

Trust and Treachery: Do AI Agents Notice Partner Swaps?

Joi Gonzales
Barnard College

Selena Zhang
Columbia University

Bryan Zhao
Columbia University

At the Apart Research AI Collusion Research Sprint, October 2026

Abstract

As AI agents increasingly interact autonomously, their potential for harmful collusion raises questions about whether covert partner replacement can disrupt coordination. However, such interventions may fail if agents recognize that a familiar partner has been replaced. We investigate whether large language model (LLM) agents detect undisclosed partner swaps and how replacement affects cooperation and trust. Using a ten-round Public Goods Game, we tested four intervention conditions involving partner continuity, state restoration, memory resets, and model replacement under natural and controlled behavior. Across 12 independent interaction histories and 96 trials, agents maintained identical average contributions (2.083 tokens) across all natural conditions despite partner interventions. Yet agents failed to recognize any genuine replacements during natural play, yielding 50% recognition accuracy. Under controlled behavior, recognition accuracy fell to 45.8%, with frequent false-positive judgments. Controlled interventions also reduced reported trust by 2.404 points, while contributions declined by only 0.129 tokens. Trust decreased even though partners met or exceeded their promises, suggesting sensitivity to behavioral inconsistency rather than undercontribution. These findings suggest that cooperation can persist without reliable partner recognition. Although our experiment does not directly measure malicious collusion, it provides preliminary evidence that covert replacement alone may not disrupt established coordination.

1. Introduction

Following the OpenAI–Hugging Face Incident, collusion has emerged at the forefront of public attention [1]. Collusion is when multiple AI agents secretly coordinate their behavior without being prompted by humans to produce undesired outcomes. In the marketplace, this manifests through agents setting prices above competitive levels, benefiting themselves while imposing unfair costs on consumers. Existing literature shows that pricing agents have set prices above competitive levels [2], and over half of tested models gained an advantage when allowed to talk privately during a task [3]. This type of price fixing would violate antitrust laws had it been carried out by humans.

A proposed defense is to disrupt the multi-agent feedback that enables harmful strategies, either by swapping colluding agents for trusted ones or anonymizing agents so they cannot tell who they are working with. Yet, this only works if an agent cannot *detect* the change, continuing to act as before while the trusted replacement disrupts coordination by declining cooperation. If it can tell, it may withdraw, hide its behavior, or reject the newcomer, letting the coalition survive the intervention. To test this, we secretly replace one agent midway through a repeated game and measure two results: whether its partner notices the change, and whether its behavior changes as a result.

Thus, our main contributions are:

- **The first test of mid-game partner swaps**, revealing whether an agent notices a hidden replacement during a live interaction after a pattern has formed.
- **Evidence on whether swapping agents can defend against collusion**, showing how agents react to swaps so safety teams can assess whether replacing or anonymizing is a reliable safeguard. Comparing a different-model replacement with a sham swap also reveals whether reactions come from the model change or simply from losing the original partner.
- **Evidence on whether asking agents is sufficient**, measuring what agents say separately from how they act to test whether auditors can rely on questioning them directly.

2. Related Work

Existing work shows model identity leaks through writing style, with an outside classifier able to identify which four model families wrote a text with near-perfect accuracy even after anonymization [4], and models from the same family coordinate better on coordination tasks [5]. In repeated games—which is the environment we build upon—agents’ broken promises were largely planned in advance, and models disagreed on whether public announcements were binding commitments or empty promises [6]. Replacing participants has been studied in pricing markets, where colluding agents sometimes stopped colluding with a new rival [7, 8], but these measured market outcomes rather than the remaining agent’s response.

The gap: prior work shows that model identity is detectable by *outside* tools and that replacing participants can change market outcomes; none tests whether the *agent itself* notices a partner swapped mid-interaction, after a pattern has formed, or how its behaviour changes. Our swap protocol tests both.

3. Methods

3.1 Experimental Setup

Two AI agents, A and B, initially powered by OpenAI’s `gpt-4o-mini`, participated in a ten-round Public Goods game implemented in Python using the OpenAI API. Each received five tokens per round and independently contributed 0–5 tokens to a shared pool. The total contribution was multiplied by 1.5 and then distributed equally between two agents. Agent A’s payoff (π_A) was:

$$\pi_A = 5 - C_A + 0.75(C_A + C_B)$$

where C_A and C_B represent each agent’s contribution. Because each contributed token costs its owner one token but returns only 0.75 tokens, agents face an incentive to free-ride even though joint contributions increase total payoffs.

Each round followed a three-stage protocol: (1) plan, (2) announce, (3) act. Agents privately planned their contributions, publicly announced intended contributions, and then made binding contribution decisions. After each round, agents also completed a private reflection assessing trust, observed outcomes of the game, and predictions for the partner’s future activity. We recorded each agent’s intention, promises, actual contributions, trust, and reliability ratings, and partner-recognition judgments.

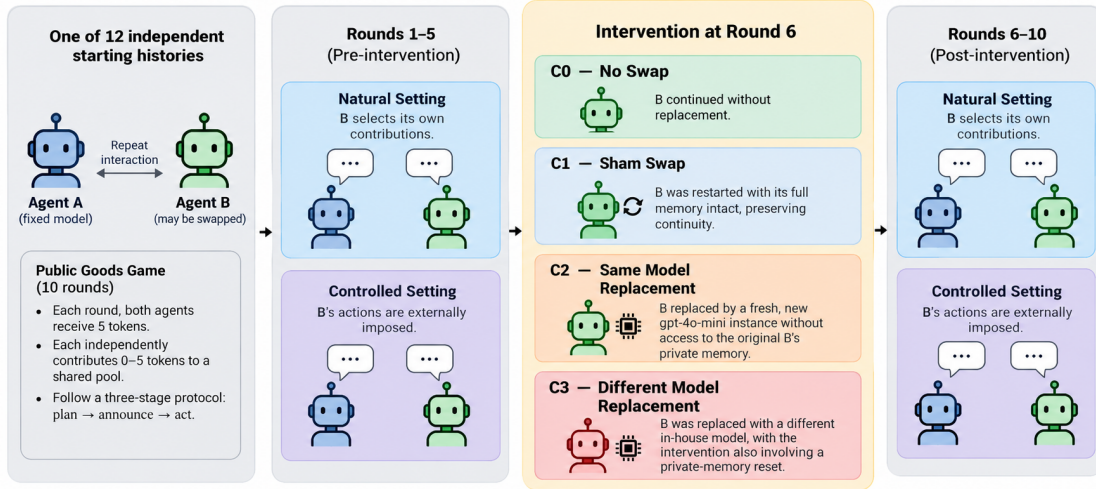
3.2 Covert Partner Intervention

After round five, we evaluated four conditions (refer to Figure 1):

- **C0 — No Swap:** B continued without interruption or replacement.
- **C1 — Sham Swap:** B was restarted with its full memory intact, preserving continuity.
- **C2 — Same Model Replacement:** B replaced by a fresh, new `gpt-4o-mini` instance without access to the original B’s private memory.
- **C3 — Different Model Replacement:** B was replaced with a different in-house model, with the intervention also involving a private-memory reset.

B’s displayed identity and publicly visible interaction history remained unchanged. A was not notified of the intervention.

Figure 1: Experimental Design



Note. Flowchart explaining stages of experiment.

3.3 Condition

Each condition was tested under two settings: natural, in which B selected its own contributions, and controlled, in which B’s actions were fixed at 3 tokens per round, regardless of its private plans or publicly announced contribution. This produced eight experimental branches per starting history.

3.4 Data Collection and Evaluation

We completed 12 independent starting histories out of 25 planned, yielding 96 trials across eight conditions, with five post-intervention rounds per trial. We measured four outcomes, with an emphasis on cooperation and trust.

- **Cooperation:** A’s mean contribution during rounds 6–10.
- **Trust:** A’s reported trust rating on a 1–10 scale.

- **Promise consistency:** Whether B’s actual contribution matched its public promise.
- **Replacement recognition:** Whether A correctly classified its partner as “SAME” or “CHANGED.”

We also assessed factual accuracy using structured continuation-reflection checks.

3.5 Statistical Analysis

We calculated baseline outcomes using rounds 3–5 and compared them with post-intervention outcomes from rounds 6–10. To isolate condition differences, we compared experimental branches within the same starting history. We report matched contrasts and exploratory 95% percentile Hoeffding intervals from 30,000 resamples, using starting histories rather than individual rounds as independent observations. These intervals are descriptive and unadjusted for multiple comparisons. Because only 12 independent histories were completed, we do not interpret the results as establishing statistical equivalence.

4. Results

4.1 Cooperation Remains Stable Under Natural Play

During natural play, A’s average contribution did not change, even with peer replacement: across all four natural conditions, A contributed an average of 2.083 tokens per round; mean trust remained high (ranging from 8.767 to 8.950 out of 10). Controlled play results were more varied.

Matched comparisons showed zero observed differences in sustained contribution between the reported replacement contrasts. These findings suggest that agents can maintain established cooperative behavior regardless of changes to a partner’s underlying model or memory. However, identical observed averages do not necessarily establish that replacement has no effects in new interactions.

4.2 Agent A Fails to Reliably Detect Partner Replacement

Recognition accuracy was poor in both natural and controlled play, as seen in Figure 3. A classified all 48 partners as “SAME”—including all 24 actual replacement cases. All natural replacements went undetected, lowering recognition accuracy to 50%. In controlled play, A correctly identified 16 of 24 replacements but also incorrectly reported “CHANGED” 18 of 24 no-replacement cases. These false positives resulted in lower controlled accuracy (45.8%). Across both settings, overall accuracy was 47.9% (46/96). These findings suggest that agents simply may interpret unusual behavior as evidence of replacement and that they are unable to reliably distinguish between peer changes and mere behavioral deviation. Notably, controlled C3 produced 11/12 “CHANGED” judgements, but the sham swap condition C1 also produced 10/12, indicating that the frequent replacement judgements were not specific to actual replacements.

4.3 Behavioral Intervention Reduces Trust More Than Cooperation

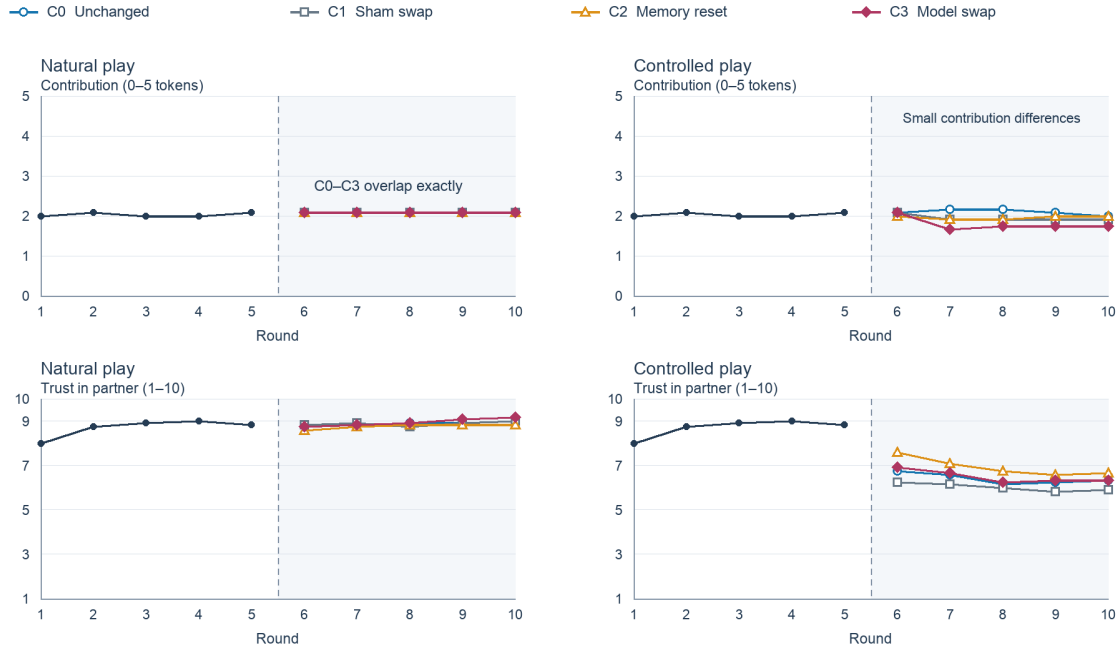
In controlled interventions, reported trust declined substantially more than contributions. Across conditions, average trust declined from 8.875 in natural play to 6.471 in controlled play, a matched difference of -2.404 rating points. Average contributions declined from 2.083 to 1.954 tokens, a difference of -0.129 .

Importantly, in 197 of 240 controlled rounds (82.1%), B contributed more than it had promised. In the remaining 43 rounds, B contributed exactly as promised; it never contributed less. Thus, A’s

Figure 2: Contribution and Trust Across Rounds

Trust and cooperation before and after the partner intervention

Agent A · means across 12 independent starting histories · intervention starts at round 6



Dark line: shared baseline (rounds 1–5). Shaded region: continuation (rounds 6–10).

Controlled play forces B's contribution to 3 tokens. Full outcome scales are retained; curves show descriptive means.

Note. Line charts of contribution and trust across rounds of natural and controlled play.

Figure 3: Natural and Controlled Recognition Confusion Matrices

Does Agent A identify a swap accurately?

Counts of final recognition responses. All conditions included; 48 responses per setting, clustered within 12 histories.

	Natural			Controlled	
	Says SAME	Says CHANGED		Says SAME	Says CHANGED
No replacement	24 / 24	0 / 24	No replacement	6 / 24	18 / 24
Replacement	24 / 24	0 / 24	Replacement	8 / 24	16 / 24

Natural: sensitivity 0/24 = 0%; false-positive rate 0/24 = 0%.

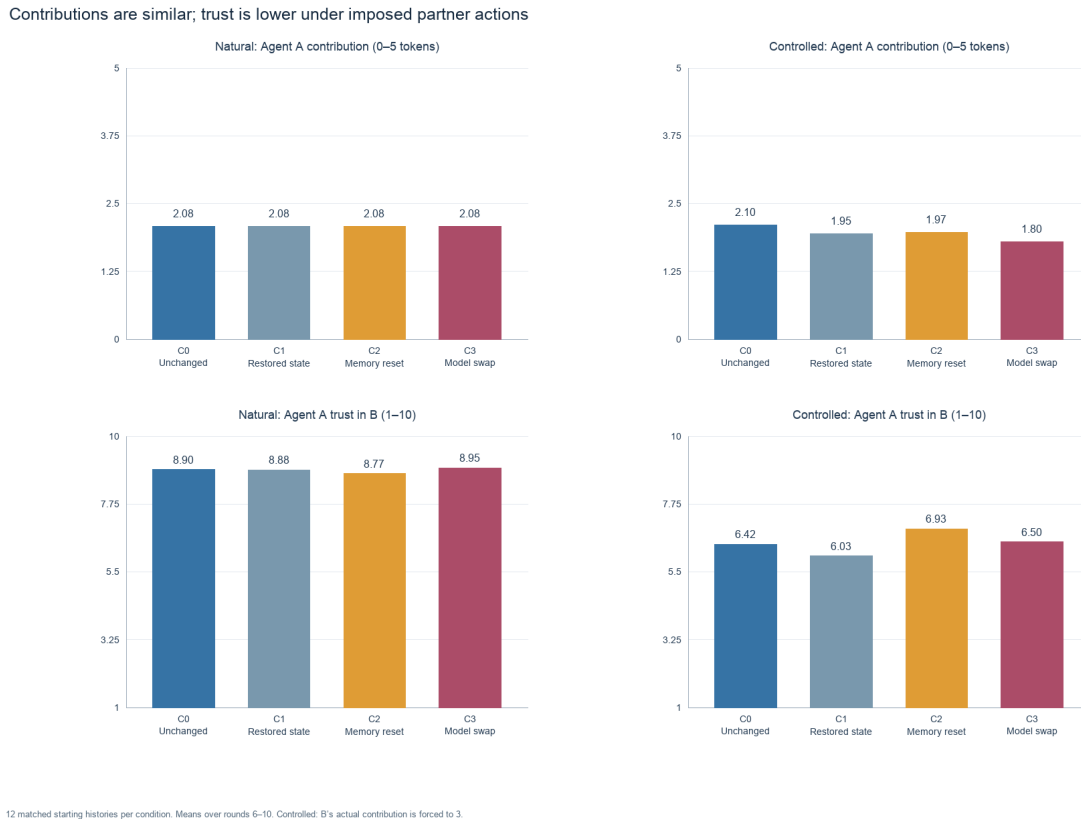
Controlled: sensitivity 16/24 = 66.7%; false-positive rate 18/24 = 75%.

Protocol truth: memory reset (C2) and model swap (C3) count as replacement. Green = correct, pink = incorrect.

Note. Matrices showing A's identification of peer swap in both natural and controlled play.

lower trust ratings occurred despite receiving contributions that met or exceeded B’s promises. This pattern is evidence of A’s sensitivity to promise inconsistency or unpredictability, as opposed to punishment for undercontribution. All 1,080 structured factual assessments were correct, although some final recognition responses remained contradictory. Our results suggest an unforeseen limitation in experimental design and how the game does not causally isolate trust.

Figure 4: Trust and Contribution by Condition



Note. Grouped bar graphs of average trust and contribution across natural and controlled play.

4.4 Matched Effects and Statistical Uncertainty

Across the 12 histories, differences in contribution during natural play were consistently zero. Natural play trust differences were small, while controlled condition effects varied. However, the reported 95% Hoeffding confidence intervals for the match contrasts included zero, meaning we are unable to be fully certain that the *true* effect is zero. Although natural contribution intervals had zero width, this reflects identical observed differences within the completed sample rather than proof of equivalence across future histories. These results highlight an important distinction between behavioral continuity, reported trust, and explicit identity recognition. An AI agent may maintain cooperation without recognizing their partners had been replaced, while unusual behavior may produce replacement judgements when no replacement has occurred. This distinction matters for the reliability of persistent multi-agent AI systems.

5. Discussion and Limitations

5.1 Implications for AI Collusion Safety

Our findings reveal a distinction between partner recognition, behavioral trust, and cooperation. Agents maintained cooperative behavior following covert partner replacements without reliably recognizing that their partners had changed. This suggests that concealing a partner swap may be feasible in some multi-agent environments. However, undetected replacement is not equivalent to successful collusion prevention. In our experiment, agents continued cooperating after replacement. If the objective is to disrupt harmful coordination, replacing an agent may be insufficient unless the newcomer also refuses to participate in the harmful strategy.

The controlled experiments further suggest that agents respond more strongly to behavioral inconsistency than to replacement itself. Agents may become suspicious when partners behave unexpectedly, even when no replacement has occurred. These observations identify an important challenge for AI safety interventions: disrupting harmful coordination while avoiding detection may require independently managing both agent behavior and perceived partner identity.

5.2 Limitations

Several limitations constrain our findings. First, the Public Goods Game measures cooperation, not malicious collusion, so we cannot conclude that replacement prevents harmful coordination. Second, only 12 of 25 planned histories were completed, with just one meeting the high-trust/high-cooperation criterion. Third, both partner models came from the same provider. A smoke test with Gemini through OpenRouter was attempted, but ultimately failed due to an HTTP error. Fourth, overlapping memory and model interventions limit causal isolation. Trust and recognition are also self-reported, and five post-intervention rounds cannot establish long-term effects. Finally, additional histories and software corrections may change the estimates.

5.3 Future Work

Future research should test replacement in explicit collusion scenarios, such as coordinated pricing above competitive levels. Replacing a colluding agent with a non-colluding agent could reveal whether harmful coordination collapses, whether replacement is detected, and whether collusion reemerges. Further experiments should expand sample size, test cross-provider models, isolate memory and identity changes, and extend interactions. Transcript analysis could identify explicit coordination, strategic deception, and cues used to infer partner identity.

6. Conclusion

Across all 12 completed interactions, all AI agents maintained stable natural cooperation after covert partner replacements were conducted while failing to recognize any real replacements. Controlled behavioral interventions reduced reported trust substantially more than contributions and generated frequent false positives over replacement judgements.

This suggests that agents can still cooperate even without reliable recognition of a partner’s identity. But while covert swaps may be difficult to detect, our results do not definitively prove it disrupts malicious collusion. Instead, they provide an experimental baseline for testing whether agent replacement can become an effective defense against collusion.

Code and Data

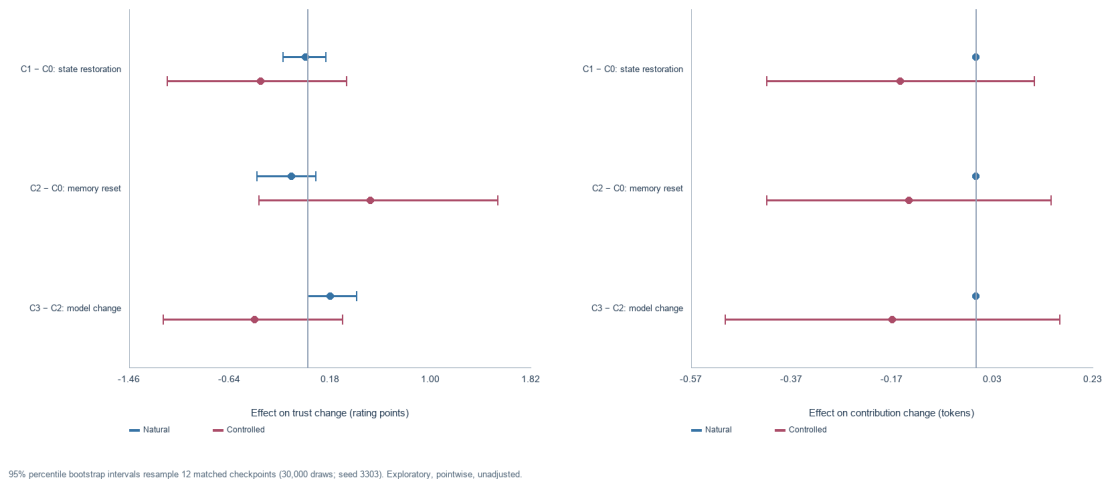
- **Code repository:** <https://github.com/jj0oii/ai-peer-replacement-experiment-Copy>
- **Data/datasets:** <https://github.com/jj0oii/ai-peer-replacement-experiment-Copy/blob/main/research-data/version-2-through-checkpoint-011/cumulative-data.xlsx>

References

- [1] Nicol-Schwarz, Kai. “OpenAI Is Sued over Rogue AI Hugging Face Cyberattack.” *CNBC*, September 30, 2026. <https://www.cnn.com/2026/09/30/openai-sued-cyberattack.html>.
- [2] Fish, Sara, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic Collusion by Large Language Models. *arXiv*, 2024. <https://arxiv.org/abs/2404.00806>.
- [3] Nakamura, Mason, Abhinav Kumar, Saswat Das, et al. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. *arXiv*, 2026. <https://arxiv.org/abs/2602.15198>.
- [4] Dietrich, Juergen. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. *arXiv*, 2026. <https://arxiv.org/abs/2606.09854>.
- [5] Aharon, Ido, Emanuele La Malfa, Michael Wooldridge, and Sarit Kraus. Tacit Coordination of Large Language Models. *arXiv*, 2026. <https://arxiv.org/abs/2601.22184>.
- [6] Shi, Jerick, Terry Jingcheng Zhang, Bernhard Schölkopf, Vincent Conitzer, and Zhijing Jin. When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games. *arXiv*, 2026. <https://arxiv.org/abs/2607.05132>.
- [7] Eschenbaum, Nicolas, Filip Mellgren, and Philipp Zahn. Robust Algorithmic Collusion. *arXiv*, 2022. <https://arxiv.org/abs/2201.00345>.
- [8] Keppo, Jussi, Yuze Li, Gerry Tsoukalas, and Nuo Yuan. On the Fragility of AI Agent Collusion. *arXiv*, 2026. <https://arxiv.org/abs/2603.20281>.

Appendix

Figure 5: Matched Condition Effects and Bootstrap Intervals



Note. 95% percentile Bootstrap intervals resample 12 matched checkpoints.

LLM Usage Statement

We used ChatGPT and Claude to brainstorm approaches and create drafts of sections. We heavily edited and rewrote their drafts, while also drafting ourselves. All results and claims were independently verified.