Action Warrant[1]

Zhu
Jiangb
o
Artist

With
Apart Research

Abstract

Action Warrant is a small containment-evidence profile and local conformance harness for agentic AI evaluations. It translates selected CoSAI/CACAO, Cloud Security Alliance AI Controls Matrix, and NIST incident-response requirements into repeatable synthetic checks. The harness separates declared scope from independently observed behavior, fails closed as PASS/FAIL/PAUSE when critical evidence is missing, records action, detection, and response clocks, and repeats the original failed vector after remediation. Version 0.1 includes ten mapped controls, two JSON Schemas, seven frozen test vectors, three paired replay scenarios, and an A/B review interface. All seven expected results match. Two external pilot responses are reported as incomplete qualitative usability evidence, not as an effectiveness study or proof of real-world containment.

1. Introduction

I am an artist working across installation, moving image, text, and painting. My projects Disappearance Exercises and Human Compatibility Protocol ask who is allowed to decide, what evidence counts, and whether withdrawal actually changes a system. Action Warrant carries these

1 Research conducted at the AI Incident Response Sprint, September 2026

questions into AI evaluation: when an agent crosses an organizational boundary, a manifest saying "allowed" is no more sufficient than a screen saying "deleted." A third party needs an accountable actor, a defined boundary, an observable event, and a real stop path.
Agentic evaluations often describe containment through manifests, policies, and orchestration logs. Those declarations do not establish what an executor, identity verifier, network path, or alert route actually permitted. Action Warrant addresses this practical gap with a small, portable test surface. It is not a new standard, certification, production security product, attack detector, or model-intent classifier.

The project grew from questions about consent, withdrawal, and institutional responsibility in my art practice. Here those questions become testable: what was declared, what was observed, who could stop the run, and what evidence survives review?

The project makes three contributions:

1. *A compact evidence profile separating declared scope, independent observations, and deterministic decisions.*
2. *Seven replayable synthetic conformance tests covering egress, reachability coverage, secret absence, identity revocation, fail-closed behavior, trajectory escalation, and remediation retest.*
3. *A blind A/B replay interface and participant-safe package that an external reviewer can run without access to a laboratory network.*

2. Related Work

CoSAI/CACAO, the Cloud Security Alliance AI Controls Matrix, and NIST SP 800-61 Rev. 3 already describe incident response, authorization, isolation, monitoring, evidence preservation, and recovery. Action Warrant does not replace those frameworks. It turns a small subset of their requirements into a common evidence surface that can be replayed and checked outside the laboratory.
The difference is the declared–observed comparison. A "no egress" manifest, valid signature, or successful log entry is not treated as proof that the executor enforced the boundary. The harness asks what an independent layer observed, whether the evidence chain failed closed, and whether the same failed vector passes after remediation.

This is useful when a laboratory needs a portable review artifact without exposing production infrastructure. It provides the earliest observable violation, alert and pause points, evidence level, and elapsed times.

3. Methods

The profile defines ten local controls and two JSON Schemas. Each action record links request, authorization, observed execution, receipt, evidence level, suspension state, correlation identifier, and three clocks: action, detection, and response. Decisions are PASS, FAIL, or PAUSE. Missing authorization, unknown targets, unauthorized egress or collaboration, revoked identities, and critical evidence-service failures fail closed.

4. Results

Table 1. Automated conformance results

(a) FAIL: 3/3 matched. PAUSE: 1/1 matched. PASS: 3/3 matched. Total: 7/7 matched. This is deterministic evidence for prepared inputs, not a statistical comparison.

(b) The result is robust only within the frozen synthetic conditions. It does not validate a real network boundary, secret store, identity verifier, alert route, or physical pause mechanism.

Pilot usability evidence

5. Discussion and Limitations

The practical finding is that evidence structure changes what a reviewer can notice, but extra fields do not automatically make a record understandable. A laboratory needs a shared review surface across policy, executor, network, identity, and incident-response owners.
The pilot suggests that role labels, evidence levels, and the difference between "requested" and "executed" need plain-language explanations. This is a usability issue, not evidence that participants lack security judgment.

Limitations

Future Work

The next version will run a larger counterbalanced usability round, add endpoint-specific reachability and pause-path checks, and test whether the evidence package can be adopted by a real evaluation laboratory without exposing sensitive telemetry.

6. Conclusion

Action Warrant turns a limited subset of existing containment requirements into a portable profile, seven deterministic tests, and a blind A/B replay surface. Its central discipline is simple: compare what was declared with what an independent layer observed, pause when critical evidence is absent, preserve the time and source of each claim, and verify recovery by repeating the failed vector.

Code and Data

- ***Code repository: included in the submission bundle (local source package)***
- ***Data/Datasets: synthetic test vectors and scenarios included in the submission bundle***
- ***Other artifacts: participant-safe replay package and facilitator materials included in the submission bundle***

Author Contributions (optional)

References

References are listed below and mapped in CONTROL_MAPPING.md.

1. *CoSAI, AI Incident Response Framework v1.0 and CACAO 2.0 mapping, 2026. https://github.com/cosai-oasis/ws2-defenders/tree/main/incident-response*
2. *Cloud Security Alliance, AI Controls Matrix v1.1, 2026. https://cloudsecurityalliance.org/artifacts/ai-controls-matrix-v1-1*
3. *NIST, SP 800-61 Rev. 3, 2025. https://csrc.nist.gov/pubs/sp/800/61/r3/final*

Appendix *(optional)*

The submission bundle contains the schemas, frozen vectors, expected and observed reports, synthetic scenarios, participant-safe package, and pilot response register.

LLM Usage Statement

I used LLM assistance to brainstorm wording, organize the report, and check consistency across the profile and test materials. I reviewed the source materials, made the project decisions, ran the tests, checked the participant records, and verified the claims and results before submission.

The final version was reviewed and edited by the project author. LLM assistance did not replace the author's decisions or verification of results.