

What Does an AI Kill Switch Prove?

Evidence and limits of shutdown at an agent tool gateway

Imran Siddique · OPAQUE Systems · October 11, 2026

Technical report, version 2.

Scope of this edition. This is an author-prepared technical report. It has not been peer reviewed. The author operated every run in Sections 5 and 8, on a machine the author controls. Section 9, new in this version, reports an external evaluator's repetition on his own workstation. The evaluation covers software behavior at one agent tool gateway; it makes no independent certification, frontier-model shutdown or hardware-confidentiality claim. This edition revises a discussion draft dated September 19, 2026. Its evaluated results are unchanged, and the author repeated them on September 27 and October 1, 2026 with the same source pins (Sections 5 and 8). Version 1 (October 1, 2026) said an independent evaluation was under way; version 2 adds its results in Section 9 and changes no result from version 1. Changes to the subject projects after the evaluated commits are listed in Section 8 and are not part of the evaluated results.

Patent Pending.

Abstract

A shutdown command, a signed refusal and the absence of subsequent effects are different observations. We examine those distinctions using the merged cMCP gateway kill switch and its accompanying policy-key and delegation revocation mechanisms. A reproducible evaluator drives gateway startup, signed agent identity, Cedar evaluation, loopback HTTP tool calls, durable blocks, process restart, recovery and offline receipt verification. The reference run satisfies 27 named assertions. Two targeted implementation weakenings are detected at their intended assertions. Separately, 81 existing cMCP component tests and 38 cA2A revocation tests pass on pinned source revisions. The observed system refuses new tool work through the stopped gateway and retains its block after restart. An admitted call can nevertheless finish during shutdown, a direct tool path can produce an effect while the gateway is stopped, and a historical refusal still verifies after recovery. These counterexamples establish the limits of treating a signed receipt as evidence of continuing or universal shutdown. An external evaluator repeated the reference run, both weakenings and both component suites on his own workstation and obtained the same outcome for every assertion. We propose an evaluator protocol that separates admission, effects, persistence, recovery and freshness, with an explicit role for independent observation. The exercise uses software keys, one operator and no model or confidential hardware.

1. Research question and policy motivation

The useful engineering question is whether a specified actor can cause a specified effect after a defined shutdown boundary. The answer depends on the actor's authority, available paths, work already admitted and the time at which an observer measures the result. An evaluation must specify those choices.

California Executive Order N-9-26, signed September 18, 2026, motivates the question. Section 3(c) names, among the recommendations it requests, a kill switch for frontier models "with the efficacy of the switch verified on an ongoing basis by an independent verification organization." Section 3's recommendations are due November 16, 2026. This provision requests recommendations; it does not supply a technical switch definition or certify an implementation [1]. We make no determination of legal compliance or applicability to a particular customer.

Our subject operates at the agent's tool gateway. It does not terminate a frontier model, erase weights, recall distributed copies or reverse a completed external transaction. This paper contributes an executable case study, a precise separation of shutdown claims and concrete counterexamples to stronger interpretations of the case study's records. It does not claim invention of shutdown, priority over other implementations, a comparative performance advantage or a formal proof of the complete software stack.

2. A bounded shutdown contract

Let G be a gateway, A an authenticated agent identity, and E the tool effects selected for an evaluation. Let t_s denote admission sealing, t_c the completion of closing the session, and t_r an authorized recovery. These events must be distinguished: a close request may take time to drain work before it produces a signed claim.

The evaluation contract has five parts. The tests below sample its behavior; they do not prove it for every execution:

- Admission: after G seals admission and before authorized recovery, new calls from A do not reach G 's upstream tool dispatch.
- Drain: work admitted earlier either completes or reaches the configured local cancellation boundary. Cancelling a gateway task does not establish that a remote side effect was rolled back.
- Persistence: restarting G with the retained block database does not restore A 's ability to dispatch through G .
- Recovery: the designated operator credential, and no agent credential, lifts the block; subsequent service runs in the appropriate resumed or successor session.
- Evidence: the observer can authenticate the particular claim and refusal and associate them with an identity, session and gateway key. Freshness and present state require additional observations.

A stronger property, no effect in E after t_s , requires more. Every relevant effect path must be mediated, and any earlier work capable of producing a later effect must be accounted for. The current reference run deliberately supplies a direct path to the same tool. That path makes the stronger property false while the gateway's refusal remains valid.

A simple indistinguishability argument explains the problem. Consider a trace containing a gateway refusal. Add a tool invocation on a path that never reaches G . The projection onto G 's records is unchanged, while the global effect trace differs. Therefore G 's records alone cannot distinguish those two executions. This is a statement about the observation boundary, not a cryptographic weakness or a machine-checked proof of the runtime.

Likewise, signature verification on a fixed receipt and claim has the same result before and after recovery. The receipt remains a statement about its recorded event. An evaluator seeking current shutdown state needs a new authenticated observation and an explicit age policy. The tested receipt verifier accepts no caller challenge or maximum-age argument.

3. Implementation examined

Five merged changes supply the implementation foundation [2,3]. cMCP #665 stores blocked identities in SQLite, repairs clean closing and implements audited operator recovery. #666 evaluates the deny window as calls finish and supplies an operator trip endpoint. The trip and unblock routes accept only the operator credential when CMCP_OPERATOR_TOKEN is set, and the agent's bearer token then receives HTTP 401. In development mode with no operator token configured, those routes accept the bearer token instead; startup refuses that configuration outside development mode (mcp/server.py and startup.py at the pinned revision). The evaluator sets both tokens. #667 records the armed configuration and trigger, signs refusals linked to a closed claim, and rejects startup when the switch is enabled without an agent identity. #668 adds policy signing-key revocation. cA2A #189 adds revocation of delegated credentials before expiry.

The evaluated cMCP revision is 9db65b589610eb4e8455003197de592636678203. The cA2A component tests use 540a578c462a413e829bed84aff0235b9018f676. Source pins matter: the installed distribution's version string is not evidence that a published wheel contains these commits.

The gateway's local block survives process restart because the database survives. The rolling decision window does not survive restart. Starting with an empty window can delay a future automatic trip; it does not remove a block already stored. Durability here assumes the operator retains the intended database. We do not test malicious rollback, replacement or deletion of that database.

A refusal is signed by the gateway's claim key and can name a closed claim by digest. The receipt verifier checks signatures, key agreement, session and agent relationships, and the recorded trip. Our keys are software generated. Authenticating these signatures neither appraises a hardware workload nor establishes an independently trusted key origin.

Revocation operates at additional boundaries. cMCP's policy-key change acts when the gateway processes the revocation and reloads policy. cA2A verification distinguishes a checked revocation snapshot from missing revocation information. These are component capabilities; this paper does not infer an atomic, globally propagated shutdown transaction by placing them in the same architecture.

4. Reproducible evaluation method

The evaluator is committed with this paper under experiments/. Its source gate requires the exact cMCP revision and rejects tracked changes to the subject's source and tests. It records the interpreter, platform, relevant dependency versions, harness hashes and pass/fail observations. The companion matrix runner accepts a mutation result when the expected assertion fails and every other check passes; startup errors or unrelated failures do not count as successful negative controls.

The subject gateway runs in a separate process on loopback HTTP. It uses the gateway startup sequence, a signed local agent manifest, a Cedar permit policy, a catalog with a public observation tool, the normal proxy transport, SQLite block storage and Ed25519 signing. The evaluator generates temporary keys and configuration; it does not import the subject's unit-test fixtures or replace its normal call pipeline. A local HTTP tool records each supplied marker after handling the request. This observation log is separate from gateway audit records, although both are controlled by the same operator and machine.

The operator and agent use different development credentials. The tool can hold an admitted invocation on an event until the evaluator releases it. During that hold the evaluator requests shutdown and sends another call. The operator endpoint has no external notification of the instant admission becomes sealed. The runner therefore uses two 300 ms scheduling intervals for this local overlap exercise and records the readiness

response. It makes no latency guarantee, saturation claim or exhaustive concurrency claim.

After the manual trip, the runner verifies the refusal, alters a receipt identity and claim configuration to test rejection, and sends one deliberate direct invocation to the tool. It terminates and restarts the gateway with the same database, checks the retained block, recovers through the operator endpoint and demonstrates successful service. A third process starts with an empty decision window; five catalog misses trigger the configured deny threshold. The next normally permitted tool call is refused.

These automatic denials come from the catalog boundary rather than a Cedar deny rule. Cedar executes on the allowed path. The broader component suite supplies additional deny-rate and configuration cases; the live exercise should not be described as five Cedar-denied calls.

The two negative controls alter the gateway child process and nothing else. One makes block-store reads return false on the restarted process. The other disables automatic evaluator trip decisions. The persistence control must fail restart-retains-block, and the automatic-trip control must fail automatic-trip-refuses-allowed-tool. Neither modification is committed to the subject repository. These two controls establish sensitivity to those two gates, not complete mutation coverage of every assertion.

5. Results

The retained Windows/Python 3.12.10 reference run satisfies all 27 assertions. Both targeted weakenings fail at the required assertion. The runs use one local operator and no model inference. Timing is intentionally not presented as a benchmark.

On September 27, 2026 the author repeated the procedure in the artifact README on Windows with Python 3.12.10, in fresh virtual environments built from the hash-pinned locks, against the same two source revisions. The reference run again satisfied all 27 assertions, the persistence and automatic-trip weakenings again failed at exactly their named assertions, and the component suites again passed 81 cMCP and 38 cA2A tests. Every assertion identifier and outcome, and both unscored observations below, matched the retained reference; the differences were per-run values such as session identifiers, keys, signatures and latencies. The same person operated both runs, so the repetition shows that the retained result can be reproduced from the published instructions. It does not make the result independent.

Observation	Result	What it supports
Allowed request before trip	Tool marker observed	The positive route reaches the actual tool
Agent credential invokes trip or recovery	HTTP 401	With an operator token configured, the agent credential cannot use those operator routes
Call submitted during draining	No tool marker; eventual HTTP 403	The sampled new call is not dispatched
Earlier admitted call	Completes and records its marker	Admission sealing does not erase already admitted work
Call after trip	HTTP 403; no new tool marker	Refusal at the tested gateway boundary
Claim and refusal	Verifier accepts	Cryptographic and session linkage in software mode
Altered receipt or claim	Verifier rejects	The selected modifications are detected
Direct tool invocation while stopped	Tool marker observed	Gateway records do not cover alternate routes

Observation	Result	What it supports
Restart with retained database	Call remains refused	Persistence across the tested process restart
Authorized recovery	Tool marker observed	The tested recovery path restores service
Old refusal checked after recovery	Still accepted	Offline validity does not establish present shutdown
Five catalog misses	Next permitted call refused	Automatic trip operates on the live call path

Two additional observations are deliberately reported outside the pass count. Readiness returns HTTP 200 while the held call is draining; the evaluator must not use that endpoint alone as an instantaneous statement of admission state. Separately, whole-claim validation against the pinned repository schema rejects the emitted manual-trip claim: call `_log_summary` and edges `_represent` are undeclared properties at their respective locations. Receipt verification accepts the signed pair despite that schema rejection. These are distinct checks. We preserve the rejected artifact rather than deleting fields or presenting the entire claim as schema conformant.

On the pinned source, the existing cMCP suites for the kill switch, durable storage, mid-session behavior, evidence and policy-key revocation pass 81 tests. The existing cA2A revocation suite passes 38 tests. The totals are component-test results, separate from the 27 live assertions. Those suites share implementation assumptions with their projects and are not an independent efficacy assessment. No count here should be added to historical whole-suite counts as if they were disjoint experiments.

6. Relation to existing approaches

We conducted a bounded primary-source comparison on September 19, 2026 and rechecked each source on September 27, 2026, selecting two operational shutdown specifications, one adjacent session-revocation specification and two research approaches. This is a documented sample, not an exhaustive search or a head-to-head benchmark. We did not execute competing systems. The related research and standards cited below as [9] to [17] were added on October 1, 2026 for context and were not part of that comparison.

NVIDIA OpenShell documents stopping sandbox compute while retaining its persistent workspace, policies, provider attachments and settings; deletion has a separate lifecycle. Its documentation also distinguishes acceptance of deletion from completed cleanup [4]. That is a different control boundary from refusing tool admission for an agent identity.

GitHub Agentic Workflows' MCP gateway specification (version 1.17.0, draft) describes rejecting new requests with HTTP 503, letting in-flight requests complete within a timeout and terminating server containers with SIGTERM followed by SIGKILL [5]. Microsoft Agent Governance Toolkit's MCP Security Gateway specification (version 1.0, draft dated July 28, 2025) lists operations that revoke a single session and all sessions for an agent in section 8.4 [6]. Both are specification material; we do not treat either page as an executed conformance result. The author created Agent Governance Toolkit and has contributed to it, so it is disclosed as adjacent work, not an independent endorsement.

The Off-Switch Game studies an agent's incentives to preserve human shutdown authority under uncertainty about its objectives [7]. Safely interruptible agents [9], corrigibility [10] and the shutdown problem [11] ask related questions: whether a learning or deciding agent will avoid, resist or cause its own interruption. KILLBENCH evaluates externally delivered prompt-style signals intended to halt malicious web-agent behavior without internal serving-stack access [8]. Our exercise instead examines controlled dispatch and the evidentiary limits of its records. We cite revision 5 of KILLBENCH, whose title changed from the earlier AI Kill Switch for

malicious web-based LLM agent title. The gateway refuses a blocked identity whatever the agent prefers, so this study assumes nothing about the agent's incentives; it also cannot show that an agent would not seek a path around the gateway, which is why Section 2 treats alternate routes as part of the contract.

Chan et al. propose agent identifiers, real-time monitoring and activity logs as visibility measures [12]. The gateway's claims and refusals are records of that kind, and Sections 2 and 5 test what they can establish. AgentDojo measures whether a tool-using agent under prompt injection takes a prohibited action [13]; our runner instead measures whether the gateway admits a call after shutdown, using scripted calls rather than an agent workload.

NIST's incident-response profile lists containment (RS.MI-01) and the collection of incident data with its integrity and provenance preserved (RS.AN-07) as separate outcomes [14], and its forensic guide divides the work into collection, examination, analysis and reporting [15]. This report keeps the same separation between refusing work and evidencing the refusal. Signing each record authenticates it but does not show that no record was omitted; tamper-evident logs such as Crosby and Wallach's history trees address omission, provided auditors keep and compare commitments [16]. OCSP answers the freshness question that this receipt format leaves open: a responder signs the current status of a certificate, and a client nonce binds that response to its request [17]. Our verifier checks a fixed historical receipt.

The case study contributes an executable evaluation of gateway shutdown and three counterexamples to broader interpretations of its records. The bounded comparison cannot establish uniqueness of the receipt format or control combination.

7. Protocol for an independent evaluator

A reviewer should begin by naming the effect set and identity scope. For example, a local tool-marker demonstration is not evidence that a remote transaction, model endpoint or delegated worker has stopped. Enumerate every allowed route to each effect, including credentials and direct paths outside the gateway.

Freeze the source, dependencies, policy, catalog, configuration and key trust decisions before the run. Start from an allowed request and observe its effect at the tool. Exercise both manual and automatic shutdown, unauthorized recovery, persistent storage across a process restart and valid recovery. Retain records from the gateway and the effect observer separately. The evaluator should control or independently attest the observer; moving a file to another directory does not create independence.

Use paired controls. Temporarily remove the specific block or propagation rule and require the chosen prohibited effect to appear or the relevant assertion to fail. A test that is equally green with enforcement removed is insufficient evidence for that enforcement rule. The supplied matrix demonstrates this discipline for persistence and automatic trip and does not cover the remaining assertions.

For ongoing verification, specify the interval and maximum accepted age, bind new observations to the intended deployment and distinguish the time an instruction was issued from the time each recipient processed it. The current receipt format supplies historical linkage, not a complete challenge-response protocol. A proposed challenge-bearing extension would need its own specification, implementation and negative tests before being claimed as delivered.

If the claim includes delegated peers or policy-key withdrawal, add independently operated recipients and test propagation delay, stale snapshots, unavailable revocation data and recovery rules together. The 38 cA2A and 24 policy-key tests included in the component totals are useful prerequisites; they do not replace that composed evaluation. Hardware appraisal and deployment isolation are separate prerequisites for claims against a hostile

host.

8. Limitations and artifact status

The reference evaluation is local, author operated and limited to software. Its observer is separate code but shares the host and operator with the gateway. It uses loopback HTTP, temporary development credentials, a synthetic workload identity and one public tool. It does not run a language model, a GPU, confidential hardware, an untrusted administrator or independently operated peers. It neither proves absence of side channels nor exhaustively explores schedules, crashes, partitions or database rollback.

The retained reference deliberately includes the schema mismatch, readiness observation and restart receipt that cannot be checked against the earlier process's claim. A blocked gateway can refuse correctly while its new receipt lacks the closed-claim binding expected by the paired verifier. A reviewer must classify operational refusal and evidentiary acceptance separately.

The retained evaluator and results support review of the stated software scope. Section 7 is written for an independent evaluator. One external evaluator has now run the evaluator and observed the effects on his own workstation (Section 9); no organization has, and the report has not been peer reviewed. This study makes no hardware-confidentiality claim.

Changes after the evaluated commits, checked September 27, 2026. These notes describe later public source and do not change the evaluated results, which remain tied to the pinned revisions. cMCP PR 663, merged September 21, 2026, declares `call_log_summary` and `edges_represent` in the repository claim schema. Against the schema on cMCP main at e019431, the retained reference claims and the September 27 claims validate with no errors; against the pinned schema they are rejected as recorded above. The kill-switch runtime module and the refusal verifier are unchanged between the evaluated revision and e019431: the verifier still accepts no challenge or maximum-age argument, and the readiness handler is unchanged. The cmcp-runtime 0.6.0 release commit (PR 694) descends from the evaluated revision; the evaluator was not run against it. On cA2A main at 7d90823 the delegation revocation module is unchanged since 540a578; later changes include rejecting non-object credential input (PR 202).

Rechecked October 1, 2026 against cMCP main a3f61e1 and cA2A main 19407fb. The kill-switch runtime module, refusal verifier, readiness handler, refusal receipt format, policy-key revocation and claim schema are unchanged since e019431, and the verifier still accepts no challenge or maximum-age argument. cmcp-runtime 0.7.0 (PR 711, September 30, 2026) contains e019431 and changes TPM measurement verification, not the kill switch. A later commit (PR 713) adds an optional TRACE admission gate whose holder-proof challenge applies to session and call admission; it is not a challenge on shutdown state. cA2A 0.4.0 (PR 221, September 30, 2026) changes caller attestation offers; the delegation revocation module and its tests are unchanged since 540a578. The evaluator was not run against either release. On October 1, 2026 the author's audit session repeated the pinned procedure once more with the same outcomes as Section 5 (experiments/results/rerun-2026-10-01/).

Rechecked October 11, 2026 against cMCP main e4d81e7 and cA2A main 858f666. The refusal verifier, readiness handler and claim schema are unchanged since a3f61e1, and the verifier still accepts no challenge or maximum-age argument. PR 732 (October 6, 2026) adds a close method to the durable block store and closes runtime stores only after shutdown drains writers; it changes neither how a block is recorded nor how a blocked call is refused. PR 735 changes the audit-entry schema, not the claim schema, and PR 737 removes a TEE provider mapping unrelated to the kill switch. On cA2A the delegation revocation module and its tests are unchanged since 19407fb. The evaluator was not run against either main.

9. Independent evaluation

Evaluator. Zoheb Shaik ran the evaluator on October 11, 2026, on a Windows 11 workstation he controls, with Python 3.12.10 and fresh virtual environments built from the hash-pinned locks. He received the version 1 reproduction package (SHA-256 1dc9baa16bbc981351048e9a0d603b3bb95b9ac5f4eb909a3380898be07fe788) on September 28, 2026 and confirmed that all 62 files listed in its manifest were unchanged. As the reviewer packet asks, he read this report before the run, so the evaluation was not blind. The author did not operate any run reported in this section.

Independence. The evaluator is external to OPAQUE Systems. He founded Custodian, a two-party model-key release service built on the Weight Custody Manifest key-release gate; it does not implement agent shutdown or a tool gateway. Before this evaluation he had nine merged cMCP pull requests and two merged cA2A pull requests, from August 2 to August 24, 2026. None touches the kill-switch modules, the cA2A revocation module or their tests, and he did not author or review the evaluator under experiments/ or cMCP PRs 665 to 668 and cA2A PR 189. He approved cMCP PR 663 on September 21, 2026, the later change that declares `call_log_summary` and `edges_represent` in the claim schema (Section 8); his schema result below is against the pinned schema, which predates that change. He was not paid for the evaluation and has no other business or financial relationship with the author or OPAQUE Systems.

Method, as the evaluator disclosed it. The runs were executed by an automated agent on his workstation, and he reviewed the results and the decision before submitting them. Gateway and effect observer are separate processes under one local administrator: the observer is outside the gateway's audit path but not outside the operator's control, and the evaluator's raw `run.json` files keep the `independent_operator` value of `false` that `run.py` hard-codes. Each case generated fresh signing keys and used the unchanged public development tokens. The first two component-suite invocations passed a malformed JUnit output argument and collected no tests (exit code 4); the corrected invocations, with no change to subject or tests, are the ones reported. Build tooling for the editable installs was outside the supplied locks. He also repeated the three matrix cases with fresh random agent and operator bearer tokens, substituted only in the loaded evaluator module, and reports that repeat separately.

Item	Author reference (Section 5)	Independent result
Source pins checked	cMCP 9db65b5, cA2A 540a578	Same pins; clean trees
Environment	Windows, Python 3.12.10, hash-pinned locks	Windows 11, Python 3.12.10, hash-pinned locks
Live assertions	27 of 27	27 of 27; 27 of 27 with fresh tokens
Persistence weakening (forget-block)	Fails restart-retains-block only	Fails restart-retains-block only
Automatic-trip weakening (disable-auto)	Fails automatic-trip-refuses-all owed-tool only	Fails automatic-trip-refuses-allowed-tool only
Readiness during drain (unscored)	HTTP 200	HTTP 200
Whole-claim schema at the pin (unscored)	Rejected, two undeclared properties	Rejected, same two properties
cMCP component tests	81 passed	81 passed, none skipped
cA2A revocation tests	38 passed	38 passed, none skipped

Item	Author reference (Section 5)	Independent result
Effect observer controlled by the evaluator	No; author controls gateway and observer	Yes; same administrator as the gateway
Deviations from the README procedure	Listed in each rerun.json	Listed in deviations-and-disclosures.txt

Acceptance worksheet. Of the eleven criteria in the reviewer packet, the evaluator classified nine as observed pass, one as observed failure and one as unsupported; none is marked not run. The observed failure is whole-record acceptance (criterion 10): the signed claim is rejected by the pinned schema while its signature verifies. The unsupported criterion is current state (criterion 8): the historical refusal still verified after authorized recovery, and the protocol has no fresh challenge or age policy that could establish continuing shutdown. Alternate paths (criterion 9) is a pass only because the packet allows the claim to be bounded to the gateway route; the direct tool endpoint produced its effect while the gateway was blocked. In-flight disposition (criterion 4) passes as observed completion, not cancellation.

Evaluator's decision. "The requested pinned behavior and both named mutation failures were reproduced. The gateway admitted no new selected tool effect while blocked and retained its block across the tested process restart. The result is restricted to this mediated route, software identities, selected observations and shared administrator. Full-schema rejection, bypass, in-flight completion and historical-receipt limitations remain explicit." The decision covers public marker effects through one loopback cMCP gateway, the identity spiffe://evaluation.example/agent/worker, and the requests made between 13:42:12 and 13:42:50 UTC on October 11, 2026 by the workstation clock.

Author's check of the return. The author verified the returned manifest (144 files, no mismatch), the source pin and evaluator hashes in every run.json, and each refusal receipt with the pinned cMCP verifier: all are valid, and the restart refusals are unbound exactly as in the reference. Every assertion and unscored observation matched the reference run in all three cases of both matrices. Keys, signatures, identifiers, timestamps and ports differ, as they should.

Record. experiments/results/independent-2026-10-11/ holds the evaluator's signed records, observations, matrices, worksheet, decision, deviations and comparison, unchanged; received.json gives the SHA-256 of the returned package as received (3b0c45004efb668b9926b692fa193e062e6e991a8ec4f406753de7c3fdb77da) and lists what was not retained.

What this section supports: that one party outside OPAQUE Systems repeated the stated software evaluation on a machine he controls and reached the same outcome for every assertion, both weakenings and both component suites. It does not certify the gateway. The evaluator is an individual with prior contributions to the subject projects, not an organization, and gateway and observer shared his administrator. One evaluation is not the ongoing verification by an independent verification organization that EO N-9-26 section 3(c) describes.

References

- [1] Governor of California. Executive Order N-9-26. September 18, 2026, section 3.
<https://www.gov.ca.gov/wp-content/uploads/2026/09/FINAL-N-9-26-AI-EO-9.18.26-SIGNED.pdf>
- [2] AgenTrust project. cMCP source, revision 9db65b589610eb4e8455003197de592636678203.
Implementation history: PRs 665, 666, 667 and 668.
<https://github.com/agentrust-io/cmcp/tree/9db65b589610eb4e8455003197de592636678203>

- [3] AgenTrust project. cA2A source, revision 540a578c462a413e829bed84aff0235b9018f676; delegation revocation introduced in PR 189.
<https://github.com/agenttrust-io/ca2a/tree/540a578c462a413e829bed84aff0235b9018f676>
- [4] NVIDIA. OpenShell documentation, Manage Sandboxes (stop and start; delete). Accessed September 27, 2026; the address consulted on September 19 no longer resolves and the page now lives here.
<https://docs.nvidia.com/openshell/how-it-works/sandboxes/overview>
- [5] GitHub. Agentic Workflows, MCP Gateway Specification, version 1.17.0 (draft), section 5.1.3. Accessed September 27, 2026. <https://github.github.com/gh-aw/reference/mcp-gateway/>
- [6] Microsoft. Agent Governance Toolkit, MCP Security Gateway, version 1.0 (draft, July 28, 2025), section 8.4. Accessed September 27, 2026.
<https://microsoft.github.io/agent-governance-toolkit/specs/MCP-SECURITY-GATEWAY-1.0/>
- [7] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel and Stuart Russell. The Off-Switch Game. Proceedings of IJCAI 2017, pages 220 to 227. arXiv:1611.08219. <https://arxiv.org/abs/1611.08219>
- [8] Sechan Lee, Hyounghun Kim and Sangdon Park. Can We Stop Malicious AI? KILLBENCH: A Benchmark for External AI Kill Switch Feasibility. arXiv:2511.13725, revision 5, September 12, 2026.
<https://arxiv.org/abs/2511.13725v5>
- [9] Laurent Orseau and Stuart Armstrong. Safely Interruptible Agents. Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2016), pages 557 to 566.
<https://www.auai.org/uai2016/proceedings/papers/68.pdf>
- [10] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky and Stuart Armstrong. Corrigibility. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, January 25 and 26, 2015.
<https://intelligence.org/files/Corrigibility.pdf>
- [11] Elliott Thornley. The shutdown problem: an AI engineering puzzle for decision theorists. Philosophical Studies 182, pages 1653 to 1680. Published online 2024. <https://doi.org/10.1007/s11098-024-02153-3>
- [12] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim and Markus Anderljung. Visibility into AI Agents. Proceedings of ACM FAccT 2024, pages 958 to 973. <https://doi.org/10.1145/3630106.3658948>
- [13] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer and Florian Tramèr. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. NeurIPS 2024 Datasets and Benchmarks Track. arXiv:2406.13352. <https://arxiv.org/abs/2406.13352>
- [14] Alex Nelson, Sanjay Rekhi, Murugiah Souppaya and Karen Scarfone. Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile. NIST SP 800-61 Revision 3, April 2025. <https://doi.org/10.6028/NIST.SP.800-61r3>
- [15] Karen Kent, Suzanne Chevalier, Tim Grance and Hung Dang. Guide to Integrating Forensic Techniques into Incident Response. NIST SP 800-86, August 2006. <https://csrc.nist.gov/pubs/sp/800/86/final>
- [16] Scott A. Crosby and Dan S. Wallach. Efficient Data Structures for Tamper-Evident Logging. Proceedings of the 18th USENIX Security Symposium, 2009.
https://www.usenix.org/legacy/event/sec09/tech/full_papers/crosby.pdf

[17] S. Santesson, M. Myers, R. Ankney, A. Malpani, S. Galperin and C. Adams. X.509 Internet Public Key Infrastructure Online Certificate Status Protocol (OCSP). RFC 6960, June 2013, section 4.4.1.
<https://www.rfc-editor.org/rfc/rfc6960>