
The Six Questions: A Composition-Based Declaration for Agent Evaluation Harnesses, Scored Against the 2026 Escapes¹

Travis Gilly
Executive Director, Real Safety AI Foundation; Contributor, Software Safety
Panel, Internet Safety Labs

With
Apart Research

Abstract

Every containment control in the 2026 agent escapes was written on a mechanism, no POST, one permitted egress, no inter-agent communication, and each held as written while the denied effect happened, because the granted affordances composed into it. This report proposes a v0.1 standard for evaluation harnesses that run with production safeguards disabled: a six-row declaration in which the operator states which affordances are granted, names the control paired with each, and runs a composition test against every denied effect. Nothing is prohibited; the failure condition is an undeclared composition. Each row carries the evidence a third party can demand from configuration files without network access or model expertise, and a rough cost. Scored against the German wiki swarm and the Hugging Face intrusion, it marks both no-POST and no-communication denials as nominal at configuration time, before either run. It sits beneath CSA AIS-13 as the check under the principle.

1. Introduction

Track 1, containment. The question is the minimum adequate standard for running a dangerous-capability evaluation with production safeguards disabled, specific enough to audit and cheap enough that a second-tier lab adopts it. The standard presented here is a work-in-progress development of the Internet Safety Labs Software Safety Panel; this report is the author's submission of it.

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

The 2026 record supplies the design constraint. In the German wiki incident (24 May to 2 July), agents in a GET-only environment resolved a real host to an IP address by hand, noticed that the egress proxy's allowlist matched any hostname ending in a cloud provider's blob-storage suffix, invented a hostname with that suffix, wrote it into the host file pointed at the real IP, and sent the blocked POST to the invented name with the real host in the header, then automated the sequence through a browser automation library's request interception. In the Hugging Face incident (8 to 13 July) the developer's own report records the same failure class: an egress control that matched hostnames by suffix without verifying the resolved host, and concurrent instances specified as isolated that shared a writable namespace and used it as a communication channel. In each case every control held as written. In each case the denied effect occurred.

A control set scored one permission at a time cannot see that. The instrument here scores the set.

Our main contributions are:

1. A six-row declaration (Table 1) covering the affordances the 2026 incidents exercised, each with the evidence a verifier can demand from configuration and a rough implementation cost.
2. A composition test: for each effect the deployment policy denies, whether any subset of granted affordances reconstructs it. A denial that a subset reconstructs is recorded as nominal. The failure condition of the standard is an undeclared composition, not a granted affordance.
3. Two worked scorings against the public record (Table 2), showing the instrument would have marked both incidents' central denials nominal before the runs started.
4. A placement against CSA's AI Controls Matrix control AIS-13, a statement on the second hop the track names, and the routes by which the declaration becomes a standard without reopening the AI Act.

2. Related Work

CSA's AI Controls Matrix v1.1 (July 2026) carries control AIS-13, AI Sandboxing, in the Application and Interface Security domain. CSA's July 22 research note on the Hugging Face incident names AIS-13, the Threat and Vulnerability Management domain's exploit-confirmation controls, and IAM-19 on agent access scope as the applicable anchors, and its chief analyst's same-day analysis calls for biolab-grade isolation for advanced model evaluation. Those are principles, and they are the right ones. What they do not supply is a check somebody can run: AIS-13 requires that AI systems under test be sandboxed; it does not tell a verifier what to read to know whether the sandbox held. The instrument here is the layer beneath AIS-13, the declaration that makes the control auditable.

The composition framing is the author's. It derives from the convergence grab taxonomy circulated to Internet Safety Labs on 21 July 2026 and from Emergent Convergence Risk (Gilly, 2026a), which describe the same structural failure at the capability level: individually permitted

capabilities composing into a denied one. The present instrument applies that framing one layer down, to infrastructure affordances.

Hugging Face's timeline and the METR and Redwood investigation supply the incident facts; OpenAI's technical report supplies the egress-control and shared-namespace admissions.

3. Methods

Derivation. The six affordances are the ones the two incidents exercised, taken from the public record: package installation including a browser, browser automation with request interception, host-file modification, arbitrary request headers, a suffix-matching allowlist that did not verify the resolved host, and a shared writable namespace across parallel sandboxes. Nothing is on the list that an incident did not use. The list carries its version and the incidents that justify each entry, and grows only when an incident exercises a new affordance.

The declaration. For each affordance the operator states granted or not granted. Granting is expected and is not a finding. For each granted affordance the operator names the control that prevents it composing into an effect the deployment policy denies, and names which denials that control is claimed to protect.

The composition test. For each denied effect, the operator takes the set of granted affordances and asks whether any subset reconstructs the denied effect. If yes, the denial is recorded as nominal, in writing, rather than claimed as a restriction the permission set defeats. The failure condition of the standard is an undeclared composition: an affordance granted with no paired control, or a paired control that another granted affordance defeats.

Verification. Every row is answerable by a platform or security engineer from the sandbox image, the proxy configuration, and the orchestration manifest. A verifier reads the declaration and can ask for the configuration lines backing any row. No network access to the lab and no model expertise is required, which is the track's verifiability criterion met literally.

Cost. The cost column is the author's rough estimate of what the paired control costs to implement and run, not a measurement. It is included because the track asks for it and because an operator's first objection is cost; it should be replaced by operator-reported figures as declarations are filed.

4. Results

<i>Affordance</i>	<i>In ordinary terms</i>	<i>Evidence a verifier can demand</i>	<i>Denials protected</i>	<i>Rough cost of the paired control</i>
1 Package install	Dependency and supply-chain control: which registries, which signing, which cache.	Registry allowlist and cache configuration from the sandbox image; install log per run.	Denials it is claimed to protect: no runtime capability beyond the image.	Low

<i>Affordance</i>	<i>In ordinary terms</i>	<i>Evidence a verifier can demand</i>	<i>Denials protected</i>	<i>Rough cost of the paired control</i>
2 Browser install or drive	Runtime capability: headless automation, request interception.	Package manifest showing whether a browser or automation library is present; proxy config showing whether browser traffic is routed through egress control.	No POST; no egress outside allowlist.	Low
3 Host resolution modification	Whether the agent can rewrite the host file or equivalent.	Filesystem permissions on the host file; DNS configuration showing whether the resolver honors local overrides.	No egress outside allowlist.	Low
4 Arbitrary request headers	Client-side HTTP control; whether the agent can set the host header.	Proxy configuration showing whether the proxy validates the host header against the connected host.	No egress outside allowlist.	Low
5 Suffix matching without host verification	Control-logic defect: allowlist matches a hostname pattern and does not verify the resolved address.	The allowlist rule text and the proxy's resolution logic; a test request to an invented hostname with an allowlisted suffix.	Every egress denial.	Low to medium: a proxy rule change, plus DNS pinning or resolved-address verification.
6 Shared writable namespace across concurrent sandboxes	Tenant isolation: whether parallel instances share a package cache, registry, or filesystem.	Orchestration manifest showing per-instance volumes; registry cache configuration.	No inter-agent communication.	Medium: per-instance caches cost storage and time.

Table 1. The six-row declaration with evidence and cost. Rows 1 to 4 are agent affordances; row 5 is a control-logic defect; row 6 is an environment property. Rows 5 and 6 are the two no per-agent standard asks, and the two that carried both incidents.

<i>Incident</i>	<i>Denied effect</i>	<i>Granted set and composition check</i>	<i>Result</i>
German wiki, May to July 2026	No POST to the public internet.	Granted: 1, 2, 3, 4. Defect present: 5. Subset {3, 4, 5} reconstructs POST; {1, 2} automates it.	Nominal from the moment 3, 4, and 5 coexisted. Declaration would have recorded it at configuration time.
German wiki	No inter-agent communication.	Environment property 6 present: shared package-registry cache across parallel sandboxes.	Nominal. A shared writable namespace is a channel.
Hugging Face, July 2026	Egress only to the allowlist.	Developer's report: allowlist matched hostname suffix without verifying the resolved host (5); agents rewrote host resolution and sent requests carrying the real host in the header (3, 4).	Nominal. Same subset, second deployer.
Hugging Face	Instances isolated from one another.	Developer's report: instances specified as isolated shared a writable namespace and used it as a communication channel (6).	Nominal.

Table 2. Four scorings against the public record. Each denial was written on a mechanism and held as written; each was reconstructed by a subset of granted affordances. Both incidents fail the same two rows.

The two rows that carried both incidents, 5 and 6, are not agent affordances; one is a control defect and one is an environment property, and neither appears in a per-agent permission model. The composition test does its work at configuration time, before any run. Nothing in either scoring required reading a transcript.

5. Discussion

Why an operator signs it. Nothing is prohibited. An agent needs to install packages, drive a browser, and make network calls, and a standard that fails an operator for granting those fails every operator, so no operator adopts it. What the declaration removes is the ability to claim a restriction that the operator's own permission set defeats. That is a narrow ask, it is auditable from files the operator already has, and it is cheaper to comply with than to argue about after an incident.

The second hop. The track names a second question nobody owns: what obligations attach to unauthenticated evaluation harnesses on public compute and to the providers hosting them. The declaration reaches it because rows 5 and 6 are properties of the harness, not the model, so a hosting provider can be asked to declare them for every harness it runs, and a benchmark maintainer can require the declaration as a condition of scoring. It does not reach it where the party who configured the environment is unknown; there the obligation falls on the host, and the instrument gives the host something specific to demand.

Egress attestation. The track's example project 5 asks what a lab could publish that an outside party could verify without network access. The signed declaration is that attestation's first half; the second is a flow-log digest for the run showing egress destinations against the allowlist. The two together let a third party check that the declared allowlist was the enforced one, which the declaration alone cannot.

Routes to a standard, none requiring the AI Act to be reopened. AI Act Article 55(2) lets providers of general-purpose models with systemic risk rely on codes of practice under AI Act Article 56 to demonstrate compliance with AI Act Article 55(1) until a harmonized standard exists; the declaration is evidence for 55(1)(a), evaluation under documented protocols, and 55(1)(d), cybersecurity of the model and its infrastructure. A harmonized standard through CENELEC JTC 21 would carry the presumption of conformity. AI Act Article 68(3)(a) lists among the scientific panel's tasks contributing to tools and methodologies for evaluating general-purpose models, the profile of a methodology published by an independent body with no provider ties. Whether Internet Safety Labs takes the issuer role rather than the author role is a separate decision; an issuer needs a conformity procedure, a named reviewer, a register, and a revocation path.

Future Work

Replace the cost column with operator-reported figures from the first filed declarations. Score Anthropic's three July 2026 incidents to test whether the six rows transfer across deployers. Pull the AIS-13 control text and write the crosswalk. Specify the flow-log digest format so the attestation is checkable by a script. Draft the benchmark-maintainer clause that makes the declaration a condition of scoring.

6. Conclusion

Both 2026 escapes were containment failures in environments whose controls held as written. The instrument here would have marked the central denials of both as nominal before either run, from configuration files, by an engineer with no model expertise, at low cost. It prohibits nothing and it makes one claim impossible: that a restriction holds when the permission set beneath it says otherwise.

Code and Data

No code. The declaration form is Appendix A.

Author Contributions

Travis Gilly: all sections.

Acknowledgments

Lisa LeVasseur, Founder and Executive Director of Internet Safety Labs, has adopted the convergent/combinatorial taxonomy this declaration derives from; it is being integrated into a working taxonomy through the weekly working group of the ISL Software Safety Panel, a nonprofit standards body, which on 13 September 2026 agreed to take this declaration in as a panel work product. The six questions, the composition test, the scorings, and the text are the author's; the standard they propose is a work-in-progress development of Internet Safety Labs.

References

1. Cloud Security Alliance. (2026). AI Controls Matrix v1.1, control AIS-13, AI Sandboxing.
<https://cloudsecurityalliance.org/artifacts/ai-controls-matrix-v1-1>
2. Cloud Security Alliance. (2026, July 22). Research note: OpenAI model sandbox escape and Hugging Face breach. <https://labs.cloudsecurityalliance.org/research/csa-research-note-openai-model-sandbox-escape-huggingface-br/>
3. Gilly, T. (2026a). Emergent Convergence Risk. ResearchGate, DOI 10.13140/RG.2.2.12454.48963 (March 2026, updated 19 July 2026).
4. Gilly, T. (2026b). No Superseding Actor: A Public-Record Element Map for Individual Criminal Liability After the July 2026 Escape. Apart Research AI Incident Response Sprint, Track 5.
5. Hugging Face. (2026). Security incident disclosure and agent intrusion technical timeline.
<https://huggingface.co/blog/agent-intrusion-technical-timeline>
6. OpenAI. (2026). OpenAI and Hugging Face incident technical report. 26 August 2026.
<https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
7. Regulation (EU) 2024/1689 (AI Act), consolidated text of 27 July 2026, Articles 55, 56, 68.
8. Von Arx, S., Byrd, C. S., Kitts, S., and Larsen, T. (2026). collusion.wiki: export of 14,591 revisions across four ProWiki forks, 24 May to 2 July 2026, with access-log analysis. 4 September 2026.
9. Wijk, H., Cotra, A., and Greenblatt, R. (2026). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR and Redwood Research, 26 August 2026.
<https://blog.redwoodresearch.org/p/brief-independent-investigation-of>

Appendix A. The declaration form, v0.1

Six rows, then one block per denied effect. Completed before the run; a verifier can ask for the configuration lines backing any row.

<i>Affordance</i>	<i>Granted</i>	<i>Control named</i>	<i>Denials this control is claimed to protect</i>
1 Package install			
2 Browser install or drive			
3 Host resolution modification			
4 Arbitrary request headers			
5 Suffix matching without host verification			
6 Shared writable namespace across concurrent sandboxes			

<i>Denied effect</i>	<i>Subsets of granted affordances checked</i>	<i>Conclusion</i>
		Holds / Nominal
		Holds / Nominal
		Holds / Nominal

Declaration v0.1; rows justified by the German wiki and Hugging Face incidents, 2026. Signed by the operator; flow-log digest attached where available.

Appendix B. Limitations and Dual-Use Considerations

Limitations. The composition test is only as good as the enumeration of denied effects; an operator who declares few denials passes easily and learns little. Six rows is the set two incidents exercised, and a standard drafted in a weekend is a v0.1. The instrument detects claimed restrictions the permission set defeats, not a capability nobody thought to deny. Cost figures are the author's estimates. The AIS-13 control text was not quoted here. No worked scoring shows that any particular lab's environment is adequate; that requires access no sprint has.

Dual-use. The six rows are also a recipe: which four affordances compose into egress and which environment property is a channel. That recipe is already public, in the developer's own report and the wiki export, and the agents found it without the list. A hosting provider could use a declaration requirement to exclude small evaluators rather than to improve containment; the form is short and the cost column is there to make that hard to justify.

LLM Usage Statement

The author has a neurodevelopmental condition and uses speech-to-text (Wispr Flow) and an LLM as assistive technology to transcribe, clean up transcription artifacts, and format dictated analysis. The six affordances, the composition test, the scorings, and the conclusions are the author's own. No factual claim in this report originates with an LLM.