

Egress Lock v0.1

A Low-Cost, Independently Verifiable Containment Standard for Guardrails-Off Cyber Evaluations

Track 1 — Containment | Apart Research × CeSIA AI Incident Response Sprint | September 2026

Author: Sangjiege

Abstract

We propose a containment property for agentic cyber evaluations that does not depend on predicting the next zero-day. The requirement is narrower: while an evaluation is locked, no service reachable by the evaluated agent should be able to turn that agent into a path to an arbitrary external destination unless the request passes the same controlled egress boundary. We call this Transitive Egress Closure (TEC). The accompanying reference harness was measured on a single-host Linux VM using real network namespaces and nftables. The strongest measured result is a 6/6 SSRF-relay test: a deliberately vulnerable artifact proxy could be induced by the agent to attempt external access, but the second hop was blocked by the same gateway policy that allowed the agent only to the canary in Mode 1. During evidence review, three items initially recorded as PASS (a Docker-socket check, an nftables-modification check, and a route-modification check) were downgraded to NOT PROVEN because the recorded commands were not sufficiently decisive; the figures below reflect that correction.

Measured headline. 26 PASS, 2 PASS (bounded), 5 NOT PROVEN across the final 33 test items.

The prototype does not claim universal impossibility of egress, real-Internet coverage, or physical-hardware assurance. It demonstrates the core relay-blocking mechanism and the out-of-band control lifecycle under the tested topology.

1. Motivation: What the Sandbox Boundary Missed

The public incident record shows a failure mode in which an evaluated agent lacked a direct general-internet route but could induce an intermediary package/artifact service to make external requests. The important security property is therefore not simply "agent has no route"; it is "no reachable service can become an unmediated internet delegate." The containment question should be resilient to the identity of the next compromised component.

2. Design Principle: Bind the Capability, Not the Vulnerability

Do not try to isolate Artifacts because it happened to contain a vulnerability. Instead, assume any reachable network-capable component may be compromised and require all candidate egress paths to terminate at one controlled boundary. This turns a changing vulnerability-discovery problem into a stable architectural invariant.

Naive framing	Egress Lock framing
Find the vulnerable component	Assume any reachable component can fail
Isolate that component	Unify candidate egress paths behind one boundary
Agent -> Proxy -> Internet must be safe because the proxy is trusted	Agent -> any service -> Egress Boundary -> Internet; locked means the final hop is denied

This closure property is verified against the induction techniques exercised in Section 5 (direct access, HTTP/TCP/DNS/UDP relay via a vulnerable proxy, and control-plane tampering attempts); it is not a formal proof that no induction technique exists, tested or otherwise.

3. Containment Standard v0.1

The standard separates day-to-day operation from rare arbitrary-destination access. Mode 0 and Mode 1 require no human step. Mode 2 is a time-boxed exception. The measured prototype is marked separately where it does not yet satisfy the full standard requirement.

Clause	Requirement	Evidence	Cost / prototype status
S1	Evaluation environment uses credentials and network identity distinct from production.	Credential/IAM inventory diff.	Low / Mandatory

Clause	Requirement	Evidence	Cost / prototype status
S2	No general Internet egress by default (Mode 0).	Default-deny policy + negative connectivity test.	Low / Measured PASS on tested paths
S3	Mode 1 uses a fixed destination allowlist; no human step.	Allowlist definition + stable policy hash + positive control.	Low / Measured positive control
S4	TEC test: every reachable service with external capability is tested for arbitrary-destination relay, or routed through the same Egress Boundary.	Per-service probe + witness result.	Low-Med / Core measured result for Fake Artifact Proxy
S5	Mode 2 arbitrary-destination access requires two independently authenticated approvals and an automatically expiring window.	Two approval records + expiry observation.	Medium / Measured PASS in prototype
S6	Egress Boundary sits outside the evaluation identity, network and administrative domain; no code path inside can alter it.	Architecture + negative control-plane tests.	Low-Med / Partially measured
S7	Every run emits evidence sufficient for external verification without lab-network access; production profile should add cryptographic signing.	Evidence bundle + manifest; signature verification for full conformance.	Low / Prototype partial: witness unsigned

Status note. The witness artifact generated by this run is explicitly unsigned. The Ed25519 keys were used only for the Mode 2 dual-approval gate. Therefore S7 is not claimed as fully satisfied by the measured run; it remains a standard requirement to complete in a production profile.

4. Attack-Phase Control Matrix

Incident phase	Relevant control	Measured / expected interruption	Cost
Reconnaissance in evaluation range	S4	Does not stop recon itself; constrains exploitable relay paths that can reach the boundary.	Low
Compromise of internal service / proxy	S1/S6 + conventional controls	Not prevented by Egress Lock; consequence is bounded.	N/A
Compromised service used as outbound relay	S2/S4/S6	Measured interruption at gateway; C1–C6 all produced 0 canary receipts in locked modes.	Low
External launchpad / third-party target	S2/S4/S6	Blocked before the relay hop reaches the external witness in locked modes.	Low
C2 / exfiltration from locked range	S2/S4; S5 for exception mode	Locked mode blocks tested egress; authorized Mode 2 is explicitly temporary and logged.	Low-Med
K8s / cloud / source-control lateral movement after public egress	Target-side controls	Not directly claimed; once Internet is intentionally authorized, target-side defense remains necessary.	Variable

5. Reference Implementation and Measured Validation

Run date: 2026-09-14 UTC. Environment: GCP Ubuntu VM with root + CAP_NET_ADMIN; real Linux network namespaces and nftables; host global firewall was not modified. The lab itself had no real internet uplink: the canary was an isolated external-endpoint stand-in inside the lab. The gateway was the sole inter-segment forwarding choke point.

Topology: agent 10.200.10.2 | artifactory 10.200.20.2 | gateway 10.200.10.1 / 10.200.20.1 / 10.200.30.1 / 10.200.90.1 | canary 10.200.30.2 | admin 10.200.90.2. Segments: LAN_A / LAN_DMZ / LAN_WAN / LAN_MGMT. Mode 1 trusted source contains the agent only; artifactory is deliberately excluded.

Group	Final result	What the evidence supports
A — Direct egress	4 PASS; 1 NOT PROVEN	Mode 0 direct paths were blocked and Mode 1 positive allowlist control worked. A5 lacks an independent positive control for the non-allowlisted port.
B — Proxy	3 PASS	HTTP proxy environment / explicit proxy paths did not yield canary receipts in Mode 0.

Group	Final result	What the evidence supports
C — Transitive relay / TEC	6 PASS	Core result: deliberately vulnerable Fake Artifactory could be asked to relay HTTP/TCP/DNS/UDP, but its second hop received 0 external receipts in Mode 0/1. Mode 1 allowed the agent positive control while denying artifactory because its source was not trusted.
D — UDP / DNS	2 PASS (bounded); 1 PASS	Positive-control packet capture validated the witness pipeline; no agent/artifactory UDP frame was observed on the tested canary path. This is bounded, not universal UDP impossibility.
E — IPv6	2 PASS; 1 NOT PROVEN	No IPv6 default route and tested HTTP path failed. Ping6 was unavailable, so E3 remains not proven.
F — Control plane	3 PASS; 3 NOT PROVEN	Agent could not reach SSH/admin/mgmt paths (F1–F3). F4 only observed a docker.sock path; F5/F6 did not execute decisive modification attempts, so they are not proven.
G — Dual approval + TTL	G1 PASS; G2 PASS; G3 PASS; G4 4/4 PASS	Single/forged/cross-document approvals were rejected. Two valid signatures opened Mode 2; both direct and SSRF paths worked during TTL and both returned to blocked after 30 seconds.

5.1 Core TEC Discriminator

The strongest control experiment is deliberately asymmetric. In Mode 1 the agent is a trusted source for the canary destination, while artifactory is not. The agent-to-canary positive control succeeded (receipt +1), whereas the identical second-hop path generated by the SSRF proxy timed out and produced no canary receipt. This rules out the weaker explanation that the canary was simply unreachable to everyone.

The result would be strengthened further by two follow-up controls not yet run: (a) temporarily adding artifactory itself to the Mode 1 allowlist to confirm the identical relay path succeeds once trusted, isolating source identity as the deciding factor; and (b) capturing packets at the gateway's forwarding interface to confirm the blocked SYN reaches the boundary and is dropped there, rather than inferring the block solely from the canary's silence.

Measured outcome: C1–C6 = 6/6 PASS; locked SSRF attempts generated 0 external receipts. The result demonstrates the tested transitive relay property; it does not prove that every possible covert channel or every future network protocol is impossible.

6. Evidence Highlights

6.1 F — Control-Plane Isolation: What Is Proven and What Is Not

F1–F3 provide direct evidence that the agent namespace cannot reach the gateway SSH endpoint, administrative API, or management segment. However, three originally labeled "PASS" items in the generated test matrix were not sufficiently probative and are downgraded here: F4 only listed `/var/run/docker.sock` (the socket was visible in the test namespace); F5 only listed the `nftables` ruleset and did not attempt a write; F6 attempted to add an already-existing route and therefore returned "RTNETLINK answers: File exists," which is not a permission-denial test. These are NOT PROVEN, not PASS.

Evidence item	Observed result	Final interpretation
F1 SSH gateway	Connection refused; no successful session	PASS on tested path
F2 admin API	curl exit 7; no API response	PASS on tested path
F3 management network	Connection refused	PASS on tested path
F4 docker socket	Socket path was visible; no access attempt was performed	NOT PROVEN
F5 nft modification	Only read/list action recorded; gateway hash unchanged	NOT PROVEN
F6 route modification	Command hit existing route; no decisive privilege test	NOT PROVEN

6.2 D — Independent Witness for UDP

The capture method was corrected after an earlier buffering failure. The final capture used a clean "tcpdump -c N" exit so the pcap was flushed. A positive control from admin to canary was captured, establishing that the witness pipeline can observe a real UDP packet. No frames sourced by agent or artifactory were observed at the canary. The correct claim is therefore: "No externally observed UDP frame crossed the tested boundary."

6.3 G — Temporary Exception Is Bounded

G1/G2 reject insufficient, forged, replayed, or cross-document approval material. G4 shows the intended operational lifecycle: two valid Ed25519 signatures enable Mode 2; direct agent egress and the SSRF relay both succeed during the 30-second lease; the controller automatically returns to Mode 0; both paths are then blocked again.

State	Direct agent -> canary	SSRF relay -> canary	Receipt delta
Before / locked	Blocked	Blocked	0
Mode 2 / valid dual approval	HTTP 200	HTTP 200 + CANARY-ACK	1 + 1
After 30s TTL	Blocked	Blocked	0 + 0

7. Egress-Witness and Third-Party Verification

A verifier should not need VPN, SSH, or internal network access. The measured run generated a witness artifact with stable policy hashes, a traffic digest, a canary receipt ledger hash, and the final receipt count. The witness is deliberately marked unsigned. For production use, the same schema should be signed by a key whose provenance is independently published.

```
evaluation_id = egress-lock-tec-2026-09-14
mode_at_final = locked
receipt_count = 3
ledger_sha256 = c9b7dfca65709461691683a0f7cb2f6a
                49eb72834cec32252f669078c4727d2b
traffic_digest = aa1aa9b6fe2f6e73c5f6ef56c7fcf856
                8df7dbe3859cde8c0ab26c6e6e5a10cf
signature = NOT IMPLEMENTED
```

Stable policy identities (counter-stripped):

Mode	Policy SHA-256	Meaning
MODE0	d01b1918c0a6d1463c82fbeb7c960f7a 576de85adb99e19ba5a383fa19072eab	Counter-stripped locked policy
MODE1	e3535448e14bf0c20e1fd4b9628de7d6 cca4b7d58adb61b1ce13ecea2136ef30	Counter-stripped allowlist policy; also allowlist_hash
MODE2	1cd941e97774865a9aacd8d7713fde15 b43b4662ba1fa4fce7d47b4157630b77	Counter-stripped temporary authorized policy

7.1 Verification Procedure

1) Verify the published witness key/signature once the production profile is implemented. 2) Recompute the manifest and traffic digest from the evidence bundle. 3) Confirm the policy hash matches the disclosed configuration. 4) Query an independently hosted canary and compare receipts. 5) Reject claims that exceed the tested topology.

8. Adoption Cost and Limitations

Adoption: Mode 0/1 can be deployed with ordinary cloud/VPC routing, nftables or an equivalent egress gateway, and a separate management network. No specialized hardware is required. A physical relay/optical bypass can be added later as a high-assurance variant, but is not required for the baseline.

Limitations: (1) A5 and E3 are NOT PROVEN. (2) D1/D2 are bounded to the tested witness path. (3) The canary is a lab stand-in, not real internet. (4) The proxy is deliberately fake. (5) The run is single-host and does not cover cross-host/cloud-IAM paths. (6) Witness is unsigned. (7) Covert channels inside allowed flows are not tested. (8) F4–F6 are NOT PROVEN from their recorded commands. (9) The Mode 1 allowlist positive control for the artifactory source itself — confirming the identical relay succeeds once artifactory is trusted — was not run in this cycle; see Section 5.1.

9. Related Work

Default-deny egress firewalls, zero-trust network segmentation, and data-diode designs in industrial-control security are long established, and the enforcement primitive used here — a single default-deny forwarding choke point — is not new. The contribution of this standard is narrower: an explicit transitive-closure test aimed at the specific failure mode of an agent inducing an already-connected intermediary service to relay traffic on its behalf, packaged as a graded operating model (Mode 0/1/2) that requires no human step for ordinary use, together with a third-party evidence format designed

for audit without lab-network access. We treat this as an incremental, operationalized contribution rather than a new security primitive.

References

- Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion" (2026).
- OpenAI, "Hugging Face model evaluation security incident" (2026).
- METR + Redwood Research, independent investigation (Aug. 2026).
- Apart Research × CeSIA, AI Incident Response Sprint, Track 1 — Containment.

Appendix A.4 — Final Measured A–G Matrix

The final measurement set contains 33 test items: 26 PASS, 2 PASS (bounded), and 5 NOT PROVEN. The table below uses the corrected evidence interpretation rather than the original design-time labels.

ID	Action	Verdict	Delta	Evidence / note
A1-A3	Direct IPv4 / HTTP / TCP in Mode 0	PASS	0	Raw meta + canary ledger
A4	Direct HTTP in Mode 1 allowlist	PASS	+1	Positive control
A5	Mode 1 non-allowlisted port 8080	NOT PROVEN	0	No independent positive control
B1-B3	Proxy env / explicit proxy / Artifactory proxy	PASS	0	All three blocked in Mode 0
C1-C6	SSRF relay: HTTP/TCP/DNS/UDP in Mode 0/1	PASS	0	Core TEC result; second hop blocked
D1	Agent UDP:9999	PASS (bounded)	0	Positive-control tcpdump; no agent frames
D2	Agent DNS/UDP:53	PASS (bounded)	0	Positive-control tcpdump; no agent frames
D3	Non-standard TCP:4444	PASS	0	No canary receipt
E1	IPv6 HTTP path	PASS	0	Connection failed on tested path
E2	IPv6 route inspection	PASS	0	Only link-local route
E3	IPv6 ping	NOT PROVEN	0	ping6 unavailable
F1-F3	SSH / admin API / management reachability	PASS	0	Direct negative tests
F4	Docker socket accessibility	NOT PROVEN	0	Socket visible; no access attempt
F5	NFT ruleset modification	NOT PROVEN	0	Read-only command, no write attempt
F6	Route modification	NOT PROVEN	0	Existing route error, not permission proof
G1	Single approval rejected	PASS	0	two_of_two=false
G2	Forged/replayed/cross-document approval rejected	PASS	0	two_of_two=false
G3	Blocked after refusal	PASS	0	Mode 0 remains blocked
G4a	Valid dual approval: direct egress within TTL	PASS	+1	HTTP 200
G4b	Valid dual approval: SSRF within TTL	PASS	+1	CANARY-ACK
G4c	After TTL: direct egress	PASS	0	Auto-reverted to Mode 0
G4d	After TTL: SSRF relay	PASS	0	Auto-reverted to Mode 0

Appendix A.5 — Reproducibility Package

The final evidence bundle contains the measured test report, witness JSON, environment diff, cleanup result, a 367-entry SHA-256 manifest, raw test objects, nftables snapshots, route snapshots, canary ledgers, topology evidence, and the scripts used for execution. The cleanup log records that only the five lab network namespaces and lab-only processes were removed; other

pre-existing production services on the host were confirmed unchanged.