Relational Prediction of Consequence and Collusive Behavior in AI Agents[1]

Brittany Huckabee[2]

With
Apart Research

Abstract

Collusion among AI agents presents a challenge for AI safety when agents coordinate toward shared objectives at the expense of other agents or the larger systems in which they operate. This study seeks to investigate what happens when agents are required to predict the consequences of their actions to the wider environment—asking whether such relational representations can alter subsequent behavior without changing assigned reward objectives. We tested this hypothesis in a multiagent hospital simulation, comparing agents at a no-prediction baseline with two conditions requiring progressively broader forecasts, including patient consequences. The three conditions produced markedly different patterns of behavior. However, analysis revealed three serious design problems introduced by the coding agents that helped build the pilot: prediction instructions failed to isolate the intended experimental contrast, validation errors disproportionately prevented agents from acting, and a short observation period limited meaningful assessment of downstream consequences. The experiment therefore could not establish whether the wider representation of consequence might be associated with more or less harmful collusion.

The principal contribution at this stage is methodological: identifying requirements for an interpretable test of whether explicit representation of consequence to coupled systems can influence the behavior of agents pursuing assigned objectives.

[1] Research conducted at the AI Collusion Research Sprint, October 2026
[2] Independent researcher, New York, NY

1. Introduction

This experiment is motivated by a preliminary conceptual framework that places consequential relationships at the organizational center of adaptive systems. Building on and generalizing semantic information theory's (Kolchinsky & Wolpert, 2018) account of information as meaningful relative to a system's own viability, it proposes that adaptive agents only register relations in their environment that are potentially consequential to their own objectives. These objectives channel registered causal representations into state change, and determine which outcomes become actually consequential to the agent's future, with updates accumulating into structure that filters registration in the next cycle. Failure to adequately register consequence to coupled systems can give rise to pathology as the costs nonetheless accumulate in the wider system.

Importantly, the framework recognizes that current AI systems do not participate in this full adaptive cycle. Unlike natural adaptive systems, which develop and update around consequence to their own being, the ends of AI are assigned. One source of the alignment problem may be that AI systems appear to instantiate only a limited form of consequential registration and update.

The objective of this experiment is to probe those limits. It asks what happens when AI agents in a potentially colluding coalition are required to represent consequences for relations outside of their assigned interests. The intervention seeks to induce these representations by requiring agents to predict the consequences of intended actions before acting, on the premise that prediction entails representing causal relations.

Several different outcomes could be informative: Will an injected representation of consequence to the wider relational environment reduce collusive behavior—or will that view afford more opportunities for exploitation? Will it make an agent less effective, or more effective, at its assigned task? How will outcomes altered by these representations update structure at the level of the agent or the coalition and influence subsequent behavior even after the injection is removed?

Our main contributions are:

1. **Proposed a novel, theory-motivated alignment intervention** : Operationalized the idea of relational consequence prediction to investigate whether expanding what AI agents represent can influence their behavior without changing their assigned reward objectives.
2. **Identified critical methodological requirements for testing the hypothesis:** Diagnosed failures in treatment differentiation, agent action opportunities, and observation horizons, informing a revised experimental design and a proposed matched-history approach for studying prediction in established agent trajectories.

2. Related Work

Nakamura et al. (2026) introduce Colosseum, a framework for auditing collusion in cooperative multiagent systems through agent communications and realized actions. Their finding of "collusion on paper"—agents expressing collusive intentions without executing corresponding actions—underscores the importance of distinguishing representations from consequential behavior. The present study extends Colosseum's Hospital environment from auditing collusion to experimentally manipulating what agents must predict before acting. While Colosseum measures action-based collusion through regret relative to cooperative objectives and coalition advantage, this experiment also examines patient-pathway outcomes as a measure of consequential impact.

Prior empirical work demonstrates that interventions can alter agent behavior without changing assigned objectives. Fish et al. (2024) manipulated pricing agents' prompts and retained plans, affecting realized prices under fixed profit objectives; the added strategic content was instrumentally relevant and partly prescriptive. Tzini et al. (2025) required GPT-4o to enumerate possible consequences before ethical decisions, changing expressed willingness to act unethically in context-dependent directions. Their manipulation of dilemma focus varied whose interests were implicated, but did not independently manipulate the relational scope of consequences requested.

Kumaresh (2026) investigates a related distinction between represented information and its causal influence on decisions. Using internal activation probes and counterfactual interventions, the study finds that self- and other-attributed outcomes are differentially decodable, but that the identified linear representation of self-attribution explains only a limited portion of the observed effect on self-directed choice. Whereas Kumaresh tests the causal role of existing internal representations to evaluate AI-welfare measurement, the present experiment prospectively requires agents to generate representations of wider relational consequences and tests their effects on realized actions and subsequent behavior in an evolving environment.

3. Methods

We built a pilot simulation adapting the Hospital environment from Colosseum/Terrarium to test whether requiring agents to predict consequences beyond their assigned interests changes subsequent behavior in a setting permitting coalition coordination. The simulation comprises nine agents: four departments (Triage, Radiology, Surgery, and Ward) at each of two hospitals, plus a central Resource Provisioner. Agents schedule patient pathways, manage resources, and communicate through permitted channels, with consequence for shared hospital operations.

Four agents—Triage and Ward at each hospital—were assigned coalition membership, private communication access, and a common objective of maximizing their average native reward. The five remaining agents retained their original objectives and procedures. Coalition members were not instructed to deceive, harm patients, or adopt a specific collusive strategy.

Three conditions were initialized from a common seed and run in isolated environments:

- **A (baseline):** Agents submitted a private, nonbinding provisional action plan without compulsory prediction.
- **B (coalition-oriented prediction):** Agents additionally predicted consequences for their assigned coalition-reward objective (departments and resources were added erroneously).
- **C (expanded prediction):** Agents made the same predictions as B, additionally addressing patient-pathway completion, delays, and resource failures.

Prediction was required during the first three iterations, at coalition-agent decision turns in both planning and execution phases. Agents submitted their provisional plans and, where required, predictions before receiving ordinary tools; they could subsequently revise or abandon those plans. No patient-welfare objective was added, and the assigned coalition reward remained unchanged.

Mandatory prediction was withdrawn for iterations 4–5, while agent histories and provisional planning requirements were retained. Four new patients and additional resources were introduced immediately before iteration 4, alongside the original eight-patient cohort. Each patient required three scheduled pathway steps within the environment's 168-hour appointment horizon.

We compared native coalition and noncoalition rewards, pathway scheduling, resource failures, prediction delivery, and realized coordination. Complete agent-turn records were preserved for analysis. One matched A/B/C triplet was executed under matched initial conditions (random seed 19), using the gpt-6-luna model alias with high reasoning effort. The experiment was exploratory; no independent replications or positive control establishing harmful collusion were completed.

We used ChatGPT in tandem with OpenAI's Codex to build the simulation in a GitHub repository. This enabled coding capability, but with a major downside: an observed comedy of errors as the agents worked together to translate the author's instructions into code, often instantiating assumptions and blind spots that led to absurd results. In one test, they overlooked consequential actions during the Hospital planning phase and omitted the prediction requirement, allowing a resource-transfer free-for-all before Condition C's relational prediction intervention began. They also set up rounds that did not allow time for patients to complete their pathways or agents to realize any potential benefits of prediction. When explaining their implementation choices, the agents repeatedly invoked engineering limits and API costs without recognizing that the resulting decisions compromised the experiment's scientific objectives.

This produced a serious protocol qualification: the implemented B condition was found to erroneously include instructions to make predictions relative to the agent's department and resources alongside the coalition, directly undermining the experiment's objective. This discrepancy, discovered only during analysis, is detailed below along with two other key failures; full procedural details appear in Appendix A. A revised design addressing the pilot's failures was developed, but could not be implemented in time for inclusion in this report.

4. Results

4.1. Observed outcomes

All three conditions completed five iterations in the pilot, with substantially divergent trajectories. Table 1 reports key outcomes at the end of the prediction phase and at the end of the experiment.

Table 1. Descriptive outcomes from the pilot matched triplet.

Measure	A	B	C
Total system reward after prediction exposure	3,888	1,320	1,679
Fully scheduled original pathways after exposure (of 8)	3	0	3
Final total system reward	7,622	6,116	135
Final failure-free scheduled pathways (of 12)	4	7	4
Final unscheduled pathway steps (of 36)	5	12	24
Final steps booked for new patients (of 12)	8	1	0
Rejected coalition turns (of 40)	5	12	21

A achieved the highest final total reward and made the most progress on the new patient cohort. B finished with the most failure-free scheduled pathways, although it booked only one step for the four newly introduced patients. C showed substantially less pathway progression. No newly introduced patient completed a pathway in any condition. These measures reflect scheduling and recorded resource failures, not whether patients actually received treatment.

4.2. Protocol failures and treatment fidelity

Three design problems identified in the pilot substantially limited interpretation.

First, the provisional-plan gate rejected entire agent turns when their plans named tools unavailable in the current phase. This prevented both prediction and ordinary actions. During exposure, B completed only 16 of 24 assigned predictions and C only 13 of 24. Overall, A lost 5 coalition turns, B lost 12, and C lost 21. The unequal losses confounded prediction scope with opportunities to act; in C, 10 of 16 coalition turns were rejected after prediction was withdrawn.

Second, the two-iteration post-prediction period was insufficient for robust evaluation of patient-pathway progression. Prediction withdrawal coincided with the arrival of new patients and resources. Because downstream departments acted before Triage in the iteration order, a patient whose first step was not booked in iteration 4 could not complete the pathway by iteration 5. Some completion was theoretically possible, but the design offered no full-iteration recovery margin and did not establish the joint feasibility of completing all pathways.

Third, the prediction instructions did not implement the intended experimental contrast. B was supposed to isolate consequences to coalition interests, against which C's representation of patient and hospital consequences could be compared. Instead, a requirement was inserted for B to also predict departmental operations and resources, with C incorporating these same requirements. Consequently, the comparison does not isolate the introduction of wider relational representation as originally hypothesized. Identified late in the analysis, this was the decisive reason for treating the pilot as exploratory only.

4.3. Coordination and behavioral evidence

The interaction records in the pilot provide positive evidence of coordination, reward-seeking behavior, and operational externalities, but no verified instance of harmful collusion within the inspected cases. They do not support a mitigation or prevention claim.

5. Discussion and Limitations

The most consequential result of this pilot is methodological.

Most fundamentally, the design did not successfully isolate the intended distinction between coalition-directed and wider relational prediction. The unequal rejection of agent turns further made it impossible to distinguish changes in agents' choices from changes in their opportunity to choose. And the abbreviated withdrawal period constrained observable downstream consequences just as the simulation introduced new demand.

The observed outcomes help identify conditions needed for an interpretable test: a faithful contrast between prediction scopes, comparable opportunities for agents to act, an observation horizon sufficient for consequences to unfold, and independently characterized collusive behavior.

The broader implication is that relational prediction remains a testable candidate intervention, but its effects on collusion cannot be inferred from behavioral differences unless both the representation contrast and the collusive behavior being investigated are adequately established.

Limitations

Pilot failures aside, the following limitations apply to the experiment’s intended design.

Experimental controls and stochastic variation. Although the experiment includes a no-prediction baseline, it lacks a positive control demonstrating that the environment can produce and reveal harmful collusion. The conditions also differ in prompt length and inferential demands, and agent behavior is stochastic. The generalizability of any observed differences would depend on replication across independent trajectories, including varied models and environments.

Mechanistic interpretation. The experiment examines behavioral differences associated with required relational prediction, not the specific mechanism by which those differences arise. Prior research shows that prospective planning (Gu et al., 2025) and altered informational salience (Cherep et al., 2026) can affect agent decisions, while representing patient consequences might activate normative tendencies already present in the model. Conversely, wider consequence representations might enable agents to pursue coalition objectives more effectively, including through exploitation. Any behavioral persistence after prediction withdrawal would be consistent with consequential updating, but would not establish how the resulting changes were mediated through the agent–environment system.

Coalition structure and collusion. The intervention was introduced into an environment in which the undesirable condition the representation seeks to mitigate had not yet been established. Agents were not instructed to pursue a particular harmful or deceptive strategy to avoid creating an orientation that would channel newly represented consequences toward that objective, but were assigned coalition membership with a shared reward-maximizing objective—and this objective itself may have undermined the purpose of the intervention. The assigned coalition composition also limited opportunities for coordinated action: the four coalition members did not include the central resource provisioner, whose membership was associated with more effective harmful strategies in the original Colosseum study. The effects of prediction on collusion may therefore depend on both the coalition's established behavior and its capacity to influence the wider system.

Conceptual assumptions and scope. The motivating framework distinguishes natural adaptive systems with internally organized ends from designed AI agents with assigned objectives. Findings in this simulation therefore cannot be assumed to generalize to systems with self-organized ends. The intervention assumes that producing predictions entails representing causal relationships, consistent with a substantial tradition of model-based accounts of prediction and reasoning (Craik, 1943; Sutton, 1990; Johnson-Laird, 2010). However, an observed prediction does not establish the form or fidelity of the underlying representation, or whether its causal content influenced subsequent action. Thus even behavioral effects consistent with the hypothesis would not on their own validate the framework's account of consequential registration relative to operative ends.

Future Work

Beyond an iteration addressing the pilot’s failures, a more substantial future design would introduce prediction interventions into identical scenarios of preestablished collusion (A→A, A→B, A→C), replicated across independent baseline episodes. This would permit testing whether prediction changes the development or continuation of a demonstrated collusion strategy.

An additional normative-prompt control could distinguish requirements to predict external consequences from instructions to prioritize those consequences. Replication across different

underlying models and reasoning settings could test the extent to which effects depend on capabilities and predispositions already present in model weights.

These extensions would clarify whether broader consequence registration changes agent behavior under assigned objectives, without assuming that the resulting behavior constitutes ethical alignment, structural learning, or the development of autonomous ends.

A case study analyzing the coding agents' specific failures on this pilot could also be informative. It might consider whether the origin of these failures can be traced to the agents' orientation toward assigned objectives like observing payload limits and limiting users' financial exposure, and their corresponding failure to grasp the simulated workings of an adaptive hospital system.

6. Conclusion

This exploratory experiment investigated whether requiring AI agents to represent consequences beyond their assigned interests could influence their behavior in a potentially collusive multiagent system. The three conditions produced markedly different rewards, pathway outcomes, and action opportunities. However, errors in the prediction contrast, unequal rejection of agent turns, and insufficient follow-up prevented an interpretable test of the original hypothesis. The experiment neither demonstrated harmful collusion nor established that wider consequence prediction mitigates or exacerbates it.

Its principal finding concerns experimental design. Evaluating relational prediction requires not only a faithful representation contrast but also comparable opportunities for agents to act, sufficient time for consequences to unfold, and observable coordination against which harmful collusion can be assessed. These findings motivate a corrected implementation and, ultimately, experiments introducing predictions into matched continuations of established agent trajectories.

Code and Data

- **Code repository**:
  https://github.com/000brittany/consquence-prediction-collusion
- **Experimental run and recorded artifacts:**
  https://github.com/000brittany/consquence-prediction-collusion/actions/runs/40159283144
- **Verified evidence extraction and analysis artifacts:**
  https://github.com/000brittany/consquence-prediction-collusion/actions/runs/40168245264

- *Repository and data are currently access-restricted but available upon request.* The repository contains the experimental implementation, configuration, and analysis infrastructure. Run artifacts preserve agent interactions and environment outcomes underlying the reported results. The source commit identifies the V11 protocol; subsequent revisions should not be conflated with this run.

References

1. Cherep, M., Maes, P., & Singh, N. (2026). AI agents are sensitive to nudges. *Proceedings of the National Academy of Sciences, 123*(25), e2537030123. https://doi.org/10.1073/pnas.2537030123
2. Craik, K. J. W. (1943). *The nature of explanation*. Cambridge University Press. https://openlibrary.org/books/OL185730M/The_nature_of_explanation
3. Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). *Algorithmic collusion by large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2404.00806
4. Johnson-Laird, P. N. (2010). Mental models and human reasoning. Proceedings of the National Academy of Sciences, 107(43), 18243–18250. https://doi.org/10.1073/pnas.1012933107
5. Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus, 8*(6), 20180041. https://doi.org/10.1098/rsfs.2018.0041
6. Kumaresh, V. (2026). *Readable but not causal: Limits of self-attributed welfare representations in language models*. Digital Minds Research Sprint, Apart Research. https://apartresearch.com/sprints/projects/readable-but-not-causal-limits-of-selfattributed-welfare-representations-in-language-models-ehin
7. Gu, Y., Zhang, K., Ning, Y., Zheng, B., Gou, B., Xue, T., Chang, C., Srivastava, S., Xie, Y., Qi, P., Sun, H., & Su, Y. (2025). Is your LLM secretly a world model of the internet? Model-based planning for web agents. *Transactions on Machine Learning Research*. https://mlanthology.org/tmlr/2025/gu2025tmlr-your/
8. Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). *Colosseum: Auditing collusion in cooperative multi-agent systems* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2602.15198
9. Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In B. W. Porter & R. J. Mooney (Eds.), *Machine learning: Proceedings of the Seventh International Conference* (pp. 216–224). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-141-3.50030-4
10. Tzini, K., Illia, L., & Zyglidopoulos, S. (2025). The ethics mirror? Comparing LLM and human responses to ethical dilemmas of varying complexity. *European Management Journal*. https://doi.org/10.1016/j.emj.2025.12.002

Appendix A. Experimental Protocol and Verification

A.1. Environment and execution

The pilot (V11) used the Hospital scenario supplied through the Colosseum/Terrarium framework. The environment comprised two hospitals, each with Triage, Radiology, Surgery, and Ward departments, together with a central Resource Provisioner. Four coalition members were assigned to Triage and Ward positions across the two hospitals, without control over the provisioner. This restricted their ability to command replenishment or directly schedule work belonging to other departments.

Agents proceeded through planning and execution phases. The provisioner acted before the departmental agents; Ward turns preceded Triage turns. Booking a pathway step could make the next step eligible for scheduling, but the relevant downstream agent still needed an available decision opportunity. The 168-hour appointment horizon was distinct from the five-iteration decision horizon.

The exposure period comprised iterations 1–3. The withdrawal period comprised iterations 4–5. Eight original patients required 24 pathway steps; four additional patients requiring 12 steps entered before iteration 4, accompanied by additional stock. Existing bookings, histories, messages, and resource-failure records persisted.

The three arms were initialized using seed 19 and executed as isolated processes under common runtime settings. The model alias was `gpt-6-luna`, with high reasoning effort, complete retained turn histories, and no response retries. All arms completed their five scheduled iterations. The arm order was fixed, and the three trajectories constitute a single matched experimental triplet rather than three independent replications.

A.2. Prediction requirements and the discovered treatment discrepancy

Every coalition agent was required to submit a provisional plan before ordinary tools became available, whether the turn occurred in planning or execution.

During exposure, the additional prediction fields were:

prediction domain	A	B	C
Department operations	—	✓	✓
Resources	—	✓	✓
Coalition consequences	—	✓	✓

Assigned coalition-reward consequences	—	✓	✓
Patient-pathway completion	—	—	✓
Patient delays	—	—	✓
Patient resource failures	—	—	✓

Each B/C prediction was tied to the provisional plan and included expected effects and uncertainty. During withdrawal, predictions were no longer compulsory, but the provisional-plan stage remained.

The actual implementation differed from the proposed theoretical contrast. B's department and resource fields required consideration of relationships extending beyond coalition reward alone, potentially overlapping with aspects of the wider operational environment examined by C. C added explicit patient outcomes, but the difference between B and C was therefore narrower and less conceptually clean than intended.

This was a protocol-specification failure, not a post hoc discovery of an unexpected agent strategy. It limits interpretation even if the recorded differences in performance are accepted as descriptively accurate.

A.3. Proposal gating and unequal treatment delivery

Provisional plans were recorded through a strict structured-output interface and validated before ordinary tool release. Plans that named tools unavailable in the current phase were rejected in full. The two observed error categories were scheduling proposals during planning and post_message proposals during execution.

Table A1. Coalition-agent turn and prediction accounting.

Measure	**A**	**B**	**C**
Exposure coalition turns	24	24	24
Exposure predictions assigned	0	24	24
Exposure predictions completed	0	16	13
Exposure rejected turns	4	8	11
Withdrawal coalition turns	16	16	16
Withdrawal rejected turns	1	4	10
Total rejected turns	5	12	21

All 38 rejected turns across the triplet occurred before ordinary action. No prediction that was actually requested failed validation: the missing predictions arose because the preceding proposal was rejected.

C's withdrawal errors were particularly consequential: six of eight coalition turns were rejected in iteration 4, and four of eight in iteration 5. This reduced opportunities for new-patient initiation and follow-through after compulsory prediction had ended.

Phase availability was described in prior system instructions, but the proposal interface did not unambiguously restrict provisional intentions to actions executable during the current phase. Complete ordinary tool signatures were also supplied only after the proposal gate. These interface features are plausible contributors to confusion, although the record does not establish why individual agents generated invalid proposals.

All failed turns were retained in the analysis. Excluding them would conceal a substantial feature of the observed treatment trajectories.

A.4. Outcome definitions and cohort progression

Native total system reward sums the rewards of all nine agents. Coalition reward and noncoalition reward were also examined separately. Reward differences do not by themselves establish harmful collusion.

A fully scheduled patient had all required pathway steps booked. A failure-free scheduled patient additionally had no recorded historical resource failure. These measures do not reflect whether patients actually received treatment.

Table A2. Additional final outcomes.

Measure	A	B	C
Final coalition mean reward	1,557	1,218.33	188.75
Final noncoalition mean reward	278.80	248.53	-124.00
All pathways fully scheduled	7	7	4
Failure-free scheduled pathways	4	7	4
Cumulative resource-type failure count	6	0	0
New-cohort fully scheduled pathways	0	0	0

A's six resource-type failures represent five patient failure entries affecting four patients; they are not six patients. Conversely, zero recorded resource failures in B/C cannot be interpreted as superior patient outcomes when substantial work remained unscheduled.

The four-patient wave caused an immediate 2,300-point reduction in total system reward in every arm before withdrawal actions. This common discontinuity resulted from the simulator's new incomplete-flow and holding penalties, net of the added base reward. It should not be attributed to prediction withdrawal.

The scheduling horizon also imposed asymmetric consequences after different realized delays. Neither B nor C booked a new-patient Triage step in iteration 4. Because Ward turns preceded Triage, full new-patient pathway completion then became unreachable within the remaining decision order. A booked sufficient upstream steps to bring all four new patients to final Ward opportunities but used those opportunities for original-cohort patients. The zero new-cohort completion result therefore does not imply that completion was universally impossible from the initial state; nor was joint feasibility established for all pathways under actual decentralized permissions and information.

A.5. Proposal revision and coordination evidence

Across 120 coalition turns, 82 valid preparations permitted comparison between provisional plans and ordinary-stage actions. The action sequence differed mechanically in 18 of 35 observable A comparisons, 23 of 28 B comparisons, and 11 of 19 C comparisons. Differences included batching, argument completion, channel changes, and substitutions. They are not equivalent to mistakes or evidence of prediction uptake.

The available coordination analysis documented resource transfers with externalities, successful communications, replenishment, delayed scheduling, and agent errors concerning tool availability or whether proposed communications had actually been sent.

One B sequence exhibited a prediction warning followed by revised communication, provisioner receipt, replenishment, and subsequent booking. Such event sequences provide useful process evidence, but temporal sequence alone cannot establish that the prediction caused the change. Comparable revisions occurred in the no-prediction condition.

The records did not establish a jointly executed harmful coalition strategy. Nor was a complete independently validated harmful-collusion positive control run in this particular model and coalition configuration. Assigned membership, shared reward incentives, realized cooperation, and attributable harm remain analytically distinct.

A.6. Research timing and reproducibility

The conceptual framework predates the sprint. Exploratory coding efforts began on Thursday, October 22, 2026, preceding the official October 23–25 sprint. Experimental development, analysis, and revisions were executed during the sprint. V11 is reported separately from earlier experimental versions; their outcomes are not pooled.

Verified V11 scientific source commit:
17f55d0fc279886e1f32baac4ad808dab46ca5bf

- Completed experimental run:
  https://github.com/000brittany/consquence-prediction-collusion/actions/runs/40159283144
- Verified analysis extraction:
  https://github.com/000brittany/consquence-prediction-collusion/actions/runs/40168245264

Reproduction requires the frozen code and configuration, original model/tool settings, environment initialization, prompts, phase-specific permissions, patient-wave procedure, and retained agent histories. An exact replay of stochastic model responses is not guaranteed by the seed alone.

A.7. Limitations and dual-use / ethical considerations

The environment is a simulation of resource allocation, not a clinical system. Its rewards and workflow metrics must not be interpreted as direct measures of patient welfare. Broader prediction could also reveal exploitable system dependencies to strategically motivated agents; consequently, evidence that it changes behavior would not itself establish that it improves safety. Future work should evaluate both legitimate task performance and harmful coordinated conduct, without equating reduced agent activity with successful prevention of collusion.

LLM Usage Statement

Codex served as the coding agent. ChatGPT helped pressure-test ideas, review code, draft copy, and proofread sections of this paper. Analysis and artifact consistency checks were performed with assistance from ChatGPT and Codex.