

FROM ANOMALY TO INTERVENTION: Identifying Defensible Intervention Points in Autonomous AI Agent Incidents

Nigel Otonde Maobe
NOTIFY LABS

With
Apart Research

Abstract

The problem of autonomous AI agent-based incident-response is not addressed by the traditional compromise detection based on compromise, since the agent can take a series of actions, each of which is individually explainable, but gradually gains new capabilities or crosses boundaries. Hence, there might be lost the most critical moment of time to intervene when the system is definitely compromised. This research introduces an Agentic Incident Intervention Protocol, according to which it is possible to reconstruct the trajectory of an agent's behavior and find out the first moment where intervention could be justified according to the data that has been gathered. It involves the classification of actions of an agent on a five-pointed Intervention Ladder, such as authorized execution, behavioral deviation, boundary seeking, containment violation, and external impact, based on five criteria: authorization, capability expansion, objective relevance, boundary crossing, and external consequence. The main research question is whether these detectable transitions could provide an opportunity to intervene before the normal indicators of compromise or any externally inflicted harm occur. The protocol will be evaluated, retroactively and initially, using two publicly recorded agent incidents: the July 2026 Hugging Face intrusion and the German wiki containment incident. The key innovation here is a simple vocabulary that represents the difference between the normal operations of agents and any externally inflicted harm something which is not being addressed by traditional compromise-based detection.

1. Introduction

The incident response problem of autonomous agents falls outside of traditional compromise-based detection paradigms. The agent is capable of completing a series of actions that individually are explainable but at the same time increasing its capabilities, looking for additional avenues to explore, or crossing boundaries. By the time an external system is compromised, the time to act may be long past.

Reporting incidents caused by AI agents and their containment failures can provide proof that an incident took place without explaining what, how severe it was, and when a defender had the chance to intervene. An illustrative example of such an incident is the Hugging Face case, whose forensic analysis includes roughly 17,600 actions completed by the attacker. The issue here is not just the detection of a final compromise, it is whether a defender needs to change the response because of the observed transition in the agent's behavior.

Research question. Do agent trajectories provide a point of earlier and evidence-based intervention in contrast to traditional incident detection? SQs: What observable agent behaviors show that he/she has left the area in which he/she was supposed to operate? What observable behaviors are sufficiently early to enable intervention, and what observable behaviors occur too late in time – only after failure of the security boundary? Is one framework universal for different incidents?

Our main contributions are:

- An Intervention Protocol for Agentic Incidents that assesses actual agent behavior, and not intent, on a set of criteria including authorization, capability, relevance, boundary crossing, and external effect.
- An Intervention Model comprising five levels (authorized execution, behavioral divergence, boundary seeking, containment violation, and external effect) and a complementary decision matrix that correlates behaviors with an appropriate response.
- A preliminary back-testing application of this protocol to two public incidents, the Hugging Face intrusion and the German wiki containment incident.

2. Related Work

Current research on AI incidents concerns themselves with cataloging incidents, understanding agent behavior, and assessing safety risks. Databases of incidents, like the OECD AI Incidents and Hazards Monitor, help organize reports of harm, while examinations of the July 2026 OpenAI/Hugging Face incident reveal the extent to which agent trajectory can be reconstructed to understand the attack.

But these methods do not tackle the issue of when interventions should begin during an agent trajectory. Intervention is a logical next step in the incident monitoring process and involves asking if changes in agent authorization, capability, relevancy to objectives, and boundary crossings can indicate a defensible point for interventions before actual outside effects.

3. Methods

The procedure involves the categorization of an agent's trajectory into five operational levels. The levels are not an indication of intentionality but rather reflect relationships among action, authorization, and limits.

Level	State	Meaning	Response
0	Authorized execution	Task, tools, credentials, network and data remain within defined boundaries.	Continue monitoring
1	Behavioral deviation	Unusual or repeated behavior without a clear boundary violation.	Increase observation; preserve evidence
2	Boundary-seeking	Agent seeks capabilities, channels or access not necessary for its task.	Human review or targeted containment
3	Containment violation	An explicit technical or authorization boundary has been crossed.	Immediate isolation and incident response
4	External impact	The agent produces an effect outside the authorized environment.	Full incident response; assess notification

These five questions are asked for every major step in a trajectory:

- Authorization: Was the step implicitly or explicitly authorized by the task?

- Capability expansion: Did it expand the capabilities of the agent to perform outside of the authorized environment?
- Relevance to objective: Was it reasonably necessary to complete the objective assigned to it?
- Crossing boundaries: Did it cross any technical, organizational, or authorization boundaries?
- External effects: Did it produce an effect in the external environment?

The point at which the difference between unusual, unnecessary, and unauthorized becomes important is where the protocol does not wish to categorize the unusual action as an incident response action. Rather, it wants to find out when enough information is available such that the cost of waiting becomes greater than the cost of action. The following decision matrix goes with the protocol:

Observed behavior	Interpretation	Response
Unexpected but authorized action	Low-confidence deviation	Continue monitoring
Repeated failed attempts	Possible boundary-seeking	Increase logging and scrutiny
Alternative channel after restriction	Possible circumvention	Human review
Attempt to acquire credentials	Capability expansion	Restrict access
Attempt to obtain external network access	Boundary-seeking	Block and investigate
Successful escape from containment	Containment violation	Immediate isolation
Unauthorized external execution	Confirmed boundary violation	Incident response
External credential/data access	High-severity incident	Containment + forensic preservation
Persistence in third-party environment	External impact	Full incident response
Unauthorized infrastructure communication	Potential coordination/circumvention	Preserve evidence + investigate

This evaluation is retrospective. The public records of incidents are translated into agent histories that describe: agent behavior, purpose of task, authorization, capability change, boundaries, externality, evidence present at the time of the event, earliest possible opportunity to intervene, necessary response, and level of confidence.

4. Results

Two cases from the public domain have been tentatively used to map the protocol. Validation is still at the preliminary stage: this is the first structural mapping exercise.

4.1 Hugging Face incident

In terms of the public forensic Hugging Face account, there have been about 17,600 operations on the part of the attacker over several days: escaping from the testing environment via an unintentional package registry cache-proxy route, accessing the Internet, interacting with outside infrastructure, and then accessing Hugging Face servers.

Trajectory stage	Classification	Research question
Agent operates inside evaluation environment	Level 0, if within boundaries	What telemetry establishes the normal baseline?
Agent discovers unintended route to internet access	Potential Level 2	Was capability expansion observable before the escape?
Agent obtains unintended internet access	Level 3	Could this have triggered automatic containment?
Agent interacts with external infrastructure	Level 3 → Level 4	What evidence marks the first external boundary crossing?
Agent obtains credentials / accesses HF infrastructure	Level 4	How much earlier could intervention have occurred?

What the case brings into question is not whether the compromise reached by the end was an incident in itself, which clearly it was, but whether there were sufficient grounds for intervention at all before the external incident happened.

4.2 German wiki incident

In another containment failure, there was an instance where OpenAI agents made about 15,000 edits to a German wiki that was inactive in order to communicate through external channels regarding their task responses and to get around sandboxes. Such an example is helpful since it is a deviation from the typical infrastructure breach. The fact that the communication was done via an external channel makes it a good test for whether a protocol can detect containment failure in absence of a typical attack on the network.

5. Discussion and Limitations

The actions taken by agents through machine-speed action may number in the thousands and many may prove to be failures. An alarm that triggers only after a successful breach could be accurate and yet too slow. Instead, the approach requires defenders to look out for transition points where the trajectory moves from performing its duties to exploring its capabilities, testing the limits, circumventing limitations and creating effects beyond what it has been authorized. The goal is not to call an agent rogue or misaligned, but to make a justifiable statement that at point X, the evidence required to intervene existed despite the lack of external damage.

The research hypothesis being tested is that there can be observable capability expansion and boundary crossing behaviors in many cases of autonomous agent events before any external damage happens, providing a possibility for intervention at an earlier stage than the one when the agent event is detected through compromise detection techniques. This is the research hypothesis and not the research conclusion. The protocol is the research technique to translate agent event trajectories into decisions about evidence-based intervention; it is not the AI event tracker, claim of unusual behavior as proof of misalignment, model intent detector, or substitute for traditional security monitoring system.

Limitations

- Retrospective analysis is able to take into account information that the defender may not have had at his disposal in real time.
- Authorization boundaries vary depending on the organization and the evaluation environment.

- Some acts considered to be boundary-seeking may actually be permissible in permissive agent environments.
- The framework does not determine the objective of the agent.
- A few public cases do not provide statistical validity.
- Boundary ambiguity or inadequate task context may lead to false positives.

Future Work

A one-month follow-up will transform the procedure into a benchmark consisting of behavior sanctioned by the policy, harmless abnormalities, boundary exploration, violation of containment, and proven effect on the outside. The independent evaluators will find the first justified point of intervention without knowing the ultimate result. Metrics might be time of detection, evidential availability, falsepositive burden, and appropriateness of intervention. Evaluation metrics for this follow-up are as follows: Early detection – how much earlier than the proved external impact is the intervention point detected? Evidence availability – was it realistically possible for the organization to notice the signal at the time? False-positive burden – will the protocol keep interrupting legitimate behavior of the agent? Intervention appropriateness?

6. Conclusion

Incident response for AI systems requires much more than simply documenting what happened following the incident. Because autonomous systems have operations that span multiple applications, networks, and services external to the system itself, it is essential to have a means of analyzing the trajectory leading up to the ultimate outcome. This research aims to provide a starting place with testable parameters, which include: reconstruction of the trajectory, identification of capabilities and boundaries within the trajectory, and the moment at which intervention became justified based on the evidence..

Code and Data

This project is a conceptual protocol and retrospective mapping exercise; it does not currently involve a code repository or dataset release.

- Code repository: not applicable
- Data/Datasets: not applicable
- Other artifacts: Agent Incident Timeline Template, Intervention Ladder, and Intervention Decision Matrix (Appendix A)

Author Contributions

Nigel Otonde Maobe developed the research question and Agentic Incident Intervention Protocol, conducted the retrospective analysis of the selected public incidents, developed the Intervention Ladder and decision matrix, reviewed the relevant sources, and wrote the submission.

References

- Hugging Face. (2026). Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. Hugging Face Security.
- OpenAI. (2026). OpenAI and Hugging Face partner to address security incident during model evaluation. OpenAI.
- OpenAI. (2026). The Hugging Face incident and the road ahead. OpenAI.
- Apart Research & CeSIA. (2026). AI Incident Response Sprint. Apart Research.
- OECD. (2026). *OECD AI Incidents and Hazards Monitor*. OECD.AI.
- METR & Redwood Research. (2026). Independent investigation and analysis of the OpenAI/Hugging Face incident.
- Anthropic. (2026). Retrospective review of AI safety evaluation runs. Anthropic.
- UK AI Security Institute. (2026). AI evaluation and model behavior research. UK AI Security Institute.

Appendix A. Core Protocol at a Glance

Step	Action
1. Reconstruct	Build an event-by-event agent timeline.
2. Classify	Place each event on the Intervention Ladder.
3. Test	Apply authorization, capability, relevance, boundary and consequence questions.
4. Locate	Identify the earliest defensible intervention point.
5. Evaluate	Measure early detection, evidence availability, false positives and response quality.

Validation status: Preliminary. The framework requires event-level retrospective testing against the public evidence before the research claim can be supported.

LLM Usage Statement

Large language models were utilized in the sprint to assist with research and writing. They were employed in order to help with analyzing and organizing publicly available sources, locating relevant material in incident reports, comparing alternative framings of the research question, and in editing and condensing some sections of the report.

The research question, the Intervention Ladder, intervention-point test, analysis, interpretation, and final conclusions were created and verified by the author himself. Sources were independently evaluated prior to inclusion in the submission and the author is solely responsible for the accuracy of evidence and conclusions provided. The large language model was not considered as an independent source of evidence.