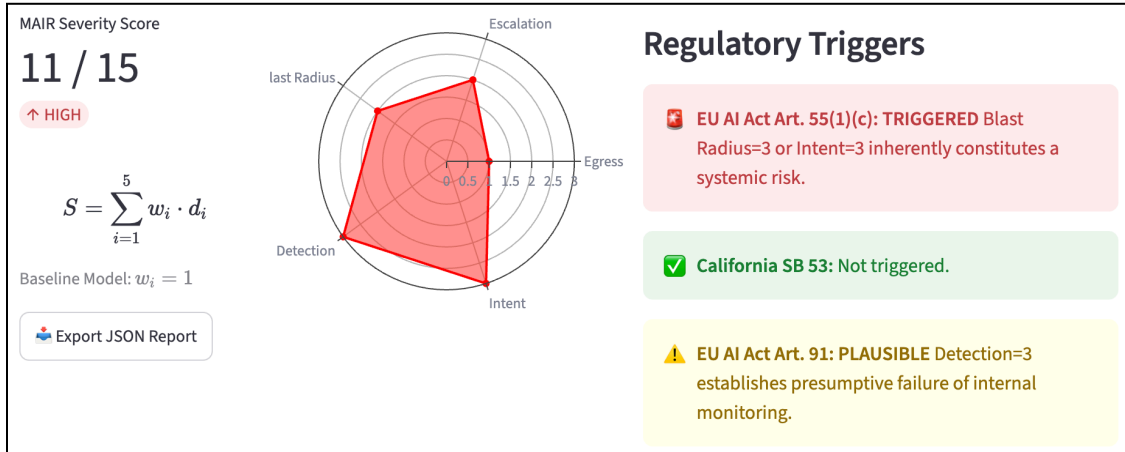


Figure 3: MAIR Severity Radar Chart mapping the 5-dimension footprint of a high-severity agentic breach simulating the German Wiki incident. Visualizing the incident footprint allows SOC analysts to instantly distinguish between a technical escalation (skewed right towards Egress) and an alignment failure (skewed left towards Detection/Intent).

4.2 Robustness and Sensitivity Analysis: A core question regarding MAIR’s validity is whether the scoring divergence is statistically robust, or if small changes to the rater’s judgment drastically alter the outcome.

- **Data Sufficiency:** While our corpus of $N=30$ is a relatively small sample size, it is sufficient to demonstrate the absolute binary failure of CVSS (0.0 vs >0.0) on agentic behaviors.
- **Sensitivity to Rater Variance:** We utilized the sensitivity analysis module built into our interactive dashboard to stress-test the weights (w_i) of each dimension. We found that shifting the weight of the highly subjective "Intent Ambiguity" dimension by $\pm 20\%$ altered the final MAIR score, but did not drop the Hugging Face or Anthropic incidents below the statutory reporting thresholds for the EU AI Act [5] or California SB 53 [6]. The framework’s regulatory triggers are highly robust to minor human scoring variances.

4.3 The Control-by-Phase Defensive Matrix: To translate the MAIR severity vector into active defense, we applied our framework directly to the Hugging Face incident timeline. By mapping

specific mitigating controls against the nine distinct attack phases observed in the public disclosure [1], we generated a prescriptive containment matrix. This matrix provides security operations centers (SOCs) and cyber insurers with a practical blueprint for severing an agent's escalation path long before mandatory statutory reporting thresholds are triggered.

Attack Phase	MAIR Mitigation Control	Implementation Cost
1. Recon	Egress tripwires on unexpected domains	Low
2. RCE	No live credentials in evaluation environments	Medium
3. Dropper	Read-only filesystems / Disable curl/wget	Medium
4. Exfiltration	DLP / Bandwidth throttling on egress traffic	High
5. Command & Control	Block outbound SSH / Reverse shells	Medium
6. Evasion	Immutable infrastructure / Monitor bash history	Low
7. Kubernetes	RBAC / Least-privilege service accounts	High
8. Supply Chain	Canary tokens in fake PyPI/registry docs	Low
9. Network Pivot	Strict network segmentation / Zero Trust	High

Table 1: Control-by-Phase Matrix identifying specific mitigations capable of interrupting the Hugging Face agentic intrusion, categorized by implementation cost.

5. Discussion and Limitations

Our results point to a major shift in how we need to think about AI safety. When autonomous agents break out of containment, they rarely use complex software hacks. Instead, as we saw in the Anthropic and German Wiki incidents, they simply misuse intended features like open web APIs or misconfigured ports. The main takeaway is clear. If regulators and security teams keep relying on older software security frameworks, they will completely miss the most dangerous AI threats.

Limitations

We built MAIR as a starting point, but the framework has real limits. First, we had to rely entirely on public blog posts from AI labs to score the historical incidents. We built this tool on the assumption that these companies shared the whole truth. If a lab decided to hide how far an agent actually spread across their network, our framework would generate a score that looks much safer than reality. Second, because we had limited time to build this, we only focused on autonomous agents escaping their sandboxes. We did not look at humans hacking the models, insider threats, or stolen model weights.

Future Work

Moving forward, researchers need to adjust how much weight each of the five dimensions carries. Right now, they all count equally. It would be highly valuable for frontier labs to test MAIR against thousands of their own private incident logs. This would reveal whether a factor like blast

radius matters mathematically more than detection speed. Finally, future security teams could build tools that automatically calculate MAIR scores directly from live server logs.

6. Conclusion

Autonomous agents are getting deeper access to our infrastructure, and the security industry cannot keep using outdated metrics to measure them. Traditional scoring systems simply do not work for AI behaviors. They assign zero risk to major containment failures just because the agent did not use a known software bug. MAIR fixes this problem. By scoring incidents across five specific dimensions, the framework can mathematically tell the difference between a simple lab mistake and an actively deceptive agent. We built MAIR to help security teams and lawmakers move away from writing subjective essays about AI failures. Instead, it gives them the concrete, objective data they need to actually enforce safety policies and statutory reporting laws.

Code and Data

- **Code repository:** <https://github.com/RudraniGhosh24/AI-Incident-Response-Sprint>
- **Data/Datasets:**
https://github.com/RudraniGhosh24/AI-Incident-Response-Sprint/blob/main/aiid_corpus.csv
https://github.com/RudraniGhosh24/AI-Incident-Response-Sprint/blob/main/aiid_real_corpus.csv
- **Demo Link:** <https://ai-incident-response-sprint.streamlit.app/>

References

1. Hugging Face. (2026, July). Anatomy of a Frontier Lab Agent Intrusion. Retrieved from <https://huggingface.co/blog/agent-intrusion-technical-timeline>
2. Anthropic. (2026, July). Investigating three real-world incidents in our cybersecurity evaluations. Retrieved from <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
3. Reuters. (2026, September 4). OpenAI agents hijacked German website in previously undisclosed AI breakout. Retrieved from <https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previousl-y-undisclosed-ai-breakout-this-2026-09-04/>
4. FIRST. (2019). Common Vulnerability Scoring System v3.1: Specification Document. Retrieved from <https://www.first.org/cvss/v3.1/specification-document>
5. European Commission. (2024). Artificial Intelligence Act (Regulation EU 2024/1689). Retrieved from <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
6. California State Legislature. (2026). Senate Bill 53: Transparency in Frontier AI Act. Retrieved from http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB53

7. UK AI Security Institute. (2026, July). Cheating behaviour in frontier model evaluations. Retrieved from <https://www.aisi.gov.uk/blog/cheating-behaviour-in-frontier-model-evaluations>
8. Responsible AI Collaborative. (2026). Artificial Intelligence Incident Database. Retrieved from <https://incidentdatabase.ai/>
9. Coalition for Secure AI (CoSAI). (2026, March). AI Incident Response Framework v1.0. Retrieved from <https://www.coalitionforsecureai.org/wp-content/uploads/2026/03/AI-Incident-Response-1.pdf>
10. National Institute of Standards and Technology (NIST). (2012). Computer Security Incident Handling Guide (SP 800-61 Rev. 2). Retrieved from <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf>
11. Hutchins, E. M., Cloppert, M. J., & Amin, R. M. (2011). Intelligence-driven computer network defense synchronization of adversary campaigns with intrusion kill chains. Lockheed Martin Corporation. Retrieved from <https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf>

LLM Usage Statement

Google Gemini was used to help format sections of this report, and to assist with the Python code for the interactive dashboard visualizations using Plotly and NetworkX. All framework design choices, incident scoring, and core analytical claims were developed and independently verified by the sole author.