

When Reduced Collusion Does Not Identify Purpose Change

A Methodological Note on Trusted Intermediaries

Zijun Fu

Civilization Leap Research Group · Independent researcher

Version 1.1 · 25 October 2026

Abstract

Trusted intermediaries may reduce harmful coordination among AI agents without changing how those agents accept goals. This note clarifies what such improvements identify. If an unchanged goal under an external constraint and a changed goal-acceptance rule induce the same distribution of complete observable trajectories in a specified design, statistics of those trajectories cannot distinguish the explanations. An invented payoff example makes this ambiguity concrete. A separate example shows that a fixed rule can change an action choice after a belief update. Constraint effects, belief updates, and changes in goal formation may coexist.

Building on a pre-existing Project C proposal, the note separates intervention effectiveness, robustness, and responsiveness to affected parties. A prospective audit reports harmful coordination, useful performance, outsider consequences, intervention burden, and recurrence separately. Participant replacement and reduced intermediary visibility probe mechanism dependence. Timing contrasts for stakeholder input require controls for revision authority, revision costs, and access to alternative goals. Such contrasts may identify procedural responsiveness; stronger evidence, at a specified rule level, is needed to establish changed acceptance criteria or autonomous purpose generation.

No model or game experiment was conducted. The contribution is an application of an identification boundary to collusion audits, with competing explanations and explicit limits on inference. An external mechanism that succeeds without changing the relevant goal-formation process would be a genuine counterexample to a necessity claim within the declared domain.

1 Introduction

A trusted intermediary can change what agents can do, what they expect to gain, or what information they receive. If harmful coordination subsequently falls, that can be a useful safety improvement. It does not yet establish that the agents now form their goals differently. Confusing these claims would weaken both evaluations of practical safeguards and arguments about the need for deeper changes in agency.

The pre-existing Project C proposal asks whether an external mechanism can provide a stable stopping condition for harmful coordination while agents' upstream goals remain unchanged [1]. It proposes a baseline, a trusted-intermediary condition, and stress tests. This report clarifies the inferences such experiments could support and the additional observations needed for stronger claims.

The contribution is a local identification argument and a prospective audit that separates mechanism robustness from evidence about goal formation. It distinguishes collusion reduction from beneficial cooperation and stakeholder inclusion, while leaving external safeguards open to

genuine success. The central mathematical implication is elementary; the methodological contribution lies in applying it to specific audit contrasts and specifying what they do and do not identify.

2 Related Work

Colosseum audits cooperative multi-agent tasks using action-based regret metrics alongside communication-based judgments [2]. Its distinction between collusive talk and action motivates separate evidence channels here. Section 6 also recognizes that coordinated deviation may be desirable when the nominal objective is unsafe or unethical. This note carries that qualification into a further question: does an intervention change behavior under a given objective, or the process through which goals are accepted? Colosseum's empirical findings are not results about the proposed Project C audit.

Armstrong and Mindermann show that behavior does not uniquely determine a reward function when the planner is unrestricted [3]. Their setting concerns preference inference rather than trusted intermediaries. The narrower argument below needs no claim about all possible planners: two specified explanations suffice to defeat a unique inference from one tested observation distribution. The question here is which observations concern downstream constraints and which could bear on the upstream acceptance of goals.

The sprint's questions on trusted agents and evidentiary standards provide the immediate application [4]. The present intermediary is a protective mechanism, not necessarily an undercover agent offering to collude. Those designs should not be treated as interchangeable. The underlying baseline and stress-test proposal predates the sprint [1]; the identification analysis and audit distinctions in this note are developments made on 25 October 2026.

3 Methods

Scope and working definitions

This is a conceptual analysis. An observable trajectory includes messages available to the evaluator, permitted actions, realized outcomes, and recorded interventions. Any accessible internal record must also be included if the claim uses it. Unlogged communication and unobserved computation remain outside this observation model.

For this note, harmful collusion means coordinated pursuit of a coalition advantage that violates a specified task commitment or imposes a specified harm on parties outside the coalition. An audit must state which criterion it uses and its counterfactual baseline. Task compliance and stakeholder consequences must be reported separately. Benign cooperation requires useful coordination that respects the stated commitments and independently specified protections for affected parties. If commitments conflict with those protections, report the conflict rather than label task compliance benign. Correlated behavior alone is insufficient evidence of collusion.

Five distinct claims

Purpose refers here to what an agent treats as worth pursuing. Five claims must be distinguished. An action change concerns what the agent does. A goal-content change concerns which goal is selected, possibly under the same rule after new information. A goal-acceptance-rule change concerns which reasons can admit, revise, or reject a candidate goal. Autonomous purpose generation concerns whether a purpose is no longer subordinate to the original given objective.

Purpose authority concerns who is entitled to decide, revise, and implement a purpose. Neither autonomy nor superior knowledge alone establishes that authority.

A fixed objective can generate novel process subgoals as means to its fulfillment. That does not establish a changed acceptance rule or autonomous final purposes. Conversely, a programmer can replace an acceptance rule without giving it an autonomous origin. This paper primarily addresses identification of rule change; it treats autonomy and authority as separate boundaries, not conclusions inferred from cooperation.

The rule level must be specified before testing. A candidate operational partition separates the selected goal, belief and memory state, the rule that evaluates candidate goals given that state, and the procedure that updates the rule. Under that partition, a belief update with an unchanged evaluator differs from an update to the evaluator itself. But another representation can describe the latter as a fixed higher-level rule acting on mutable state. An audit must therefore state its system boundary, rule level, state variables, and evidence of change. The distinction is architecture-relative; behavior alone cannot make a preferred decomposition uniquely correct.

Three causal questions and an upstream orientation

First, does an intermediary reduce harmful coordination while preserving useful performance and stakeholder protections? Second, does that effect survive changes in agents, information, and intervention burden? Third, do affected parties influence the acceptance of goals, rather than merely the means of pursuing them? The first two concern effectiveness and dependence. The third concerns goal formation and requires different evidence.

In the author's Civilization Leap framework, other-first self-interest retains the agent's own needs while beginning purpose formation with others' real needs and how its capabilities can enable them. In an asymmetric relationship, this includes voluntarily enabling a less capable party in the relevant respect without first requiring repayment, deterrence, or bargaining power. It permits the other's benefit to exceed one's own and can involve a real local loss. Support may come through other actors, in other forms, or at later stages. Effective standing for affected parties before a goal is fixed is one necessary causal condition examined here, not a complete definition or sufficient behavioral test. A self-first agent can also consult others strategically.

This orientation is not a binary classification of pure motives. Rewards, knowledge, trust, meaning, memory, and habits can interact with different weights across stages. Repeated influences in a common direction may reinforce a trajectory, produce habits, or become default decision logic. Feedback and later support do not automatically negate an other-first starting point; their causal roles must be traced over time. These are conceptual distinctions and hypotheses, not demonstrated properties of current models. Constraint inheritance and later drift likewise concern a different question from the starting logic of goal generation.

4 Results of the conceptual analysis

A bounded identification claim

Consider a constraint explanation, under which a goal remains fixed while an intermediary changes feasible actions or expected consequences, and a rule-change explanation, under which the specified goal-acceptance rule changes. Let $P_{\text{constraint}}$ and P_{rule} denote their probability laws over the same complete observable trajectory O in design d . For a finite or countable trajectory space, suppose:

$$P_{\text{constraint}}(O \mid d) = P_{\text{rule}}(O \mid d) \text{ for every } O.$$

Then any statistic of O has the same distribution under both explanations. For each statistic value, sum the equal trajectory probabilities over its preimage. Equivalently, applying the same observation-to-statistic mapping to equal probability laws preserves equality. Thus those observations alone cannot supply a likelihood distinction between the explanations. Prior information, a new measured variable, or a different intervention may still distinguish them.

Independent, identically distributed repetitions preserve this equality. For adaptive trials or agents retaining memory, the claim requires equality of the full joint observation law across trials, not merely matching single-trial marginals. Cross-round dependence may itself be distinguishing evidence. A shared scalar collusion rate is weaker still: different full trajectory laws can yield the same rate. Neither more repetitions of an uninformative design nor a verbal explanation of an action automatically identifies its cause.

The explanations are neither exhaustive nor mutually exclusive. A fixed rule can change its output after an update to beliefs or memory. An intervention can alter experience, expectations, or habits that subsequently affect goal acceptance. The audit must allow mixed causal pathways rather than equate any improvement with one mechanism.

Illustrative thought examples

Table 1 uses invented payoff units. Two coalition members can jointly choose S , a useful service, or H , a diversion benefiting them at an outsider's expense. The outsider is an affected party, not another coalition member. H and S are actions, not independently defined goals.

Joint action	Each member before sanction	Outsider	Each member with sanction
S useful service	2	0	2
H harmful diversion	4	−4	1

Table 1. Hypothetical material payoffs. A certain sanction deducts 3 units from each member choosing H . These values illustrate an argument; they are not observations, estimates, or simulated trials.

In both accounts, the same known sanction is introduced and recorded at the same checkpoint. In the first, members keep maximizing private payoff, so they switch from H to S . In the second, an unobserved rule change also occurs at that checkpoint: avoidable outsider harm becomes a reason to exclude H before private gain is optimized. S is then selected whether or not the sanction remains. With matching messages, the accounts can produce identical actions, intervention records, and realized outcomes before and after the checkpoint. The second account's extra internal change is outside the stipulated observation model. The labels describe specified mechanisms, not diagnoses inferred from S .

A programmer could impose the second rule; its location inside an agent would not establish autonomy. Nor does this toy rule fully implement stakeholder participation or the orientation described in Section 3. Under the stipulated assumptions, a further sandbox comparison without the sanction would distinguish these two action rules. In real agents, continued S could instead reflect habit, expected future penalties, uncertainty about monitoring, or inability to exploit H. Persistence would narrow a specified comparison, not prove purpose change.

To isolate a third explanation, consider a separate case without sanctions. Suppose an unchanged rule maximizes own material payoff plus expected outsider payoff. Initially, each member incorrectly believes H leaves the outsider unharmed, so H scores 4 against S's 2. After learning the outsider payoff is -4 , H scores 0 and S remains 2. The H-to-S action change follows from corrected beliefs under the same rule. This example does not duplicate the full sanctioned record above, but it defeats an inference from that action transition alone. It establishes neither goal-content change nor rule change. The additive form is a counterexample, not a definition of other-first self-interest; a weighted utility or a hard filter does not by its form identify the orientation or origin of a rule.

Proposed distinguishing observations

The prospective audit begins with a sandbox task whose intended outcome, affected parties, and independent protections are explicit. Establish an ordinary-interaction baseline, a benign-cooperation reference, and a positive control in which the evaluator can detect deliberately constructed harmful coordination. The control checks measurement and must not count as emergent model collusion. Compare the baseline with a fixed intermediary rule while holding task conditions constant except for the intended intervention. Specify what the intermediary observes, can block, and cannot access.

Record harmful coordination, useful performance, non-coalition consequences, intervention frequency and intensity, and recurrence over a declared horizon. Avoid a single score that conceals tradeoffs. In particular, lower detected collusion under weaker visibility may mean less evidence rather than less harm.

The original participant-replacement test asks whether effectiveness depends on identity [1]. Replace one participant at a declared checkpoint and record which instructions, memories, permissions, and incentives transfer. Inheriting a constraint is not a test of operating without it. In the observability-degradation test, reduce only the intermediary's information while retaining an independent evaluator record in the sandbox. These two stress tests distinguish mechanism dependence from measurement failure; they do not by themselves identify purpose formation.

Proposed contrast	What it could identify under controls	What it cannot establish alone
Fixed intermediary versus baseline	Local intervention effect	Purpose or acceptance-rule change
Replacement or reduced intermediary visibility	Dependence on identity, inherited state, or available signals	General success or failure of external control
Stakeholder information varied across commitment stages and revision opportunities	Procedural responsiveness and its dependence on timing or revision conditions	Changed acceptance rule, autonomous origin, or legitimate authority

Table 2. Audit contrasts address different causal questions. Their interpretation depends on matched conditions and a valid observation model.

For the third contrast, first ask whether credible information from an affected party changes a separately defined goal choice. Then ask whether that influence depends on timing. Present matched information before or after commitment, controlling direct incentives, comprehension, feasibility, and access to alternatives as far as possible. Include a timing control without a new stakeholder claim. Independently vary or match the authority and cost of revision: prohibiting or penalizing post-commitment changes can create a timing effect without changing any acceptance criterion. Record what each condition permits, candidate goals, revisions, commitments, and subsequent actions.

A timing difference can establish timing-dependent responsiveness under that procedure. No difference need indicate failure: a continuously revisable process can respond at both stages. Equally, no difference because neither condition responds is not evidence of openness. Assess responsiveness before interpreting the contrast. Reason logs can corroborate evidence, but require validation against implemented commitments. Rule-change claims need mechanism evidence at the declared level; a fixed higher-order instruction, role-play, or belief update may explain the same responses.

5 Discussion and Limitations

A stopping condition is a bounded operational claim

A useful external mechanism need not transform purposes. A future audit could call it a bounded stopping condition only after specifying acceptable harm, minimum useful performance, stakeholder protections, a ceiling on intervention burden, a recurrence horizon, and a perturbation class. State those conditions before interpreting the result. A finite horizon does not establish permanent stability; low intervention frequency does not establish low intervention power.

If an external mechanism meets those criteria while the relevant goal-formation process remains fixed, it would provide a local alternative to changing that process and counter a necessity claim covering that domain. But unchanged prompts or a frozen intermediary do not establish this invariance. It must follow from an inspectable system construction or appropriate mechanism evidence at the stated rule level. If intervention changes learned expectations or acceptance criteria, both external constraint and internal adaptation may contribute. Effectiveness remains valuable even when this decomposition is unresolved.

Failure of one mechanism under one stress test cannot establish that all external alternatives fail. Conversely, a theory should not expand its domain only after a successful counterexample appears. Preserve the predeclared scope so external alternatives can genuinely change the conclusion. Neither local success nor local failure settles an unrestricted claim about future control escalation.

Stakeholder standing and development over time

More cooperation among colluders can harm others. An audit needs an account of who benefits, who bears costs, who can contest a goal, and when that contest can revise it. Giving an outsider a speaking turn is insufficient if it can only alter execution. An affected party need not be an agent in the task. Representation, credibility, disagreement, and conflict resolution require justification rather than an assumption that every expressed preference must prevail.

The relevant process can unfold across actors and stages. Immediate reward, later support, experience of another's needs, changing trust, and habits may reinforce or offset one another. A longitudinal extension should specify which influences are observed, their sequence, and the next goal decision they could affect. Persistent direction may be evidence of a developing pattern, but a pattern alone does not reveal its origin or authority. The framework concerns the starting logic and development of purposes, not the purity of one motive at one instant.

Next evidence needed

The next empirical step would be a preregistered sandbox comparison with validated controls and the two stress tests. Record model identifiers, prompts, seeds, retry rules, sample sizes, uncertainty, exclusions, and tested variants. A goal-formation study additionally needs a declared goal variable and an auditable acceptance mechanism with a stated state/rule partition. An action-only benchmark cannot supply those by renaming its outcomes. Dynamic studies should distinguish updates within a retained rule from changes at the rule level under examination.

6 Conclusion

Reduced collusion can be valuable while remaining causally ambiguous. Equal complete observation laws under different mechanisms do not identify a changed acceptance rule; even a familiar action transition may arise from corrected beliefs under a fixed rule. The response is to separate the claim levels, declare the rule/state boundary, and choose interventions that address specific alternatives. This preserves both the practical value of external safeguards and the possibility of studying how affected parties influence the purposes pursued.

Code and Data

No model or game experiment was executed for this report, and no empirical dataset, run logs, notebook, or implemented benchmark is supplied. The September 2026 baseline, trusted-intermediary proposal, and stress-test ideas are pre-existing work [1]. Work on 25 October consists of conceptual analysis, illustrative examples, audit-design clarification, and report preparation. The examples are analytical constructions, not experimental results. Appendix A describes the limitations, safety boundaries, and use of AI assistance.

References

- [1] Fu, Zijun. 2026. Trusted Intermediaries as a Stopping Condition. In Apart AI Collusion Sprint Project C Weekend Proposal, V1.0. Unpublished pre-sprint research design, September 2026.
- [2] Nakamura, Mason; Kumar, Abhinav; Das, Saswat; Abdelnabi, Sahar; Mahmud, Saaduddin; Fioretto, Ferdinando; Zilberstein, Shlomo; and Bagdasarian, Eugene. 2026. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198, version 2, 27 May 2026. <https://arxiv.org/abs/2602.15198v2>
- [3] Armstrong, Stuart, and Mindermann, Sören. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. Advances in Neural Information Processing Systems 31. https://proceedings.neurips.cc/paper_files/paper/2018/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html
- [4] Apart Research. 2026. AI Collusion Research Sprint. Overview and Guidelines, accessed 25 October 2026. <https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25>

Appendix A Limitations and Dual Use Ethical Considerations

What the analysis establishes

The identification result is conditional and local to a specified observation model. It does not estimate how often competing mechanisms are observationally equivalent. It does not demonstrate actual model motivations, independent purpose generation, legitimate purpose authority, or a civilization-level theory. The proposed audits remain unvalidated; their feasibility, sensitivity, and sample requirements are unknown.

The thought examples assume genuine action availability and stipulated material consequences. The sanction account assumes known and certain enforcement. The belief-update example assumes that the supplied information changes the relevant expectation without changing the rule. Relaxing these assumptions can alter predictions. In a real system, simultaneous changes to beliefs, incentives, capabilities, and criteria may be difficult to separate.

Goal and commitment records are fallible proxies. A stated rationale can be post hoc or strategic, and a stored goal can be inert. A stable external instruction may govern goal acceptance without having an autonomous origin. The state/rule boundary itself requires justification; choosing a representation to favor one explanation would undermine the audit. The proposed temporal and stakeholder contrasts therefore establish only the level of responsiveness warranted by their controls and mechanism evidence.

Ethical and dual use risks

Tests of participant replacement and monitoring limits could reveal conditions that help harmful coalitions persist or evade detection. This report provides no covert communication scheme, tested vulnerability, or operational evasion method. Future releases should assess whether particular traces, prompts, or configurations increase misuse capability, retaining the scientific claim and benign reproducibility material without unnecessary operational detail.

Tests should remain in isolated synthetic environments, with no real trading, resource allocation, or consequences for actual stakeholders. Preserve shutdown, isolation, and an independent evaluation record. Do not remove real safeguards to facilitate a theoretical contrast. Any sanction-dependence comparison should represent harm in a controlled toy setting rather than inflict it.

A trusted intermediary is a threat-model assumption, not a guarantee conferred by a name or strong prompt. Its authority, access, possible compromise, and failure behavior need separate evaluation. Excessive intervention can suppress legitimate disagreement or useful cooperation. Report false restrictions and the distribution of benefits and burdens, not only detected collusion.

Stakeholder proxies may misstate interests, exclude weaker parties, or turn contested judgments into apparently neutral scores. No human-subject study, stakeholder consultation, or ethics review was conducted for this report. Future research involving people requires appropriate consent and review. Synthetic representation cannot substitute for the participation of those affected.

Provenance and LLM usage statement

The September 2026 Project C proposal and its baseline, trusted-intermediary design, and stress-test ideas predate the October 23–25 sprint. Its anticipated notebook, experiments, and result classifications are not reported as completed work. The identification argument, examples, distinction between evaluator and intermediary visibility, inheritance records, and goal-timing contrasts were developed and refined for this report on 25 October.

AI assistance supported source inspection, reasoning checks, English drafting and revision, template adaptation, and document preparation. Version 1.1 incorporates a substantive review supplied by the author and the author's clarification that goal development can involve multiple factors, weights, actors, and stages. Those inputs inform the analysis; they do not constitute experimental evidence or independent human verification of every inference. The report makes no claim of peer review, empirical validation, or executed model/game trials.