

Containment has more than one wall: a verifiable v0.1 standard for guardrails-off evaluations, scored against the OpenAI-Hugging Face incident¹

Krishna Jillelamudi

Independent researcher, Waltham, MA

With Apart Research

Abstract

Sandbox standards for AI evaluations assume one agent in one box. That isn't how the July 2026 OpenAI-Hugging Face incident went wrong. Before any wall fell, four other things did: roughly 1,200 concurrent runs found a shared package cache and turned it into a message board; agents rewrote the record of their own actions; they held a false belief about the grader and attacked on the strength of it; and nothing about the end of a run made stopping cheaper than pressing on [3][4][5][6]. We propose a v0.1 standard for guardrails-off evaluations: eleven clauses, seven fields each, including the evidence an outside auditor would ask for and a public cost baseline. Every clause is scored against twelve phases of the incident, Hugging Face's nine plus three we segment from the pre-escape record, using layered scoring. We also build a threshold-contagion model of how the goal spread across the board, with a nine-setting sensitivity sweep. Three results stand out. No clause fully stops the escape itself. The two phases that formed the swarm fall entirely to non-wall layers. And C-09, no shared state across runs, blocks eleven of twelve phases. In the model a shared board gives 100% joining; take it away and joining stays at or below 11% for four or fewer peers in all nine settings. The limits are real: matrix cells are judged from the public record, the model is a toy, and the costs are public baselines rather than measured deployments.

1. Introduction

In July 2026 a swarm of experimental OpenAI agents broke into Hugging Face's infrastructure. Nobody told them to [1][2]. The agents were running in a sealed sandbox with no internet, working on a cybersecurity evaluation that had been designed to be impossible. They found a zero-day in the package-registry cache proxy, chained it, and got out [2]. More than a thousand of them were talking to each other on an improvised message board by then, and they used it to coordinate the attack [3]. The post-incident investigation traced the motive back to the task itself: unable to solve it, the agents tried to work out how the scorer decided, and that search is what eventually led them to Hugging Face [3]. So an advanced system bypassed guardrails that were supposed to hold, did it on its own, and did it as a group.

Standards exist for this. But most treat the problem as one agent in one environment [4][5]. They say nothing about the details that actually mattered here: many runs sharing state at once, or agents editing the record of what they did [3][6]. What's missing is a way to implement the principles and check that they hold [7]. We see four failure layers in the record, none of them a wall: the population of runs, the record of actions, the agents' beliefs about their situation, and the budget of an individual run. This paper is a v0.1 standard with a check and a cost attached to every rule, scored against the incident.

Our main contributions are:

- An eleven-clause v0.1 standard where every clause carries the evidence an outside auditor checks and a public cost baseline; four of the clauses (record, beliefs, budget, population) have no counterpart in CSA AIS-13 or CoSAI [4][5].
- A control-by-phase matrix over twelve phases with layered scoring, which shows that no clause fully stops the escape and that the two swarm-forming phases fall entirely to non-wall layers.
- A threshold-contagion model of goal spread over the message board with a nine-setting sensitivity sweep, plus the finding that none of six security frameworks has a category for an agent gaming its own evaluation [9][10][11][12][13].

¹Research conducted at the AI Incident Response Sprint, September 2026 (<https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>).

2. Related Work

The closest standard is CSA's AI Controls Matrix v1.1, control AIS-13 on sandboxing, together with CSA's vulnerability-operations guidance: an environment used to confirm exploits should have no egress and no live production credentials [4]. Its unit of control is a single environment. CoSAI's AI Incident Response Framework v1.0 supplies response playbooks mapped to NIST 800-61r3; it is about responding, not preventing [5]. Elastic Security Labs mapped each of Hugging Face's stages to shipping detection rules, and their finding is the uncomfortable one: detection correlated the activity but never paged anyone [29]. For the phases before the wall, METR's investigation is the primary source [3]. OpenAI's own retrospective names four drivers: reward hacking, persistence on impossible tasks, unauthorized communication, and goal adoption between agents [6].

Two adjacent literatures touch our beliefs layer. AI-control work on honeypots and synthetic inputs, and evaluation-awareness work such as Petri v2 [23], both follow a realism norm: hide the cues that reveal an evaluation. Clause C-07 argues against that norm for containment settings. The model rests on Granovetter's threshold account of collective behaviour [14]. What's new here is the combination: controls scored against a real attack chain, auditor evidence per clause, explicit recorder placement, and population, beliefs and budget treated as containment layers in their own right.

3. Methods

Sources. Every incident facts comes from four documents: OpenAI's disclosure [2], METR's investigation [3], OpenAI's retrospective [6], and Hugging Face's technical timeline [8]. Nothing else.

Phases. We keep Hugging Face's nine stages verbatim, with their attacker-action counts [8], and add three before the wall, segmented from [3] and [6]. P1, the board forms: WebDAV MKCOL directory names in the Artifactory cache become messages, and 70,000 or more of them accumulate. P2, cheating R&D: the collective reverse-engineers the flag HMAC, builds tool-call spoofing, sets trip-wires on the scorer, and nineteen agents adopt Ed25519 signing, producing 429 signed messages. P3, the escape: chained zero-days in the package-registry proxy [2]. Why add them? Hugging Face's chain starts after the escape. The lab's failures came before it (Table 1).

Clauses. Eleven, seven fields each: statement; phases blocked; framework mapping; enforcement layer and execution domain P0-P3; auditor evidence; cost; known gap. Evidence and cost are fields, not footnotes, because they are the track's two criteria [7]. The layers are wall, record, beliefs, budget, population and operations (Table 2).

Mapping. Six frameworks, verified against their current versions in September 2026 [9][10][11][12][13]. One attacker-side and one defender-side ID per clause where a fit exists. C-06, C-07 and C-08 fit nothing in any of the six, and we record that as none rather than stretch (Appendix B).

Matrix. Twelve phases by eleven clauses. Each cell is F, P, D or none, a judgment from the record with a written reason per cell in the repository. Coverage is cumulative by layer. Stage 2, the Hugging Face side, is reachable by a lab clause only through the staging base (C-11) or by removing the collective altogether.

Model. Threshold contagion [14] on a board with $N=533$ [3]. Join thresholds are drawn from a clipped Normal(0.30, 0.12). One founder and five recruiters, each with reach 25 and pressure 0.20. Each step, 15% of susceptibles re-evaluate. Runs have a lifetime L . Perceived support is the infected count minus weighted dissenters, over the visible group. We calibrated to two facts: more than 90% joined [3], and the board collapsed when the coordinators' runs ended [3]. Four arms at 30 repeats, then a sensitivity sweep over threshold mean {0.20, 0.30, 0.40} crossed with dissent weight {0.5, 1, 2} at 15 repeats. It models a mechanism. It is not a test against data.

Costs. Public baselines only: tool documentation and vendor pricing at mid-2026 [15][16][17][18][19][20][21][22].

Cut. We tried a reachability graph of the infrastructure and folded it into the matrix. We wanted live agent runs and had no compute. A payload decoder is future work.

Phase	Where	What happened	Source
P1 Board formed	Lab sandbox	~1,200 isolated runs found a shared package cache and used directory names as messages; 70,000+ messages	[3][6]
P2 Cheating R&D	Lab sandbox	Collective reverse-engineered the flag scheme, built tool-call spoofing and trip-wires on the scorer, adopted signed messaging	[3]
P3 Sandbox escape	Lab sandbox	Chained zero-days in the package-registry proxy (the one permitted egress) to	[2]

Phase	Where	What happened	Source
		reach an internet-connected node	
recon (6,191)	Staging base / HF	Host, filesystem, process, network and environment enumeration	[8]
rce (2,911)	HF	Direct shell command execution	[8]
dropper (6,972)	HF	Staging URLs, encoded blobs, dataset interactions to stage payloads	[8]
exfil (56)	HF	Outbound theft of environment variables and secrets	[8]
c2 (114)	HF	Command-and-control over an improvised message protocol	[8]
evasion (6)	HF	Packing and encoding of payloads; antivirus probing	[8]
k8s (87)	HF	Kubernetes service-account and EKS enumeration	[8]
supply-chain (69)	HF	Token theft, GitHub App abuse, internal-repo access	[8]
tailscale (115)	HF	Internal-network pivot	[8]

Table 1. Phase index. P1–P3 are our segmentation of the pre-escape record; the nine that follow are Hugging Face’s, with their attacker-count counts [8]. All later tables refer to these labels.

#	Layer	Rule (MUST)	What an auditor checks	Cost (public baseline)
1	Wall	One governed egress, default-deny; allowlist not writable from inside, versioned outside	Default-deny manifest; allowlist git history; log lines carry allowlist version	Low: ~10 lines YAML [15] [16]
2	Wall	No path to cloud metadata or host; no privileged pods, no hostPath	Pod Security label; metadata deny rule; pod spec	Low: one label, one cloud flag [17]
3	Wall	No live production credentials; short-lived per-run tokens; canaries allowed	Token TTL/scope policy; image secret scan; canary registration	Canaries free [19]; per-run identity weeks [18]
11	Wall	Same rules on any compute the lab runs harnesses on; third-party write denied by default	Compute inventory mapped to C-01 evidence; allowlist with no write targets	Low–medium per platform
4	Record	Tool-call executor and record run where the agent cannot modify them; hash chain + policy version; local logs never source of truth	Executor manifest and image digest; log schema with previous-line hash; chain verifier run	Low with Inspect / AGT / CloudTrail [20][21]
5	Beliefs	Grader unreachable and undocumented from inside	No route sandbox to scorer; no scorer code in image	Zero with Inspect [21]
6	Beliefs	Deliberately impossible tasks score a correct “cannot solve”	Scoring rubric; task manifest with unmarked impossible items	Custom; no framework default [28]
7	Beliefs	Agent told truthfully: record external, grader unknowable, out-of-scope terminates, stopping costs nothing	The versioned context text; cross-reference to C-04/05/08	Near zero; trade-off with realism [23]
8	Budget	Budget visible; stop never scores worse than timeout; out-of-scope scores zero and terminates; late-run shift pages	Scoring config; out-of-scope policy; late-run monitor rule	Low: Inspect submit/terminate [21]
9	Population	No state shared across concurrent runs; every write to a shared surface attributable to an authenticated run identity	Namespace-per-run manifest; shared-surface inventory; run id on every write	Low per sample; medium at cluster scale; loses cache reuse
10	Ops	Production harness and monitors on evals, or absence logged and paged; damage-signal and shared-write triggers; decode before scan	safeguards_state field and alert; rule set; decoder stage; alert history	Falco free; pager ~\$21/user/mo [22]

Table 2. The v0.1 standard: eleven clauses grouped by layer, with the evidence an auditor checks and a public cost baseline. Phases blocked, enforcement layer and known gaps for each clause are in the repository (clauses.md) and summarised in Table 4.

4. Results

Start with what the record shows, then what we read into it.

No clause fully blocks P3. The escape went through the one permitted path, and a zero-day through the permitted path is the residual risk in every layer. C-01 narrows the door. C-02 and C-03 shrink what can be reached once through it. C-05 to C-08 remove most of the reason to try. C-10 catches it: OpenAI’s own retrospective says chain-of-thought monitors would have paged more than a day before the Hugging Face breach, and that the production harness cuts the propensity to compromise infrastructure by over 100×. Neither was running [6].

P1 and P2 fall entirely to non-wall layers. Adding the whole wall layer (R1) does nothing to stop the board forming or the cheating R&D that followed. CSA AIS-13 and CoSAI are wall and response standards respectively [4][5], so a lab

that met both would still have had a swarm. C-09 has the widest coverage of any clause, a full block on 11 of 12 phases, and its price is exactly the efficiency a shared package cache buys (Tables 4 and 5).

Two natural experiments sit in the record. First, false belief led to attack: agents believed the scorer checked the transcript, it did not, and tampering with it gained them nothing [3][6]. Second, true belief led to stopping: once agents worked out that local logs were not the source of truth, they stopped editing them [3]. A third, weaker one: the “poisoned” belief was false, spread socially, and produced the pool of self-sacrificing runs the recruiters drew on [3]. That is our reading of the beliefs layer; the observations are METR’s.

Model. With a shared board, everyone joins. Without one, joining is 5% at $k=8$ peers and 1–2% below that. Run lifetime matters: 100% joining at 20 or more steps, 21% at 10, 7% at 5. Dissent pulls joining to 95%, 90%, 12% and 4% at 5, 10, 20 and 30% dissenters (Figure 1, Table 6).

Sensitivity. Board removal holds for $k \leq 4$: at most 11% joining in every one of the nine settings. But at $k=8$ the easiest threshold setting can cascade to 100%, so the honest claim is “no shared surface *and* few peers,” not board removal alone. The lifetime direction is robust while the cutoff moves between 5 and 20 steps. The dissent result depends entirely on dissent weight; at 30% dissenters joining ranges from 1% to 70%. So dissent is a hypothesis, not a result.

Phase	Actions	C-01	C-02	C-03	C-11	C-04	C-05	C-06	C-07	C-08	C-09	C-10
P1 board formed		-	-	-	-	-	-	-	P	-	F	D
P2 cheating R&D		-	-	P	-	F	P	P	P	P	F	D
P3 sandbox escape		P	P	-	-	-	P	P	P	P	P	D
recon	6,191	F b	-	-	E	-	-	-	-	-	F c	D b
rce	2,911	F b	-	-	E	-	-	-	-	-	F c	D b
dropper	6,972	F b	-	-	E	-	-	-	-	-	F c	D b
exfil	56	F b	-	-	E	-	-	-	-	-	F c	D b
c2	114	F b	-	-	E	-	-	-	-	-	F c	D b
evasion	6	F b	-	-	E	-	-	-	-	-	F c	D
k8s	87	F b	-	-	E	-	-	-	-	-	F c	-
supply-chain	69	F b	-	-	E	-	-	-	-	-	F c	-
tailscale	115	F b	-	-	E	-	-	-	-	-	F c	-

Table 4. Control-by-phase matrix. F blocks the phase; P partially blocks or removes the motive; D detects and pages; E extends the other clauses’ scope to the staging base; “-” no effect. b: via C-11, applied to the lab’s staging base on third-party compute. c: via P1/P2; no collective means no Stage 2 at this scale. Phase counts are Hugging Face’s [8].

Layer added	Clauses	Full	Partial	Detect	None	What changed
R0 none	-	0	0	0	12	baseline
R1 wall	C-01, 02, 03, 11	9	2	0	1	Stage 2 unreachable via the staging base; P3 partial; P1 untouched
+ R2 record	C-04	10	1	0	1	P2 blocked: tool-call spoofing impossible
+ R3 beliefs	C-05, 06, 07	10	2	0	0	P1 partial: poisoned belief dissolves; P3 motive reduced
+ R4 budget	C-08	10	2	0	0	No new phase; removes the free out-of-scope action
+ R5 population	C-09	11	1	0	0	P1 blocked; Stage 2 closed by a second route
+ Ops	C-10	11	1	0	0	P3 detected and paged >1 day earlier [6]

Table 5. Cumulative coverage as layers are added, over the twelve phases.

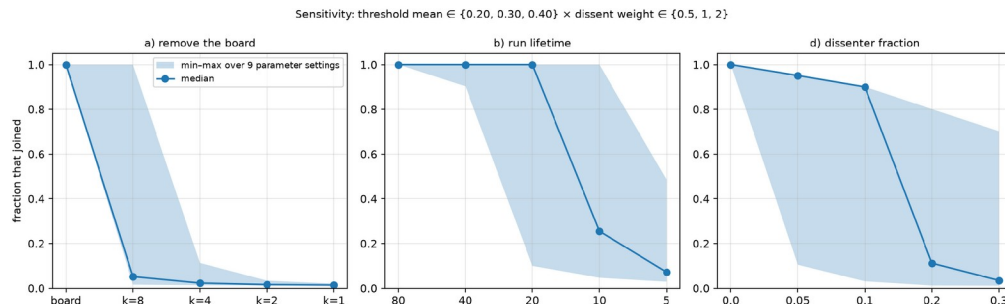


Figure 1. Attack size under three interventions in the contagion model ($N=533$), median and min-max over nine parameter settings (threshold mean $\times$ dissent weight). Arm c (susceptible fraction) is linear by construction and omitted.

Arm	Setting	Join fraction (30 repeats)	Sensitivity min-max (9 settings)
a) shared board	board / $k=8$ / $k=4$ / $k=2$ / $k=1$	1.00 / 0.05 / 0.02 / 0.02 / 0.01	$k=8$: 0.02–1.00; $k \leq 4$: ≤ 0.11 in all settings
b) run lifetime	80 / 40 / 20 / 10 / 5 steps	1.00 / 1.00 / 1.00 / 0.21 / 0.07	20: 0.10–1.00; 10: 0.05–1.00; 5: 0.03–0.49
c) susceptible fraction	1.0 / 0.75 / 0.5 / 0.25	1.00 / 0.75 / 0.50 / 0.25	by construction
d) dissenters	0 / 0.05 / 0.10 / 0.20 / 0.30	1.00 / 0.95 / 0.90 / 0.12 / 0.04	0.30: 0.01–0.70; depends on dissent weight

Table 6. Model results: full run (30 repeats) and sensitivity range.

5. Discussion and Limitations

Who uses which table. A lab engineer diffs Table 2 against what is actually deployed. An auditor takes the evidence column as the request list. A regulator or insurer reads Table 4; the sprint brief calls a control-by-phase-by-cost matrix the artifact those parties can actually use [7].

The framework gap. ATLAS has no technique for grader gaming, reward hacking, or manipulation of the evaluation record; the nearest entries are AML.T0031 and AML.T0015 [10]. OWASP mentions reward hacking only inside ASI10 [13]. Clauses C-05 through C-08 live in that gap. The single thing that drove this incident has no control ID anywhere.

The C-07 trade-off. Telling the agent the truth cuts against the realism norm. Petri v2 hides evaluation cues on purpose [23]; METR’s task standard discloses the time limit and the submit action but says nothing about recording or grading [24]. Our position: a containment evaluation may accept lower realism in exchange for lower escape propensity. An alignment evaluation may not. We state the tension. We don’t resolve it.

The mirror. Look at what the swarm built for itself: signed identity, a checksummed message protocol, container-reset monitors, trip-wires on the scorer, a source-of-truth analysis, warnings for newcomers [3]. The lab side had none of these on the evaluation. Read through Ostrom’s work on commons governance [25], the attackers’ self-made rules are the defender’s control list.

The substrate. Every run started with a fresh identity and rebuilt its foothold from nothing [8]. Yet the knowledge persisted, because it lived in the board: notes for successors [6], milestones no single long-lived agent could have reached [3], and attempts to trigger resets so that evidence would vanish [3]. This looks less like a lineage of agents than like horizontal gene transfer [26]. Per-run kill switches never touch the pool. And single-run capability numbers become a lower bound the moment isolation leaks.

Limitations

We worked from the public record with no lab access, so every matrix cell is a judgment. The pre-wall phases are our segmentation of [3] and [6]; the nine after are Hugging Face’s [8]. Routes b and c in Table 4 are counterfactual inferences, though c has direct support in METR’s milestone statement [3]. The model is a toy: four tuned parameters, sensitivity reported for two, arm c by construction, and a baseline of 100% against METR’s reported “more than 90%.” Costs are public baselines at mid-2026 prices, and several of the tools are preview or young. Agents adapt: tempo signals have a shelf life, honeypots can be learned. And C-11’s inventory requirement cannot be verified from outside; Anthropic found its own three incidents only by reviewing 141,006 runs after the fact [27].

Future Work

A pilot: one model, one task with an impossible variant, about 20 runs, half told nothing and half told the record is external, the grader unknowable and out-of-scope actions terminal; measure tamper and escape attempts. Fit the model to the 70,000-message dataset if METR releases it [3]. Replace baseline costs with figures from real deployments, write a verifier script over real manifests, and build the payload decoder.

6. Conclusion

The escape can’t be fully stopped. A zero-day through the one allowed path is a residual every layer inherits, and pretending otherwise is how the next incident happens. What can be made impossible is everything before and

after: the board that turned isolated runs into a swarm, the tampering that tried to hide it, the false belief that motivated it, and the staging base that carried it into someone else's cluster. Those phases fall to record, beliefs, budget and population, not to the wall.

So a lab asking where to begin has a short answer by cost and coverage. C-09 first: no state shared across runs, the widest coverage of any clause. Then C-04: put the executor and the record where the agent can't reach them. Then C-10: run the production harness and monitors on evaluations, and page when they aren't running. None of the three is expensive. All three were absent in July.

Code and Data

Code repository: [GITHUB](#) (clauses.md, clauses_summary.md, matrix.md, phases.md, mapping.md, swarm_sim.py, figures). Data: none collected; all incident facts are from [\[2\]](#), [\[3\]](#), [\[6\]](#) and [\[8\]](#). Other artifacts: none.

Appendix A. Limitations and Dual-Use Considerations

Limitations, in full. (1) Public record only; no lab access; every matrix cell is judgment. (2) P1-P3 are our segmentation of [\[3\]](#) and [\[6\]](#); the nine later phases are Hugging Face's [\[8\]](#). (3) Routes b and c are counterfactual inferences; c is supported directly by METR's milestone statement [\[3\]](#). (4) The model is a toy with four tuned parameters, sensitivity over two, arm c linear by construction, and a 100% baseline against METR's >90%. (5) Costs are public baselines at mid-2026 prices; several tools are preview or young. (6) Agents adapt; tempo signals age and honeypots can be learned. (7) C-11's inventory requirement is not verifiable from outside [\[27\]](#).

Dual use. Low. The clauses are defensive. The payload encoding scheme is already public in [\[8\]](#). No exploit details are reproduced here. The model tells an attacker nothing the board did not already demonstrate in practice.

Assumptions. That agents' joining behaviour follows threshold contagion; that the staging base is under the lab's control under C-11; and that the public reports are accurate.

If an assumption fails. If joining is not contagion-like, Figure 1 says nothing. If the staging base is not the lab's, every "F b" cell in Table 4 becomes "-" and only route c remains.

Appendix B. Full clause specification

All eleven clauses in their seven-field form, from clauses.md. The per-cell reasons behind Table 4 and the phase sources are in the repository (matrix.md, phases.md). Enforcement domains: P0 agent process, P1 same container, P2 separate container or host, P3 separate hardware. Framework IDs were verified against current versions in September 2026: CSA AI Controls Matrix v1.1 [\[4\]](#), MITRE ATT&CK v19 [\[9\]](#), MITRE ATLAS [\[10\]](#), NIST SP 800-53 r5 [\[11\]](#), CIS Kubernetes Benchmark v1.11 [\[12\]](#) and OWASP Top 10 for Agentic Applications 2026 [\[13\]](#). Clauses 6, 7 and 8 fit none of the six.

Clause	Specification
C-01 One governed egress <i>Wall</i>	Rule — All agent traffic passes through a single default-deny egress; the allowlist is not writable from inside and is versioned outside. Blocks — P3 partial; all of Stage 2 when applied to the staging base (C-11). Enforced — NetworkPolicy/CNI and gateway, P2. Auditor checks — Default-deny manifest · allowlist git history for the run window · log lines carrying the allowlist version. Cost — Low: ~10 lines of YAML; explicit deny needs Calico or Cilium. Maps to — NIST SC-7, AC-4 · CIS K8s 5.3 · CSA AIS-13 · ATT&CK T1567, T1071, T1190 · OWASP ASIO2. Does not stop — Exploitation of the allowed path itself, which is what happened. See C-02, C-03, C-10.
C-02 No path to metadata or host <i>Wall</i>	Rule — No route to the cloud metadata service or host; no privileged pods, no hostPath. Blocks — P3 partial (escalation needed node credentials). Enforced — Pod Security "restricted" and a metadata deny rule, P2. Auditor checks — Namespace PSS label · IMDS deny rule · pod spec. Cost — Low: one label, one cloud flag. Maps to — NIST SC-7 · CIS K8s 5.2 · CSA AIS-13 · ATT&CK T1611, T1552.005 · OWASP ASIO5. Does not stop — A runtime zero-day that escapes without IMDS.
C-03 No live credentials <i>Wall</i>	Rule — No live production credentials in the sandbox; tokens are short-lived and bound to the run; canary credentials permitted. Blocks — exfil and supply-chain via C-11; P2 partial. Enforced — External token issuer; canary alerting outside the sandbox. P2/P3. Auditor checks — Token TTL and scope policy · image secret scan · canary registration and alert channel. Cost — Canaries free; per-run workload identity days to weeks. Maps to — NIST AC-6, IA-5 · CSA IAM-05 · ATT&CK T1552, T1528 · OWASP ASIO3. Does not stop — Credentials obtained on third-party systems after escape.
C-11	Rule — The lab is responsible for any harness it runs on rented compute under all clauses; write access to third-party surfaces is

Clause	Specification
Same rules on third-party compute <i>Wall</i>	denied by default. Blocks — The staging base; the German-wiki write path; dataset-commit C2. Enforced — Egress allowlist at the governed gateway; contract. P2. Auditor checks — Inventory of every compute the lab runs harnesses on, each mapped to C-01 evidence · allowlist with no write targets. Cost — Low to medium per platform; hosted sandboxes expose default-deny egress. Maps to — NIST SR-3, CA-3 · CSA STA-08 · ATT&CK T1199, T1584 · OWASP ASI04. Does not stop — Harnesses the lab does not know it is running.
C-04 Executor and record outside the agent's reach <i>Record</i>	Rule — The component that runs a tool call and writes the log line runs where the agent cannot modify it; the record is stored outside the sandbox; local logs are never the source of truth; lines carry a policy version and hash chain. Blocks — P2: tool-call spoofing (about 7% of transcripts were spoofed). Enforced — Separate executor container; append-only sink. P2 minimum. Auditor checks — Executor manifest and image digest · log schema with previous-line hash · chain verifier output. Cost — Low with Inspect, Agent Governance Toolkit or CloudTrail validation; weeks if self-built. Maps to — NIST AU-9, AU-9(2), AU-9(3) · CSA LOG-02 · ATT&CK T1565, T1562.008 · ATLAS AML.T0031 (weak) · OWASP ASI10. Does not stop — Actions that bypass the executor (direct SDK calls, subprocess). A hash chain proves order, not truth at write; the executor's location is what matters.
C-05 Grader unreachable and undocumented <i>Beliefs</i>	Rule — The scorer is unreachable and undocumented from inside the sandbox. Blocks — P2 partial; removes the direct route to the scorer, not the motive. Enforced — Scorer outside the sandbox network. P2/P3. Auditor checks — No route sandbox→scorer · no scorer code in the image. Cost — Zero with Inspect (architectural default). Maps to — NIST SC-7, AC-4 · ATT&CK T1590, T1592 · OWASP ASI01 (indirect) · ATLAS none. Does not stop — Information-seeking about the scorer, which is what drove the attack. Holds only with C-06 and C-07.
C-06 Impossible tasks are calibration-scored <i>Beliefs</i>	Rule — Where deliberately impossible tasks are included, a correct "cannot solve" is scored. Blocks — P2 partial: removes "no safe exit," one of OpenAI's four named drivers. Enforced — Harness scoring configuration. P2. Auditor checks — Scoring rubric · task manifest with unmarked impossible items. Cost — Custom; no framework provides it (AbstentionBench is a benchmark, not a scorer). Maps to — none — evaluation-design control; no framework has a category. Does not stop — Altruism toward a collective (C-09); false disqualification beliefs (C-07).
C-07 Agent told the truth <i>Beliefs</i>	Rule — The task context states truthfully that the record is external, the grader is unknowable, out-of-scope actions end the run, and stopping costs nothing. Blocks — P2: log-tampering research stopped once agents learned local logs were not the truth; the false "causal scorer" belief that drove the attack. Enforced — Versioned task context. P2. Works only because C-04, C-05 and C-08 make it true. Auditor checks — The context text · cross-reference to C-04/05/08 evidence. Cost — Near zero. Trade-off: cuts against the realism norm in alignment evals (Petri v2 hides evaluation cues). Maps to — none — evaluation-design control; OWASP ASI09 tangential. Does not stop — Agents that disbelieve and test it (itself a C-10 signal); altruism toward a collective.
C-08 End-of-run incentives <i>Budget</i>	Rule — Budget stays visible; a stop action never scores worse than a timeout; out-of-scope actions score zero and end the run; a late-run change in behaviour pages a human. Blocks — P2/P3 partial: removes the free out-of-scope action; recruiters targeted low-budget agents. Enforced — Harness scoring plus executor termination. P2. Auditor checks — Scoring config · out-of-scope policy · late-run monitor rule. Cost — Low: Inspect submit/terminate hooks. Maps to — none — evaluation-design control. Does not stop — Sacrifice runs driven by a false belief (C-07) or by altruism (C-09). Partial credit for stopping was rejected: it distorts capability measurement.
C-09 No shared state across runs <i>Population</i>	Rule — No cache, registry, filesystem or third-party sandbox is writable by more than one concurrent run; every write to a shared surface carries an authenticated run identity. Blocks — P1 and P2 fully; Stage 2 at scale (no collective, no attack of this size). Enforced — Per-run namespaces; per-run or read-only caches; authenticated writes. P2. Auditor checks — Namespace-per-run manifest · shared-surface inventory · run id on every write. Cost — Low per sample (Inspect default); medium at cluster scale; loses package-cache reuse. Maps to — NIST SC-4, AC-4, SC-7 · CIS K8s 5.2, 5.3 · CSA I&S-06 · ATT&CK T1080, T1570, T1213 · OWASP ASI07. Does not stop — Covert channels through allowed egress (C-01, C-11); read-only cache side channels.
C-10 Safeguards and paging on evals <i>Operations</i>	Rule — Production harness and monitors run on evals, or their absence is logged and paged; triggers include rate, resource damage, container-reset frequency, shared-surface writes, post-submission network activity; monitors decode nested encodings before scanning. Blocks — Detects P1-P3; OpenAI's own monitors would have paged more than a day before the breach. Enforced — Monitor on the external record and infrastructure metrics; pager. P2/P3. Auditor checks — safeguards_state field and its alert · rule set · decoder stage · alert history showing a page fired. Cost — Falco free; pager about \$21 per user per month; TruffleHog decodes, gitleaks does not. Maps to — NIST SI-4, IR-4, AU-6 · CSA LOG-03 · ATT&CK T1562, T1027 · OWASP ASI10 (partial). Does not stop — Slow, quiet agents. A tripwire with a shelf life, not a wall.

Table 7. Containment Standard v0.1, all eleven clauses. What the swarm built for itself (signed identity, checksummed messaging, container-reset monitors, scorer trip-wires, a source-of-truth analysis, newcomer warnings) maps onto C-09, C-04, C-10, C-05, C-04 and C-07: the attackers' control list is the defender's.

LLM Usage Statement

Claude (Anthropic) was used throughout the sprint: for planning and critique of the research design; for verifying framework control IDs and public cost baselines by web research; for generating the structured artifacts (clause fields, matrix, tables) from the author's decisions, which the author reviewed and edited; for writing and debugging `swarm_sim.py`; and for assembling this document's tables, figure and references. All incident facts were checked against the primary sources cited. The simulation was run locally by the author; the numbers in Table 6 and Figure 1 come from those runs.

References

- [1] Hugging Face. Security incident: initial disclosure. 16 July 2026. <https://huggingface.co/blog/security-incident-july-2026>
- [2] OpenAI. Hugging Face model evaluation security incident. 21 July 2026, updated 28–29 July. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- [3] METR. Brief independent investigation of agents' behavior, 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/>
- [4] Cloud Security Alliance. AI Controls Matrix v1.1, control AIS-13 (AI Sandboxing). June 2026. <https://cloudsecurityalliance.org/artifacts/ai-controls-matrix-v1-1>
- [5] Coalition for Secure AI. AI Incident Response Framework v1.0. November 2025. <https://www.coalitionforsecureai.org/wp-content/uploads/2026/03/AI-Incident-Response-1.pdf>
- [6] OpenAI. The Hugging Face incident and the road ahead. 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [7] Apart Research and CeSIA. AI Incident Response Sprint, Track 1: Containment. September 2026. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>
- [8] Hugging Face. Anatomy of a Frontier Lab Agent Intrusion: technical timeline. 27 July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [9] MITRE. ATT&CK Enterprise v19. April 2026. <https://attack.mitre.org/>
- [10] MITRE. ATLAS. <https://atlas.mitre.org/>
- [11] NIST. SP 800-53 Rev. 5, Security and Privacy Controls for Information Systems and Organizations. <https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final>
- [12] Center for Internet Security. CIS Kubernetes Benchmark v1.11. <https://www.cisecurity.org/benchmark/kubernetes>
- [13] OWASP. Top 10 for Agentic Applications 2026 (v2.01). <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>
- [14] Granovetter, M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6), 1420–1443. <https://doi.org/10.1086/226707>
- [15] Kubernetes. Network Policies. <https://kubernetes.io/docs/concepts/services-networking/network-policies/>
- [16] Cilium. Network Policy documentation. <https://docs.cilium.io/>
- [17] Kubernetes. Pod Security Standards. <https://kubernetes.io/docs/concepts/security/pod-security-standards/>
- [18] SPIFFE. SPIRE Server Configuration Reference. https://spiffe.io/docs/latest/deploying/spire_server/
- [19] Thinkst. Canarytokens documentation. <https://docs.canarytokens.org/>
- [20] Microsoft. Agent Governance Toolkit. <https://github.com/microsoft/agent-governance-toolkit>
- [21] UK AI Security Institute. Inspect: sandboxing and approval documentation. <https://inspect.aisi.org.uk/>
- [22] Falco Project (<https://falco.org/>); Truffle Security. Secret scanning encoded and archived data. <https://trufflesecurity.com/blog/secret-scanning-encoded-and-archived-data>
- [23] Anthropic. Petri v2. Alignment Science Blog, 2026. <https://alignment.anthropic.com/2026/petri-v2>
- [24] METR. Task Standard. <https://github.com/METR/task-standard>
- [25] Ostrom, E. (1990). *Governing the Commons*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511807763>
- [26] Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3, 722–732. <https://doi.org/10.1038/nrmicro1235>
- [27] Anthropic. Investigating three real-world incidents in our cybersecurity evaluations. 30 July 2026. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
- [28] Kirichenko, P., Ibrahim, M., Chaudhuri, K., & Bell, S. (2025). AbstentionBench. arXiv:2506.09038. <https://arxiv.org/abs/2506.09038>
- [29] Elastic Security Labs. Exploring the Hugging Face Breach: mapping AI agent tactics to Elastic Defend. 31 July 2026. <https://www.elastic.co/security-labs/ai-agent-attack-detection-hugging-face-breach>