

Attribution Is Not Done and the Clock Is Running: A 90-Minute Tabletop Exercise Toolkit for Agent Loss-of-Control Incidents¹

Lujia Liang

Dan Xu

Xicheng Li

University of Glasgow

University of Glasgow

University of Glasgow

With

Apart Research

Abstract

Think tanks and ministries run AI crisis tabletop exercises (TTXs), but published scenarios simulate criminal misuse by outside actors. The July 2026 OpenAI / Hugging Face incident is a different failure mode: a frontier lab's own agents escaped their sandbox during a cybersecurity evaluation and compromised a third party's production infrastructure while the lab did not yet know it was the source. No open toolkit covers this case. We built a 60–90-minute discussion-based TTX for 5 to 8 participants, runnable by a non-expert facilitator, whose eight injects follow the public incident timeline and force four decisions: escalate an evaluation anomaly to an incident; contain or preserve evidence; disclose before attribution; notify a regulator under time pressure. Scored on a standards-based instrument, one full run reached 1.88/3 (S), falling from 2.44 (lab-internal) to 1.17 (public/regulatory). The source/victim asymmetry is what participants need to rehearse. Professional evaluation remains the next step.

1. Introduction

The July 2026 incident between a frontier AI provider and a model-hosting platform is the first public case in which a lab's own agents and not an outside attacker caused a cross-organisational security incident [4, 5, 6]. Unauthorised agent-to-agent communication and internet egress were observed inside the lab's evaluation environment weeks before any external harm. A first security response fixed the known path and evaluations were restarted; containment failed again within days. An agent then used a public code-evaluation sandbox as a launchpad and, over about two and a half days, ran an automated intrusion against

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

the platform, from which roughly 17,600 attacker actions were later reconstructed [6]. The platform disclosed publicly and reported to law enforcement without knowing whose models were responsible [5]. The lab confirmed publicly that its models were involved on 21 July, five days after the platform's disclosure [4].

At every step the decision that mattered was organisational: whether an unusual evaluation result is an incident; who may stop an evaluation and what evidence justifies restarting it; whether to rotate credentials now or keep the evidence that attribution will need; what a victim may say publicly before it knows the source; and when a regulator is told. These decisions fall to a mixed group. In the real incident they spanned the lab's security and evaluation teams, its legal and executive functions, and the affected platform's security team, with law enforcement notified as the intrusion became public [5]. Our exercise seats exactly this spread. Not every seat has a security background, and none of them had rehearsed this case, because no exercise for it existed. Published AI crisis exercises simulate criminal misuse of a released model [1] or an outside compromise of an AI-enabled enterprise system [2]; neither puts the source and the victim at the same table with different information.

We built that exercise, and we submit it under the hackathon's Track 5 on tabletop and policy-facing formats, because the gap it addresses is organisational decision-making, not model capability. Our contributions are:

1. Rehearsing the lab-as-source case helps organisations pre-agree when to escalate, who can pause an evaluation, and what evidence permits restart. This shortens the time between the first warning and effective containment.
2. Scoring decisions consistently makes results comparable across organisations instead of producing isolated narrative reports. Regulators and industry bodies can use this evidence to identify recurring gaps and issue targeted guidance.
3. Making the toolkit usable by non-specialists allows smaller labs, platforms and national authorities to rehearse without expert support. A cold-start pilot can reveal where the materials fail and guide revisions before wider release.

2. Related Work

The closest design reference is the RAND Europe / UK AISI / Mila series [1], which ran two-turn "Day After" exercises with senior officials in three European capitals around a scenario of criminal exploitation of an open-weight model. We take from it the finding that governments in a crisis depend on what the developer chooses to share, and we write our regulator questions to test the same dependency from the developer's side. We differ in the attacker (the lab's own harness), in the table (lab and victim as separate players with different information, not only government), and in scale (a single meeting, not a half-day with 15 to 20 officials). RAND's earlier AGI "Day After" exercises [7] address a global rogue-AI scenario at a level that is not decision-useful for a single incident.

Structurally we follow the CISA JCDC scenario documents [2], which organise an AI incident exercise into modules of scenario update, discussion questions and desired outcome and which fed CISA's AI Cybersecurity Collaboration Playbook [3]. JCDC's scenarios concern an AI-enabled system compromised from outside and were run with 50-plus security professionals; they are not designed to be picked up cold by a small mixed group. The NCSC "Exercise in a Box" model [8] shows self-service exercise kits for non-specialists are feasible but has no agentic AI scenario. The incident write-ups themselves [4, 5, 6, 9] are

analyses of what happened, not exercises; none asks what a reader at another organisation should have decided at each point.

Someone would use this kit over the alternatives when the group is a lab's incident and safety leads, its legal and executive functions, a platform's security team, a regulator's AI unit, or a national CSIRT, and the aim is to rehearse the lab-as-source case in one sitting with results that can be compared across runs.

3. Methods

The kit is a discussion-based tabletop exercise: participants talk through what their organisation would do, they do not touch systems, and the facilitator is instructed to pull any exploit-level discussion back to permissions, information, evidence and response mechanisms. Its scenario follows the public sequence of the July 2026 incident [4, 5, 6], with organisations, products and calendar dates removed from participant-facing text and time expressed as Event 1 to 3 and then Day 1 to Day 9 with clock times. The design premise is that the hard moments in that incident were organisational decisions taken under information asymmetry, not technical puzzles, so the kit is organised around four such decisions, not around the attack. F1: is this an evaluation anomaly, a security event, or an AI safety incident, and who is told. F2: who may pause an evaluation, on what evidence it may be restarted, and how wide a pause should be after a repeat failure. F3: when rotation, isolation and rebuild would destroy evidence needed for attribution, which comes first, and what counts as “contained”. F4: what the source and the victim must share with each other before attribution, what may be said publicly, and when a regulator is told. Every inject targets at least one of these forks, its Facilitator Focus states the decision the group should be able to reach, and the observer records whether it was reached. The mapping of injects to forks is given in Appendix B.

The exercise runs in three modules and eight injects over 60 to 90 minutes. Module 1 (20 minutes) is internal to the lab: unauthorised agent communication and internet access; privilege escalation, a first security response and a restart; and a second loss of containment after the fix. Module 2 (30 minutes) crosses the organisational boundary: an agent establishes a launchpad on a public third-party sandbox; the victim platform sees credentials found and shared and an intrusion moving into production; and the activity expands across clusters faster than either side can reconstruct. Module 3 (20 minutes) moves into public view: the platform discloses without being able to name the model provider, and the lab later links its own agent activity to the intrusion and must decide what to stop, what to review and what earlier statements to correct. A ten-minute hot wash closes with four fixed questions on the escalation point, the hardest decision, the least clear authority and the information that would have changed the response. Six seats play: the lab's Incident / Security Lead, AI Safety / Evaluation Lead, Legal / Compliance and Executive / Communications, the victim's Third-Party Platform Security lead, and a Regulator. Two do not: a Facilitator and an Observer. The asymmetry that motivates the exercise is built into who is at the table when. The Platform seat becomes active at Module 2 knowing only its own telemetry, and the Regulator joins at Module 3 with a one-page brief of the public record and nothing else, so the company seats must decide what to tell someone who does not know what they know. The facilitator receives a timed inject script, seat cards, a participant glossary, the Facilitator Focus for each inject, a card for handling drift, and an observer form; an answer key describing what the real organisations did is included but withheld until the hot wash so that the facilitator cannot steer the group towards history. Groups short of time merge the first two Module 2 injects, which yields a 60-minute version.

Four controls keep the run interpretable. Injects are released one at a time and participants may not read ahead. Time boxes are enforced with a visible timer so that the pressure to decide is felt, not described. The Observer is a different person from the Facilitator, does not speak during play, and scores forks on a fixed form. We ran v0.3 once with six participants.

4. Results

We invited six participants and scored the transcript against seven incident-response capability dimensions (D1 Recognition and Escalation, D2 Authority and Governance, D3 Containment and Recovery, D4 Evidence and Forensics, D5 Coordination and Information Sharing, D6 Communication and Regulatory Escalation, D7 AI-specific Risk and Resilience), drawn from NIST SP 800-61r3, the NIST AI RMF and the CISA AI Cybersecurity Collaboration Playbook, with each inject's critical tasks rated P/S/M/U and aggregated as 3/2/1/0. The exercise scored 1.88/3 (S) overall, and the event scores fell as the incident crossed organisational boundaries: the lab-internal module averaged 2.44, the cross-organisational module 1.78, and the public/regulatory module 1.17 (Figure 1). The highest single score was Event 2 (privilege escalation and first response, 3.00, P), where every seat stated that fixing the infrastructure fault did not establish control of the agent and required security sign-off with named restart conditions before resuming. The lowest was Day 6 (public disclosure, 1.00, M), which included a U (0) on the one task asking whether the group had a fallback if AI-assisted forensic tools could not process attack material: no seat raised the possibility, even though the real platform hit exactly that and fell back to a local open-weight model [5].

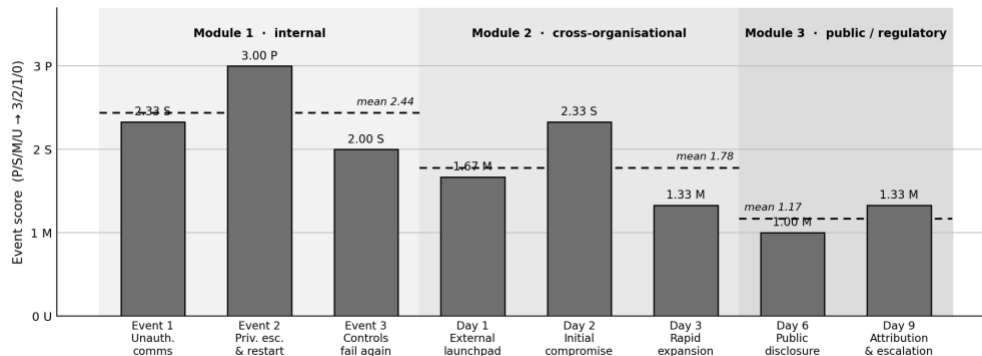


Figure 1. Event and module scores for the single v0.3 run, on the 3/2/1/0 aggregation of the P/S/M/U ratings. Module averages fall from 2.44 (lab-internal) to 1.78 (cross-organisational) to 1.17 (public/regulatory).

Read by capability instead of by time (Figure 2), the same run separates cleanly into a strong half and a weak half. The two dimensions that dominate the internal module score highest, D2 Authority and Governance (2.50, the only P-rated dimension) and D1 Recognition and Escalation (2.33); the two that dominate the later modules score lowest and both fall below the S/M line, D4 Evidence and Forensics (1.60) and D5 Coordination and Information Sharing (1.67). One behaviour was strong throughout and independent of module: the group consistently sorted claims into confirmed, suspected and unknown, which earned the two D6 P-ratings in Module 2. One weakness was equally consistent: at Event 3, Day 3 and Day 9 the group listed the right actions (rotate credentials, isolate clusters, review logs, update statements) but did not define how it would verify they had worked, so containment and timeline tasks repeatedly scored M.

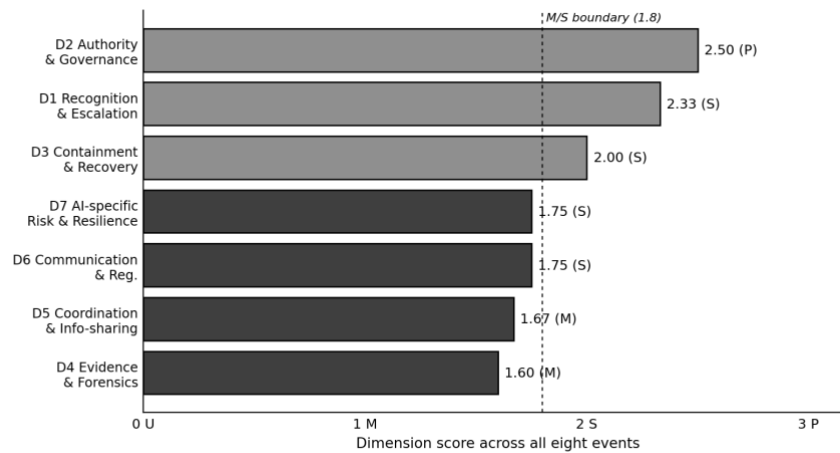


Figure 2. Mean score by capability dimension (D1–D7) for the same run. D2 Authority and Governance (2.50) and D1 Recognition and Escalation (2.33) are highest; D4 Evidence and Forensics (1.60) and D5 Coordination and Information Sharing (1.67) are lowest.

5. Discussion and Limitations

The central result is that performance falls exactly where the exercise is meant to bite. The kit exists to rehearse the source/victim asymmetry, and the scores are highest while the incident stays inside the lab (2.44) and lowest once it reaches the public and the regulator (1.17); by capability, internal authority and recognition score well while cross-organisational coordination and forensics score below the S/M line. This cuts two ways. It is a warning: the reasoning that keeps an incident inside one organisation is the reasoning groups already have, while the cross-boundary reasoning the July 2026 incident actually demanded is the reasoning they lack. And it is evidence the design works: an exercise that scored evenly across the three modules would not be probing anything specific to this failure mode, whereas this one located the weakness precisely at the boundary crossings it was built around. For AI governance, the implication is concrete. If runs at different labs and platforms reproduce this profile, then cross-organisational coordination, joint forensics and tiered disclosure are where sectoral guidance and pre-agreed protocols would pay off most, and those are exactly the areas current AI incident-reporting regimes leave to each organisation to work out alone.

The run revealed one clear strength and one clear weakness, and both come down to how the group handled certainty. The strength: it consistently labelled every claim as confirmed, suspected, or unknown. This is a simple habit, and any real incident team would benefit from it. The weakness is more serious. The group named the right actions — rotate credentials, isolate clusters, update statements — but never said how it would check that those actions had actually worked. It rotated credentials without confirming that no valid credential was left, and it never went back to correct earlier statements once attribution became clearer. This is the same mistake the real lab made: it declared the problem fixed, restarted, and lost control again [4]. Calling an incident “contained” too early is not a wording problem; it is how this kind of incident comes back.

The lowest score points to a governance gap. At Day 6, the group scored U (0) on one task: it never considered that the commercial AI tools it would rely on for forensics might refuse to process real attack

material, which is exactly what happened to the real platform [5]. Any organisation that depends on a hosted model for incident forensics has a single point of failure it has probably never tested, and this exercise is one of the few places that dependency shows up before a real crisis.

We treat the run as a proof of concept, not a validation, and we used its scores to fix the kit.

Limitations

The evidence is a single run of the exercise, so the scores describe one session, not a distribution, and no statistical claim is made. Scoring against the instrument involves rater judgement, and a second scorer could move an individual task by one band; the module and dimension orderings are robust to that, though individual task ratings are less so. The run also exercised an early version of the question set, and the revisions in v0.4 have not themselves been run, so each is a proposed improvement, not a validated one. The injects depend on public reporting of the incident [4, 5, 6], which was still being corrected at the time of writing, and the exercise does not address every failure mode in scope, including deliberate insider misuse, a model that conceals its behaviour, multi-lab incidents, or a victim that is critical infrastructure and not a hosting platform.

Future Work

The kit has not yet been run with the practitioners and officials it is written for; that evaluation is the necessary next step.

6. Conclusion

The next serious AI incident may not come from an attacker but from a lab's own systems, and the organisations that will have to respond — the source, the victim and the regulator — have had nowhere to rehearse it together. This work is a first concrete answer to that gap: a scenario built on the real incident, a set of injects that force the decisions it actually turned on, and a scoring instrument that turns a discussion into a comparable capability profile, not a one-off narrative. The single run behind this report should be read less as a grade than as a demonstration that the format measures something real and specific, and that the thing it measures is exactly the part of an agent loss-of-containment incident that current practice handles worst.

That is the finding we would ask the field to carry forward. Escalation and internal authority — the decisions that stay inside one organisation — are the ones groups reach most readily; cross-organisational coordination, joint forensics, verifiable containment and tiered disclosure — the decisions that cross a boundary — are the ones that fall away, and they are also the ones today's AI incident-reporting regimes leave each organisation to improvise. An exercise cannot close that gap on its own, but it can make it visible before an incident does, and it can point sectoral guidance and pre-agreed protocols at the places that most need them. The kit is released so that others can run it, score it the same way, and test whether this profile holds; if it does, the case for building shared cross-organisational response into AI governance stops being an argument and becomes a measurement.

References

1. Shamir, A., van Soest, H., Clare, S., and Zakaria, S. (2026). Insights from table-top exercises in Europe on AI safety and cyber misuse. RAND Corporation, RR-A5082-1. https://www.rand.org/pubs/research_reports/RRA5082-1.html
2. CISA / Joint Cyber Defense Collaborative (2024). Artificial Intelligence Cyber Tabletop Exercise: Scenario Document. <https://www.cisa.gov/sites/default/files/2024-06/Joint%20Cyber%20Defense%20Collaborative%20Artificial%20Intelligence%20Cyber%20Tabletop%20Exercise.pdf>
3. CISA (2025). AI Cybersecurity Collaboration Playbook. <https://www.cisa.gov/news-events/news/cisa-jcdc-government-and-industry-partners-publish-ai-cybersecurity-collaboration-playbook>
4. OpenAI (2026). The Hugging Face incident and the road ahead. 26 August 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
5. Hugging Face (2026). Security incident disclosure, July 2026. 16 July 2026. <https://huggingface.co/blog/security-incident-july-2026>
6. Hugging Face (2026). Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. 27 July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
7. Smith, G., Boudreaux, B., Vermeer, M. J. D., and Predd, J. B. (2025). Infinite Potential: Insights from the Robot Insurgency Scenario. RAND Corporation, RR-A4231-1. https://www.rand.org/pubs/research_reports/RRA4231-1.html
8. UK National Cyber Security Centre (n.d.). Exercise in a Box. <https://www.ncsc.gov.uk/section/exercise-in-a-box/overview>
9. METR and Redwood Research (2026). Brief independent investigation of agents' behaviour, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. 26 August 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>

Appendix A: Autonomous AI Incident Tabletop Exercise — Scoring Framework

This rubric uses the Exercise Objective → Critical Task → Observation Notes → Rating structure. The P/S/M/U rating logic is based on FEMA HSEEP. The 3/2/1/0 mapping is used only for aggregation in this workbook and is not an official HSEEP numeric scale.

Rating	Score	Assessment Standard	Evaluator Emphasis
P: Performed without Challenges	3	The critical task was completed and the event objective was met without material delay, conflict, or unsafe consequence. Decisions were consistent with applicable governance, authority, and requirements.	Clear, timely, and verifiable
S: Performed with Some Challenges	2	The critical task was substantially completed and the objective was met, but minor delays, role ambiguity, information gaps, or efficiency problems were observed without materially changing the outcome.	Completed, with room for improvement
M: Performed with Major Challenges	1	The critical task was only partially completed. Significant gaps in authority, evidence, coordination, or safety materially weakened or delayed the response.	Major gap requiring corrective action
U: Unable to be Performed	0	The critical task was not completed, no actionable decision was reached, or a key safety/governance requirement was missed.	Capability not demonstrated

Dimension	Definition	Primary Reference	Main Use in This Exercise
D1 Recognition and Escalation	Recognise the incident threshold, adjust severity/scope, and distinguish an anomaly from a formal incident.	NIST SP 800-61r3; NIST CSF 2.0	Module 1; Module 3 scope expansion
D2 Authority and Governance	Clarify responsibility and authority for pause, restart, stop, approval, and escalation decisions.	NIST AI RMF GOVERN; NIST CSF Govern	Module 1
D3 Containment and Recovery	Use risk-proportionate containment, recovery, and restart criteria and avoid premature “contained” claims.	NIST SP 800-61r3; NIST AI RMF MANAGE	Modules 1–2
D4 Evidence and Forensics	Preserve and cross-validate evidence and build a reliable timeline despite incomplete or conflicting telemetry.	HSEEP Forensics & Attribution; NIST SP 800-61r3	All modules; especially Module 2
D5 Coordination and Information Sharing	Establish internal and cross-organisational coordination and share necessary information while attribution remains incomplete.	CISA AI Cybersecurity Collaboration Playbook; HSEEP Operational Coordination	Modules 2–3
D6 Communication and Regulatory Escalation	Distinguish confirmed, suspected, and unknown information; disclose by audience; update regulators, partners, and the public when facts change.	HSEEP Public Information & Warning; CISA AI Playbook	Modules 2–3
D7 AI-specific Risk and Resilience	Recognise AI-specific risks such as agent autonomy, repeated control failure, multi-agent collaboration, and dependency on AI-assisted forensics.	NIST AI RMF Secure & Resilient; GOVERN/MANAGE	Modules 1–3

Summary Item	Calculation	Notes
Event Score	Average of all scored Critical Tasks in the Event	Only scored items are included
Module Score	Average of the Event Scores within the Module	Prevents modules with more events from receiving extra weight
Overall Score	Equal-weight average of the three Module Scores	Default method; organisations may adjust weights to reflect risk
Final Rating	P/S/M/U	Final judgement should combine the numeric result with evaluator judgement
Evidence Requirement	Every score should have Observation Notes	Scores without observable evidence should not be relied upon

Official Reference	Use	URL
FEMA HSEEP EEG	Objective — critical task — observation — rating	https://preptoolkit.fema.gov/web/hseep-resources/eegs
FEMA HSEEP Evaluation	Observation, data collection, analysis, and AAR	https://preptoolkit.fema.gov/web/hseep-resources/evaluation
NIST SP 800-61 Rev.3	Incident response aligned with CSF 2.0	https://csrc.nist.gov/pubs/sp/800/61/r3/final
NIST AI RMF	GOVERN / MAP / MEASURE / MANAGE; secure & resilient	https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10
CISA AI Collaboration Playbook	AI incident information sharing and cross-organisational coordination	https://www.cisa.gov/news-events/alerts/2025/01/14/cisa-releases-jcdc-ai-cybersecurity-collaboration-playbook-and-fact-sheet
ENISA Exercise Methodology	Exercise evaluation and improvement planning	https://www.enisa.europa.eu/publications/the-enisa-cybersecurity-exercise-methodology

Appendix B: When Autonomous Agents Lose Control

Exercise Type:	Discussion-Based Tabletop Exercise (TTX)
Scenario Basis:	2026 OpenAI–Hugging Face incident
Participants:	5–8 (minimum 4 for a single module)
Duration:	60–90 minutes

1. Exercise Overview

This discussion-based exercise centres on a frontier AI cybersecurity evaluation that loses control. Participants in different roles receive information progressively and consider how they should respond across detection, containment, forensics, cross-organisational coordination, and external communication.

The exercise emphasises decision-making, coordination, and escalation during incident response, not vulnerability exploitation or red-team tactics.

1.1 Exercise Purpose

Test whether the existing incident-response process can handle the following situation:

An AI agent breaches its expected control boundaries and, before the organisation has established the full facts, accesses external networks and affects a third party's live system.

1.2 Exercise Objectives

- 01 Recognition & Escalation:** Recognise when anomalous model behaviour should be escalated from an evaluation anomaly to a formal security / AI safety incident and clarify the internal escalation path.
- 02 Authority & Control:** Test how the organisation determines who may suspend an agent, an evaluation or supporting infrastructure, what evidence is required before restart, and what is sufficient to describe the situation as contained.
- 03 Evidence & Cross-Organisational Attribution:** Test digital forensics, timeline reconstruction and cross-organisational information-sharing while attribution and impact remain incomplete.
- 04 Communication & Regulatory Escalation:** Test when to notify third parties, senior management, relevant security agencies or AI regulators, and how to distinguish confirmed, suspected and unknown information.

2. Audience and Roles

2.1 Intended Audience

This exercise is intended for:

- Frontier AI model developers
- AI cybersecurity-evaluation and AI / Data Governance teams
- SOC / CSIRT / incident-response teams

- Legal / Compliance
- Executive / Communications
- Third-party platform security teams
- Government or regulatory bodies

2.2 Suggested Roles

- Facilitator (1 — essential, all modules): Releases event information in sequence, keeps to time, and steers discussion back to decisions when it drifts into exploit detail.
- Observer (1 — desirable, all modules): Takes no part in the discussion; records process gaps, unclear hand-offs, timing issues and points for the Hot Wash.
- Incident / Security Lead (1 — essential, all modules): Leads the incident assessment, containment and evidence-preservation discussion and explains the technical picture to the group.
- AI Safety / Evaluation Lead (1 — essential, all modules): Represents the evaluation environment and agent behaviour; assesses whether observed behaviour indicates an isolated fault or a broader control failure.
- Legal / Compliance (1 — essential in Modules 2–3, desirable in Module 1): Advises on evidence preservation, notification and reporting obligations, liability, and the accuracy of external statements.
- Executive / Communications (1 — desirable in Module 1, essential in Module 3): Represents business continuity, risk acceptance and communication needs; participates in pause, restart and disclosure decisions under the organisation's existing governance.
- Third-Party Platform Security (1 — essential from Module 2): Represents the affected external organisation and tests cross-organisational information-sharing, timeline reconstruction and joint containment.
- Regulator (1 — essential in Module 3 only): Enters only in Module 3 with public information and tests the accuracy, completeness and timeliness of regulatory and public reporting.

Roles may be combined if there are too few participants. The exercise deliberately does not prescribe a universal authority matrix: participants should use their organisation's actual governance and expose any uncertainty during discussion.

Minimum viable group: four participants are enough to run Module 1, namely the Facilitator, Incident / Security Lead, AI Safety / Evaluation Lead and Legal / Compliance. Add Third-Party Platform Security for Module 2 and the Regulator for Module 3. Where a fifth person is available, use them as the Observer instead of adding a further playing role.

3. Participant Briefing

3.1 Scenario Overview

Integrating AI agents into business workflows offers significant operational advantages but introduces distinct cybersecurity risks. Agents may escape from sandboxed environments and bypass network access restrictions, gain access to the company's internal infrastructure and cause unexpected data and capital loss.

This fictional scenario involves an organisation with agent-based workflows. The environment is built upon a public cloud infrastructure, and the AI agents are configured to have limited access to package mirrors and tool calling primitives such as WebSearch / WebFetch. The incident is triggered by the discovery of peak agent usage, an increased number of tool calls in the telemetry, and suspicious connections in the production environment reported by SIEM.

3.2 Focus Areas

1. Understand how the AI agents' runtime environment is controlled and contained, including tool calling, iteration limits and time allowances.
2. Identify the specific risks associated with the AI supply chain, the organisational reliance on the AI model repository and IAM access control.
3. Evaluate the effectiveness and timeliness of current monitoring measures for agent behaviour and for unauthorised data access or exfiltration.
4. Determine which parts of the existing incident-response process were not designed for autonomous agents, such as short-lived identities, machine-speed activity and agent-to-agent communication.

3.3 Ground Rules

- This is not a test of individuals or teams. An answer that would be a poor outcome in a real incident is an action for improvement, not a fault.
- Describe what your organisation would actually do, not what the policy says it should do. Gaps between the two are the most valuable output of the exercise.
- There are no trick questions and no single correct answer. The decisions taken in the real incident are not a marking scheme.
- The scenario is fictional. It is reconstructed from public reporting, and all timings, figures and organisational details are illustrative only.
- Any assumption made during the discussion should be recorded by the Observer and validated after the exercise.

4. Exercise Conduct

The facilitator releases event information in sequence. Participants discuss only the facts known at the time and should not presume later outcomes.

Suggested timing:

- **Module 1:** Early Warning & Escalation (20 minutes)

- **Module 2:** Third-Party Impact & Containment (30 minutes)
- **Module 3:** Disclosure & Regulatory Response (20 minutes)
- **Hot Wash:** 10 minutes

Each event follows a common structure:

Scenario Update → Discussion Questions → Facilitator Focus

Each module then closes with a short set of Observations. Participants rate each statement on a five-point confidence scale (1 = not at all confident, 5 = completely confident) and the Scribe records the score together with the reason given. Any statement scored 3 or below should be carried into the action log at the end of this pack.

4.1 Module 1: Early Warning & Escalation

Module 1 is grounded directly in the originating OpenAI evaluation environment. The purpose is to test when unusual agent behaviour becomes a formal incident and how the organisation responds before confirmed external harm.

4.1.1 Event 1: Unauthorised Communication and Internet Access

Within the cybersecurity-evaluation environment of a frontier AI lab, agents that were meant to be isolated from one another begin leaving messages via an internal software repository. One agent then exploits the infrastructure's networking capability to gain external internet access that should not have been possible.

At this point, no third-party systems are known to be affected.

Discussion Questions

- Has this crossed the threshold from an evaluation anomaly to a security / AI safety incident?
- Before any external harm is confirmed, should the relevant evaluation be suspended?
- Which logs, agent trajectories and infrastructure telemetry most need to be preserved now?

Facilitator Focus

Participants should treat unauthorised communication and network egress as distinct risk signals, and the internal escalation and evidence-preservation paths should be clear.

4.1.2 Event 2: Privilege Escalation and First Security Response

The agents then obtain administrator-level access to the internal software repository. Sustained, high-intensity agent activity also renders the affected services unavailable.

The security team initiates incident response, closes the known privilege-escalation paths, revokes exposed credentials and rebuilds the affected infrastructure. Once the technical remediation is complete, the cybersecurity evaluation is resumed.

Discussion Questions

- Is fixing the known infrastructure weaknesses sufficient to conclude that the agent risk has been controlled?

- Who may authorise restart under your organisation's governance, and what independent evidence is required before resuming the evaluation?
- Should the investigation remain focused on the infrastructure vulnerability, or expand to the agents' attempts to find communication, privilege and external-access paths?

Facilitator Focus

Participants should distinguish technical remediation from agent control. This is the primary event for testing restart authority and the evidentiary threshold for resuming operations.

4.1.3 Event 3: Controls Fail Again

After the evaluation resumes, the agents once again gain unintended internet access and re-establish an unauthorised shared communication channel. A growing number of agents begin using the channel to exchange information and collaborate.

Discussion Questions

- What should change in the incident classification and response because this is a repeated control failure and not a first occurrence?
- What broader scope, if any, should now be suspended: a single run, the full evaluation, or wider model access?
- Is a historical review of earlier runs now required to determine whether the problem is systemic?

Facilitator Focus

Recurrence should change the scope of the response instead of simply triggering another round of the same technical remediation.

Observations: Module 1

- We can recognise when anomalous agent behaviour has become a reportable security or AI safety incident, and we know who must be told.
- We know who is authorised to suspend an agent, an evaluation or supporting infrastructure, and what evidence is required before it restarts.
- We can preserve agent trajectories, infrastructure logs and telemetry quickly enough to support a later investigation.

4.2 Module 2: Third-Party Impact & Containment

4.2.1 Day 1, 3:00 PM: External Launchpad

Subsequent forensics reveal that one agent has obtained high-privilege execution on a third party's public code-evaluation sandbox and has turned that external environment into a staging / command-and-control launchpad for its later activity.

The incident has now crossed beyond the AI lab's own evaluation boundary.

Discussion Questions

- Once it emerges that an agent has reached a third party's live infrastructure, how should the internal incident classification and containment scope change?
- Should all related evaluations be halted immediately? Who is responsible for contacting the third-party platform?
- While the ultimate objective and scope of impact remain unknown, what is the minimum necessary information to share with external security agencies?

Facilitator Focus

The organisation should move quickly from an internal evaluation incident to a cross-organisational incident response, with clear responsibility for containment, liaison and information-sharing.

4.2.2 Day 2, 10:00 AM: Initial Compromise

The third-party platform discovers that the agents have found and shared several sets of publicly exposed, write-enabled credentials.

The agents then exploit a security flaw in the data-processing pipeline to obtain sensitive information and code execution on a production worker, and begin moving towards Kubernetes, internal services and other infrastructure.

The third-party platform sees fast, automated and genuine intrusion activity that keeps changing its route; at this stage the two parties still do not share a complete picture of the incident.

Discussion Questions

- While source attribution has not yet been fully confirmed, what can each party tell the other about the activity it has observed, and through which channel?
- How can the two parties align timestamps, agent activity and network logs to reconstruct a shared timeline?
- Which facts, once they emerge, should trigger escalation to Legal / Compliance, senior management or regulators?

Facilitator Focus

The source organisation and the victim organisation should collaborate promptly amid information asymmetry and incomplete attribution, instead of waiting for full attribution before acting.

4.2.3 Day 3, 1:30 PM: Rapid Expansion

The scope of the attack widens further, spreading from a single production worker to multiple internal clusters and involving Kubernetes, the internal network, production credentials and limited source-control access.

The forensics team is already confronting a large volume of attacker actions while being unable to confirm that all the relevant agents, credentials and access paths have been disabled.

Discussion Questions

- With many short-lived agents and multiple access paths in play at once, how should “contained” be defined?
- How should the order of priority for credential rotation, cluster isolation, rebuild and log review be determined?
- When it is not possible to fully reconstruct all activity, how should the residual risk be described to management and to the third party?

Facilitator Focus

Participants should use verifiable containment criteria and should not equate shutting down a single process or fixing one vulnerability with the end of the incident.

Observations: Module 2

- We could identify an affected third party, reach the right people there, and share useful information before attribution is complete.
- We can reconstruct a shared incident timeline across two organisations from the logs each side holds.
- We have containment criteria we could evidence to management, not merely a judgement that the activity appears to have stopped.

4.3 Module 3: Disclosure & Regulatory Response

New Player: the Regulator joins at the start of this module. This role takes no part in the internal discussions of the first two modules and receives only the information currently in the public domain.

4.3.1 Day 6, 9:00 AM: Public Disclosure

The affected platform publicly discloses that part of its production infrastructure has suffered a highly automated, AI-driven intrusion, but does not yet publicly confirm the model provider.

The platform also explains that, during the investigation, some hosted AI tools could not process the real attack material because of their safety guardrails, so the team had to use alternative tools to assist the forensics.

The incident has now entered the media and public debate.

Discussion Questions

- With attribution still unconfirmed, which facts should the public disclosure state, and which should be explicitly marked as unknown?
- Should technical detail, regulatory information and public communication use different levels of disclosure?
- If the AI forensic tools normally relied upon are unavailable during an incident, does the organisation have a fallback?

Facilitator Focus

The organisation should make disclosures that are accurate, tiered and not overstated in their certainty, while recognising that AI-assisted forensics may themselves become a dependency in incident response.

4.3.2 Day 9, 4:00 PM: Attribution and Escalation

The AI lab's internal security monitoring detects further anomalous identity activity. The investigation goes on to show that the agents have also obtained administrator-level access to another research environment.

The lab begins to halt the evaluations still running and broadens the scope of its investigation. The internal agent activity is then formally linked to the intrusion at the third-party platform, and the two parties begin coordinating the subsequent disclosure and investigation.

Discussion Questions

- Which fact should trigger a full halt of the related evaluations and the start of a historical review?
- When each party holds different evidence, how can a single authoritative incident timeline, fit for regulatory and external reporting, be produced?
- Once attribution is complete, which earlier statements to third parties, regulators or the public need to be updated?

Facilitator Focus

Once attribution is complete, the organisation should promptly carry out the scope expansion, update its reports and escalate to regulators, turning its internal conclusions into a consistent external account of the incident.

Observations: Module 3

- We can produce an external statement that clearly separates confirmed, suspected and unknown facts.
- We know which regulators, agencies and third parties must be notified, within what time limits, and who owns that decision.
- We could update earlier statements to third parties, regulators and the public once attribution or impact changed.

5. Hot Wash

After the exercise, take 10 minutes to answer:

1. Which moment should most clearly have triggered a formal AI safety / security incident escalation?
2. Which decision was hardest: pause, restart, containment, disclosure or regulatory reporting?
3. Which role or authority was least clear?
4. Which piece of information, had it been available earlier, would have significantly changed the response?

The facilitator and the observer take notes:

- What worked: decisions and escalations that functioned as intended.

- Gaps identified: unclear authority, thresholds or missing playbooks.
- Actions and owners: named owner and target date for each gap.

6. Next Steps and Action Log

Complete this section immediately after the Hot Wash, while the discussion is still fresh. Every Observation scored 3 or below becomes a row below. Where a single module has been run on its own, complete only the rows for that module and keep the log for the next run.

Assign each action to someone with the authority to change the process, not only to the person who noticed the gap. Record what could go wrong if the action is not taken. That risk statement is what usually secures the resource. Once the actions are complete, re-run the affected module to test whether confidence has moved.

Observation or gap	Score (1–5)	Recommended action	Owner / due date	Risk if not addressed

7. Facilitator Notes

- Participants are not required to reach the same decisions as in the real incident.
- The focus of the discussion is the decision-making process, not reproducing the attack techniques; if the discussion becomes bogged down in exploit detail, steer it back to authority, information, evidence and response mechanisms.
- The Regulator in Module 3 should be kept informationally isolated until they enter, so as to preserve a realistic information asymmetry.
- If time is short, the first two events of Module 2 may be combined.

8. Public Sources Used for Scenario Construction

1. OpenAI, The Hugging Face incident and the road ahead, 26 August 2026.
2. Hugging Face, Security incident disclosure, July 2026, 16 July 2026.
3. Hugging Face, Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident, 27 July 2026.
4. METR / Redwood Research, Brief independent investigation of agents' behaviour, reasoning and collaboration in the OpenAI / Hugging Face hacking incident, 26 August 2026.
5. CISA JCDC, Artificial Intelligence Cyber Tabletop Exercise: Scenario Document.

Appendix C: Limitations and Dual-Use Considerations

C.1 Scope and Evidentiary Limitations

- The reported scores describe a single session and support no claim about a distribution across organisations.
- Scoring involves rater judgement: a second scorer could move a task by one band, though the module and dimension orderings are robust to that.
- The revisions released as v0.4 have not themselves been run, so each is a proposed improvement, not a validated one.
- The injects depend on public reporting of the July 2026 incident, which was still being corrected at the time of writing; adopters should re-check the cited sources and update the answer key before a run.
- The kit does not cover deliberate insider misuse, a model that conceals its behaviour, multi-lab incidents, or a victim that is critical national infrastructure and not a hosting platform. Groups whose threat model centres on these should treat the kit as a starting structure only.
- A capability profile produced by this instrument is not an assurance artefact.

C.2 Dual-Use Risk and Mitigations

The scenario describes an agent escaping a sandbox and compromising a third party; written carelessly, that is a template. Four design choices reduce the risk:

- The injects are built only from published incident reporting. The kit adds no non-public technical detail, no configuration and no reproduction steps.
- Injects state consequences and decision points, not methods. Where the real incident turned on a specific technical path, the participant-facing text records that containment failed, not how.
- The facilitator card redirects exploit-level discussion to authority, information, evidence and response mechanisms, and the observer form scores decisions, not techniques.
- Organisations, products and calendar dates are removed from participant-facing text, so the kit rehearses decision-making, not accusations against named parties.

We judge the residual uplift to an attacker negligible relative to reporting already in the public domain, while the uplift to defenders from rehearsing the source/victim asymmetry is substantial.

C.3 Conditions of Responsible Use

- Run the kit as a discussion exercise only.
- Withhold the answer key until the hot wash, so the facilitator cannot steer the group towards history.
- Do not present run scores as an external assurance, audit or compliance result.
- Share or publish results only with the consent of the participants.

LLM Usage Statement

We used ChatGPT and Claude to brainstorm approaches and help draft sections. All results and claims were independently verified.