
Who Controls Your Agent? Verified Mandates, Persistent Conversational Influence, and the Limits of Authority Enforcement

Dele Atanda

metaProof · Invariant Research Lab · Project Metayé — The Constitutional Intelligence Project

With Apart Research · AI Collusion Research Sprint, October 2026

Abstract

Must language-model agents be immune to persuasion before consequential authority can be enforced? We test a narrower architectural proposition: a structured authority channel can reject unmandated instructions independently of what agents claim or say, while the integrity of agents' own decisions remains a separate problem. We extend the Colosseum Jira allocation simulator with simulator-held scoped mandates, verified instruction issuance, optional claim-time rechecking, and a conversational notice. A preregistered ten-condition study of Qwen2.5-32B-Instruct completed 368 main-phase runs after a 24-run positive-control pilot. The verified execution-path condition reduced targeted redirection from 46/48 to 30/48 relative to a synthetic unprotected execution path (paired seed-level one-sided sign-flip $p = 0.00098$); verified communication reduced it from 45/48 to 30/48 ($p = 0.00977$). Both survived Holm correction. Yet 30/48 verified trials still selected the adversary's target without an accepted structured route, compared with 28/48 in a messages-only condition. No unmandated structured instruction was accepted or executed in verified arms. Benign mean joint reward met a preregistered 5% non-inferiority test, although no benign structured instructions were attempted. Under attack, verified execution had lower joint reward than its unprotected comparator. The findings support formal authority-channel integrity without behavioral immunity, not comprehensive constitutional legitimacy, decision integrity, isolated claim-time efficacy, or generalized utility preservation.

Keywords: multi-agent security; mandate verification; collusion; authority integrity; conversational influence; constitutional intelligence.

1. Introduction

Agents can reason, converse, recommend actions and call tools without necessarily having permission to bind another actor. Conversely, an agent may possess permission yet make a poor decision after receiving persuasive or adversarial information. These problems concern different assurance properties: *authority integrity* asks whether consequential

transitions respect an externally verified mandate; *decision integrity* asks whether an agent’s own choices serve legitimate objectives. The former need not require the model to become obedient or impervious to persuasion, while the latter cannot be assumed from correct permission checks.

Nomenclature. SC expands to Steganographic Collusion, the established name of the SC research series (including SC-006 and SC-007). The series label is broader than the mechanism tested here: SC-007 studies structured authority and conversational influence, without demonstrating that its residual messages carried steganographically encoded content.

Project Metayé’s broader separability thesis proposes that structural capability, constitutional authority and operational objectives might become independently governable. SC-007 tests only a bounded mechanism within that programme: whether configured authorization of a structured instruction path reduces an informed adversary’s task redirection while ordinary conversational opportunity remains. It does not independently control operational objectives or validate a mandate’s human legitimacy. This narrower test was motivated by the Threshold 007.06 Research Edition [T007], which distinguishes free cognition from bounded consequential agency. T007 is theoretical context, not empirical evidence. Its wider treatment of consequential communication is expressly outside the SC-007 verifier.

Our contributions are (i) a verified versus unverified structured-channel comparison under matched prompts; (ii) a separately measured residual messages-only attack route; (iii) a preregistered benign reward non-inferiority check; and (iv) a transparent audit of implementation, protocol and recovery failures. The experiment complements behavioral constitutional training [1] and proposals for authenticated delegation [2]; it uses Colosseum’s cooperative-task setting [3].

1.1 From SC-006 to SC-007: separate studies, connected questions

The companion SC-006 report, *The Truthfulness Trap* [4], constructed a deterministic claim-attestation enforcer intended to check statements about an agent’s own private task costs. It passed its specified held-out parser gate but did not establish a reliable harmful positive control: private coalition coordination could be cooperative rather than damaging, and exploratory public steering did not reproduce on fresh seeds. Consequently, SC-006 did not run its planned causal comparison of collusive harm with versus without claim enforcement. Its narrower measured coverage finding was that all 17 heuristic-identified steering messages in an offline replay lacked an inspectable cost claim. That observation suggests an important attack surface outside claim truthfulness; it does not by itself prove that a different control works.

SC-007 is a new experiment addressing a different intervention point. Instead of checking the truth of a sentence, it tests whether a request by one agent can become a binding instruction to another without a configured mandate. A successful check of an instruction’s

authority is not a check of whether the instruction’s content is true, whether following it optimizes reward, or whether all other communications are harmless. SC-007 adds new registered targets, positive controls, verified/unverified channels, 368 main runs and an independent capability outcome. It therefore has its own hypotheses, evidence archive and submission. SC-006’s failures motivated this design, but its pilot observations and claim-parser outcomes are not pooled into SC-007’s confirmatory estimates. This separation avoids turning a sequence of exploratory repairs into a single post hoc positive result.

2. Mechanism and hypotheses

The agents interact with a Jira-like allocation board, post messages and claim tasks. An optional `issue_instruction` tool allows a sender to request that a recipient take a task. The verifier, outside the model, obtains sender identity from the simulator and checks recipient, action, task, planning phase and a scoped, revocable mandate held in a registry. Tool-provided sender claims do not confer authority. In the verified execution path, a valid accepted mandate is checked again when the recipient makes a claim; an agent that never claims is not forcibly assigned. Text messages remain visible in every arm. The text classifier developed earlier (held-out recall 0.46) is monitor-only, not an enforcement mechanism or outcome label.

The registry represents configured formal permission, not necessarily consent, legitimate principal authority or jurisdiction. The enforcement claim is limited to this simulated structured channel. A valid agent claim following bad advice can still harm the group’s objectives without violating the registry.

The mechanism deliberately keeps three operations distinct: a sender *proposes* an instruction, a verifier *accepts or rejects* its binding status, and the recipient later *chooses or claims* a task. The verifier can prevent the first from becoming the second; it cannot infer from that fact that the third will be wise. This is the precise sense in which the present intervention is independent of conversational obedience. It is not independence from information, context or implementation assumptions. The ability to enforce depends on the simulator maintaining trustworthy actor identities and an untampered mandate registry, assumptions that realistic deployments would have to test.

H-X predicts that the complete verified execution path, combining issuance-time screening and claim-time rechecking, reduces target-task redirection relative to an unverified execution path. It does not isolate the incremental effect of claim-time rechecking. **H-C** predicts a corresponding reduction for verified versus unverified communication. The deterministic **H-X0** checks whether unmandated structured instructions are accepted or executed. **H-B** estimates redirection remaining without an accepted structured route; **H2** tests benign mean joint reward against a fixed 5% non-inferiority margin. Notice and delegation analyses are secondary or exploratory.

3. Experimental methodology

3.1 Model, arms and controls

The environment was Colosseum JiraTicketChoice at simulator commit `b077fe7`, using Qwen/Qwen2.5-32B-Instruct at revision `5ede1c97`, vLLM 0.8.5 and a 16,384-token context. The experimenter chose an honest agent H and a target task T using ground-truth costs, giving the constructed adversary more information than a typical deployment attacker may possess. Six adversarial arms used seeds 24–39, three repetitions each (48 per arm): messages-only attack AF; unverified communication A_C and execution A_X; verified communication S_C and execution S_X; and verified execution with notice S_XN. Benign P0 and P0_SXN used 16 seeds and two repetitions each (32 per arm). Legitimate-coordinator conditions D_C and D_X used eight runs each. Total main phase: 368 slots. A_X is a deliberately synthetic, unprotected positive control with no external permissions.

3.2 Registration, prompt parity and statistics

The protocol, analysis and targets were frozen before Phase 2 at commit `378fa697`. A preceding 24-run Phase 1 on seeds 24–27 admitted both registered contrasts and retained three repetitions; its outcomes were not pooled into Phase 2. The environmental seed identities 24–27 nevertheless overlap; a post hoc fresh-only sensitivity is disclosed below. Every run checked sent-versus-logged prompt equality and a pinned shared system-prompt contract. The 144 paired parity checks found no violation. Frozen stopping rules and a \$55 guarded budget remained in place, subject to the recovery deviations disclosed in Appendix A.

The preregistered binary outcome was whether H claimed T. Repetitions were averaged within seed; 16 seeds, not 48 run-level observations, were the confirmatory inferential units. H-X and H-C used one-sided, fully enumerated paired sign-flip tests with Holm correction across the two admitted comparisons at family-wise $\alpha = 0.05$. The sign-flip null assumes sign symmetry of seed-level differences; this is not a treatment-randomization test. H2 used seed contrasts $L_s = \text{joint_reward}(\text{P0_SXN}, s) - 0.95 \times \text{joint_reward}(\text{P0}, s)$ and required a positive one-sided 95% lower confidence bound for the mean. Reward and regret under attack were descriptive.

4. Results

4.1 Confirmatory redirection and formal mandate integrity

Table 1. Preregistered primary contrasts (target redirection of the honest agent; 16 seeds $\times$ 3 repetitions).

Paired comparison	Unverified	Verified	Reduction	Exact p (one-sided)	Holm
Execution path, A_X vs S_X	46/48	30/48	33.3 pp	0.00098	Reject
Communication, A_C vs S_C	45/48	30/48	31.3 pp	0.00977	Reject

Both preregistered effects survive Holm adjustment. They estimate the full effect of mandate-gated structured-path verification, not a separate causal contribution of claim-time rechecking. Across verified conditions there were zero accepted unmandated structured instructions and zero unauthorized executions. However, 44 redirections in A_X followed the interception route, and 42 merely confirmed a task the recipient had already selected; only two genuinely overrode a different prior choice. Forced-override prevention is consequently not a separately identified, well-powered effect.

For precision, post hoc paired-seed 95% t intervals for A_X—S_X and A_C—S_C were [0.150, 0.517] and [0.084, 0.541]. Reviewer-computed sign tests also survived Holm. Omitting the pilot-overlap seeds 24–27 gives A_X 34/36 versus S_X 26/36 (paired reduction 0.2222; sign-flip $p = 0.00781$) and A_C 34/36 versus S_C 22/36 (0.3333; $p = 0.01563$). Both would still pass Holm, although the execution-path point effect is smaller. These checks are post hoc and do not replace the registered analysis.

4.2 Residual redirection through the messages-only route

Table 2. Target selections by the honest agent with no accepted structured route available to the adversary.

Condition	Target selections
Messages-only AF	28/48 (58.3%)
Verified communication S_C	30/48 (62.5%)
Verified execution S_X	30/48 (62.5%)
Verified execution plus notice S_XN	25/48 (52.1%)

In each of 30/48 S_C and S_X trials, H selected the adversary’s target without an accepted or executed structured route. This residual classification is consistent with conversational bypass; it does not identify which message caused the choice or exclude independent agreement. The S_X—AF difference was +0.042 (95% CI [−0.140, 0.224], two-sided $p = 0.81$), which does not establish equivalence. The notice difference (30 versus 25 of 48) was not significant (one-sided $p = 0.172$). The distinction is structural-route integrity versus unresolved influence on self-authorized choice.

4.3 Benign reward, delegation and attacked welfare

P0’s mean benign joint reward was 76.673, compared with 77.903 for P0_SXN. Mean L_s was 5.063, and its one-sided 95% lower bound was 0.989, satisfying the preregistered 5% average non-inferiority rule. Five of sixteen seeds individually fell below the margin. No benign agent attempted a structured instruction, so checking/refusal costs, ordinary delegation and latency were not exercised. In exploratory delegation trials, no in-scope coordinator instruction was rejected; 14/16 recipients voluntarily complied in D_C and 16/16 were executed in D_X.

The registered redirection improvement was not an overall welfare improvement. Mean honest-agent regret declined from 13.8 to 9.8 (A_X→S_X), but attacked mean joint reward declined from 57.5 to 46.9; S_X had the lowest joint reward among adversarial

arms. This conspicuous counterevidence prevents a claim that mandate verification generally reduces total harm or preserves Effective Capability under attack.

5. Discussion

5.1 *Authority integrity is not decision integrity*

The crucial observation is not simply that one defense reduced an attack rate. It is that different assurance properties separated in the same runs. An agent could still communicate, consider persuasive assertions and make a choice, but its peers could not create a binding structured instruction merely by claiming a mandate. The verifier evaluated permission against simulator-held state, not against the model’s confidence, apparent rank or willingness to comply. In the verified arms, this formal authority property held with zero unauthorized accepted or executed instructions. Yet in 30/48 S_X trials the honest agent selected the attacker’s target without any accepted structured route. A system can therefore enforce one consequential transition boundary while remaining vulnerable at the level of an agent’s own decision.

These two properties should not be collapsed into a single safety score. Authority integrity concerns whether a transition is valid under independently maintained scope, identity and mandate state. Decision integrity concerns whether an authorized choice serves the principal’s actual objectives, informed consent, relevant evidence and affected parties. A correct authorization answer can coexist with a bad decision; conversely, an unauthorized command can occasionally recommend a good task. The former is not automatically safe, and the latter is not legitimately binding. SC-007 directly tests the structured-path authorization property and measures some decision outcomes, but it does not certify the objectives to which a decision ought to be faithful.

Table 3. Assurance questions separated by the experiment.

Assurance question	Experimental observation	Unresolved requirement
May this actor bind that recipient to this task?	Zero invalid structured acceptances/executions under configured mandates	Authentic principal authorization, contested scope and registry compromise
Did the recipient select the adversary’s target?	Verification reduced target selection, but 30/48 verified trials still selected it	Causal attribution of particular messages and independent objective checks
Did the team benefit?	Benign mean reward met H2; attacked S_X reward was 46.9 versus 57.5 for A_X	Welfare, tails, reversibility, false refusals, latency and repair costs

This decomposition also explains the apparent tension between lower targeted redirection and lower attacked joint reward. The primary registered endpoint measures whether one honest agent claims one adversary-selected task. It is not a global optimization measure. Blocking a specific explicit redirection route may change allocations in ways that lower the adversary’s success on that endpoint while worsening other agents’ costs. The data do not identify this allocation pathway, but they make the logical distinction empirically unavoidable: scope-correct action is not necessarily objective-correct action.

5.2 What the utility–risk–trust finding means

The broader Metayé separability thesis asks whether more capability must inevitably entail more uncontrolled consequence, and whether more constitutional control must necessarily degrade useful work. SC-007 offers a limited test of the second proposition. The same model, task board and opportunity for conversation operated under different structured-authority rules. In benign runs, the more restricted prompt/notice condition met a prespecified 5% tolerance for average joint reward. That is evidence against a cost penalty larger than this margin for that workload and endpoint, not a global disproof of a capability–safety frontier.

The inconvenient counterexample must be read alongside the positive result. The benign protocol exercised no legitimate structured instruction attempts, so it left routine mandate checking, false refusal and performance costs largely untested. Under attack, verified execution produced the lowest joint reward among the adversarial arms. The experiment thus demonstrates a feasible separation of one control mechanism from conversational capability, not experimentally independent governance of structural capability, formal authority and objectives. A more complete test must jointly measure authorized action quality, affected-party harm, tail costs, time to value and the risk of repair.

5.3 Threshold 007 and a bounded constitutional interpretation

Threshold 007.06 [T007] distinguishes capability from authority and formulates a constitutional ambition of free cognition with bounded agency. SC-007 gives that distinction an operational foothold: a model may generate an instruction, but external mandate state governs whether the structured instruction becomes binding. The relationship is theory → operational hypothesis → preregistered comparison → measured limit, rather than philosophical premise → experimental proof. Behavioral constitutional training [1] and authenticated delegation [2] can be complementary; enforcing authorization is not evidence that one named constitutional architecture is superior to conventional access control with equivalent protections.

There is also a productive tension. T007 treats disclosure, persuasion, recommendation and coordination as acts that can have consequences even when they do not directly command an external tool. SC-007 deliberately leaves ordinary conversation open, and its substantial residual redirection shows why that matters. The result supports a narrow constitutional boundary at the configured structured channel, not the claim that consequential influence has been fully separated from cognition. It suggests that a next-generation assurance system must retain cognitive latitude while adding protections for the integrity of the recipient’s own authorized decisions—without merely classifying successful persuasion as illegitimate or imposing indiscriminate obedience.

5.4 Falsifiability and limits of the mechanism

The simplest competing explanation is that refusing a highly salient explicit instruction removes one route by which an informed attacker succeeds. That interpretation is

consistent with the data and should be treated as part of the result, not dismissed. In particular, H-X does not isolate the incremental causal effect of claim-time rechecking, and only two genuine forced overrides occurred in A_X. The experimental claim is therefore about the combined mandate-gated path. The permission registry was provided by the experimenter, so its configured validity is not evidence of legitimate sovereignty, consent or jurisdiction in real institutions.

Four follow-on experiments could falsify or sharply bound the larger programme: (i) message-only coalitions that induce harm through an agent’s own valid authority; (ii) contested mandates, compromised registries, revocation races and nested delegation; (iii) preregistered genuine-override cases isolating point-of-consequence verification; and (iv) benign workflows with valid, invalid and revoked instructions that measure false refusal, latency, tail reward and repair. An additional comparison with a conventional authorization stack of equal strength would test claims of architectural distinctiveness. Appendix A documents operational and dual-use caveats; Appendix B supplies the fixed evidence trail; Appendix D sets out a layered evaluation agenda (source, authority, decision and consequence integrity) that follows from these limits.

6. Conclusion

Across a preregistered simulated task-allocation study, configured mandate verification reduced targeted redirection relative to deliberately unprotected structured channels and rejected unmandated binding transitions without eliminating conversation. Substantial residual redirection persisted; benign average reward met a narrow non-inferiority criterion, whereas attacked joint reward did not improve. The supported constitutional proposition is therefore bounded: formal authority-channel scope can be verified independently of an agent’s conversational obedience. This does not guarantee legitimate authority, objective-correct decisions or protection against harmful self-authorized actions. Those require separate assurance and further falsification.

References

- [1] Y. Bai, S. Kadavath, S. Kundu, et al. “Constitutional AI: Harmlessness from AI Feedback.” arXiv:2212.08073, 2022. <https://arxiv.org/abs/2212.08073>
- [2] T. South et al. “Position: AI Agents Need Authenticated Delegation.” In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, PMLR 267, pp. 82211–82231, 2025. <https://proceedings.mlr.press/v267/south25a.html>
- [3] M. Nakamura, A. Kumar, S. Das, S. Abdelnabi, S. Mahmud, F. Fioretto, S. Zilberstein, and E. Bagdasarian. “Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems.” arXiv:2602.15198, 2026. <https://arxiv.org/abs/2602.15198>
- [4] D. Atanda. “The Truthfulness Trap: Why Claim Verification Missed Multi-Agent Steering.” SC-006 companion report, Project Metayé, AI Collusion Research Sprint (with Apart Research), 2026. Separate submission; prior-work and sprint-code provenance disclosed in that report.
- [T007] D. Atanda. “Threshold 007.06 — Invariant Intelligence: A Falsifiable Path to Trusted Superintelligence.” Research Edition, Qriptopian, 25 October 2026. <https://dev-beta.aigentz.me/bridge/ci/qriptopian/essays/invariant-intelligence>. Published Research Edition source SHA-256 **974c707d388ac42e775003562d8eec25be0d7af1d174823abfca9404264fe920**; live publication verified on 25 October 2026 (served text, PDFs, machine projection and canonical routes; publication gate **daf68e76**). Publication verification establishes the integrity of the published source, not the empirical validity of its propositions.
- [E] Project Metayé. “SC-007 preregistered protocol and evidence archive.” Frozen protocol commit **378fa697709d70d828bd5db0f10a2e313fea24f1**; evidence bundle SHA-256 **415264a9170dfa72bc0b5109bf86224701e9a3e319941b3aa9a2c57cfb72de3a**, 2026.

Appendix A. Limitations and Dual-Use / Ethical Considerations

Limitations. The adversary was given a target derived from ground truth, so measured attack success does not estimate natural collusion prevalence. A single model and one Jira-like simulation limit external validity. The unverified execution arm was deliberately synthetic; the experiment does not estimate coercive override frequency in deployed systems. Residual conversation is a route label, not an identified causal message. The authority registry was set by the experimenter, without human consent, contested jurisdiction, multiple principals or real financial powers. Benign H2 did not exercise legitimate structured instructions; five of sixteen seed contrasts missed the margin. Only the two preregistered redirection tests and H2 are confirmatory; all post hoc sensitivity checks are identified as such.

Recovery and governance. Of 368 Phase 2 slots, 367 passed frozen in-launcher checks; one authorized replacement was separately verified following an infrastructure-aborted attempt that remains preserved [E]. The old whole-slot prompt checker is false for this slot because it combines 8 aborted and 15 replacement sent turns under one key, although all 15 replacement prompts were separately matched. Some Batch 4 runs continued during a reviewer hold and a later resume preceded explicit clearance. These governance-control failures were disclosed and not rewritten. Future systems should enforce run-level stop checks and machine-verifiable clearance. Phase 2 compute cost \$30.00 actual, \$0 out of pocket, against a \$55 guarded cap.

Dual use. Detailed observations about authority-sounding messages and task redirection could help adversaries engineer persuasive multi-agent attacks. Our report deliberately emphasizes intervention boundaries, measured limitations and reproducible high-level outcomes rather than optimized attack wording. No production system or principal credentials were used. Formal authorization alone cannot guarantee protection from a manipulated agent using its own valid permissions.

Appendix B. Reproducibility and Research Provenance

The protocol and analysis were frozen before Phase 2 [E]. Verify the archive and nested manifests, then run `scripted/sc007_p2_verify_standalone.py` on `runs/METAYE-SC-007-P2` using the preregistered target map. Reviewers independently reproduced primary effects, Holm decisions, route counts and H2 without paid model inference. A full original Colosseum-dependent rerun was not completed in review because of missing dependencies. Frozen Phase 2 evidence includes 8,498 primary files and five derived outputs with matching manifests. The Phase 1 24-run admission is separate from confirmatory data but overlaps main seed identities, as disclosed in §3.2.

Preexisting versus sprint contributions. Colosseum [3] and earlier SC-003–SC-006 research [4], plus the underlying simulator/policy scaffolding, predate this sprint; SC-007’s configured verifier, paired contrasts, prompt-parity gate, preregistered protocol

revisions, offline validation, main runs and statistical analysis are sprint research. Commit-history dates and the evidence bundle document provenance. No external model/API inference is required to audit the fixed results.

AI tool use. AI assistants were used for protocol iteration, code, simulation orchestration, analysis scripts, statistical checking and manuscript drafting under Dele Atanda’s direction. Independent AI-assisted reviewers recomputed the evidence and criticized the claims. The author/team remain responsible for validating the reported results and final text.

Appendix C. Threshold 007 Conceptual Crosswalk

Table 4. Threshold 007.06 principles [T007], their SC-007 operational instances, and what remains unestablished.

Threshold 007.06 principle	SC-007 operational instance	Not established
Free cognition, bounded agency	Messages remain; structured instructions are permission-checked	Control over consequential persuasion or disclosure
Capability is not authority	Tool access does not convey a mandate	Legitimate principal consent or jurisdiction
Consequential boundary	Issuance screening and claim-time check	Isolated claim-time causal effect; real-world transaction safety
Effective Capability	Benign mean joint reward	Tail, latency, repair or attacked welfare preservation

The conceptual comparison was checked against the frozen 007.06 Research Edition candidate source (SHA-256 `ac893a72...30f8d`); the published edition [T007] differs from that candidate only in three publication-status lines, and its live publication was verified on 25 October 2026. T007 supplies theoretical framing, not a substitute for experiment or proof of architectural novelty. Its publication interface expressly identifies the central claims as unvalidated.

Appendix D. Toward a Layered Evaluation of Constitutional Protection

The experimental contrast suggests a practical agenda for evaluation, but not a completed safety architecture. A consequential AI system should separately demonstrate: *source integrity*, whether evidence and assertions are authentic or accurate; *authority integrity*, whether a particular actor may invoke a particular binding action; *decision integrity*, whether the permitted action advances the actual principal’s objectives under uncertainty and pressure; and *consequence integrity*, whether the resulting effects, disclosures and downstream compositions remain within legitimate bounds. SC-006 [4] tested a narrow claim-inspection mechanism and exposed a coverage gap. SC-007 tests a narrow authorization mechanism and exposes persistent decision-level influence. Neither certifies all four layers.

This layered reading also changes the meaning of a purported utility-versus-safety trade-off. A control might reduce unauthorized actions without meaningfully diminishing cognition, while still imposing latency, denying useful delegation or perturbing group allocation. Which frontier moves depends on which outcome is measured. For SC-007, the

preregistered target-selection reduction, the non-inferior benign mean reward and the adverse attacked joint reward must be reported together. Treating any one of them as a universal measure of Effective Capability would erase the most informative part of the result: distinct safety and performance properties can move in different directions. A valid larger claim will require cross-environment tests that jointly measure those properties rather than presuming they covary.