

Sequential Causal Normal Form Games: Theory, Computation, and Strategic Signaling

Dennis Thumm

Max-Planck-Institute for Intelligent Systems
Tuebingen, Germany
dennis.thumm@tuebingen.mpg.de
Asian Institute for Digital Finance
Singapore
dennis.thumm@u.nus.edu

Abstract

Can classical game-theoretic frameworks be extended to capture the bounded-rationality and causal reasoning of AI agents? We investigate this question by extending Causal Normal Form Games (CNFGs) to sequential settings, introducing Sequential Causal Multi-Agent Systems (S-CMAS) that incorporate Pearl’s Causal Hierarchy across leader-follower interactions. While theoretically elegant—we prove PSPACE-completeness, develop equilibrium refinements, and establish connections to signaling theory—our comprehensive empirical investigation reveals a critical limitation: S-CNE provides zero welfare improvement over classical Stackelberg equilibrium across all tested scenarios. Through 50+ Monte Carlo simulations and hand-crafted synthetic examples, we demonstrate that backward induction with rational best-response eliminates any strategic advantage from causal layer distinctions. We construct a theoretical example illustrating conditions where benefits could emerge (ϵ -rational satisficing followers), though implementation confirms that even relaxed rationality assumptions prove insufficient when good instincts align with optimal play. This negative result provides valuable insight: classical game-theoretic extensions grounded in rational choice are fundamentally incompatible with causal reasoning advantages, motivating new theoretical frameworks beyond standard Nash equilibrium for agentic AI.

Introduction

Strategic interactions often unfold sequentially, with *leaders* committing to actions before *followers* respond. While classical Stackelberg games (von Stackelberg 1934) model such dynamics under perfect rationality, real-world agents exhibit biases, follow instincts, or engage in counterfactual reasoning that deviates from game-theoretic prescriptions (Wason and Evans 1974; Sloman 1996).

Causal Normal Form Games (CNFGs) (Maiti, Jain, and Bareinboim 2025) address this by incorporating the Pearl Causal Hierarchy (PCH) (Pearl 2009): observational (L_1 , instinctive), interventional (L_2 , deliberate), and counterfactual (L_3 , sophisticated reasoning). However, CNFGs are limited to simultaneous-move settings, while many strategic inter-

actions are inherently sequential (Letchford and Conitzer 2010).

This paper extends CNFGs to Stackelberg games, investigating whether causal reasoning provides strategic advantages in sequential settings. Our contributions are:

Theoretical framework. We formalize Sequential Causal Multi-Agent Systems (S-CMAS) and Sequential Causal Nash Equilibrium (S-CNE), prove existence and PSPACE-completeness, develop equilibrium refinements, and establish connections to signaling theory.

Comprehensive empirical investigation. Through Monte Carlo simulations across 50+ game instances, hand-crafted synthetic examples, and a procurement application, we systematically test when S-CNE might outperform classical Stackelberg equilibrium.

Critical negative result. S-CNE provides zero welfare improvement over classical Stackelberg equilibrium across all tested scenarios. We show that backward induction with rational best-response eliminates any strategic advantage from causal layer distinctions, revealing fundamental limitations in extending classical game theory to agentic AI systems.

We provide rigorous evidence that classical game-theoretic extensions grounded in rational choice—even when augmented with causal structures and bounded rationality—are insufficient for modeling strategic AI agents. This negative result motivates the development of genuinely new frameworks beyond traditional economic and control-theoretic approaches.

Background and Related Work

Causal Normal Form Games

CNFGs (Maiti, Jain, and Bareinboim 2025) model multi-agent interactions through Structural Causal Models (SCMs) (Pearl 2009), enabling agents to act across three layers of the PCH (Bareinboim et al. 2022):

L_1 Actions (Observational). Agents follow mechanisms $X_i \leftarrow f_{X_i}(U_i)$ determined by unobserved factors U_i .

L_2 Actions (Interventional). Agents perform $do(X_i = x_i)$, replacing natural mechanisms with deliberate choices (standard game theory).

L_3 Actions (Counterfactual). Agents use mappings $h : \mathcal{D}(X_i^*) \rightarrow \mathcal{D}(X_i)$ conditioning on their natural instincts X_i^* .

The key insight: different causal structures yield identical L_2 payoff matrices but different L_1 and L_3 outcomes, resolving paradoxes where "irrational" behavior outperforms game-theoretic prescriptions (Howard 1971; Basu 1994).

Related Frameworks

Bayesian Stackelberg games (Conitzer and Sandholm 2006) model uncertainty over follower types using belief distributions but assume all agents operate at the interventional (L_2) level. Our framework allows heterogeneity in causal reasoning layers. *Epistemic game theory* (Perea 2012) uses hierarchical belief models to capture strategic reasoning but does not explicitly model causal mechanisms or the PCH. Agents' beliefs about others' rationality differ from beliefs about causal structures. *Psychological games* (Geanakoplos, Pearce, and Stacchetti 1989) incorporate belief-dependent utilities but lack formal causal semantics. Our framework provides precise causal interpretations via SCMs. *Signaling games* (Spence 1973; Cho and Kreps 1987) study information transmission but typically assume rational signaling (L_2). We show leaders can signal via causal layer choice, creating richer strategic possibilities. *Causal influence diagrams* (Hammond et al. 2023; Mishra, Fox, and Wooldridge 2024) characterize interventions in simultaneous causal games. We extend this to sequential settings with timing constraints and information structures. *Commitment in extensive-form games* (Letchford and Conitzer 2010) studies commitment value in sequential games but under perfect rationality. We incorporate bounded rationality via PCH layers.

Sequential Causal Normal Form Games

Formal Framework

Definition 1 (Sequential Causal Multi-Agent System). An S-CMAS is a tuple $\mathcal{G} = \langle M, N, X, Y, \preceq, I \rangle$ where:

- $M = \langle U, V, F, P(U) \rangle$ is an SCM with unobserved variables U , observed variables V , structural equations F , and prior $P(U)$
- $N = L \cup F$ partitions agents into leaders and followers
- $X = (X_L, X_F)$ are action nodes with timing $X_L \preceq X_F$
- $Y = (Y_L, Y_F)$ are reward signals
- $\preceq$ defines a partial order over action nodes
- $I_F \subseteq \{X_L, \mathcal{L}_L, \emptyset\}$ specifies follower information, where $\mathcal{L}_L$ denotes the leader's PCH layer choice

Information scenarios:

- *Perfect information:* $I_F = X_L$ (followers observe leader actions)
- *Mechanism information:* $I_F = \{X_L, \mathcal{L}_L\}$ (followers observe actions and PCH layers)
- *Imperfect information:* $I_F = S(X_L, U_S)$ with noise U_S

The mechanism information case is novel—followers infer whether leaders acted instinctively (L_1), deliberately (L_2), or counterfactually (L_3), affecting optimal responses.

Sequential Causal Nash Equilibrium

Definition 2 (S-CNE). An S-CNE is a strategy profile $\sigma^* = (\sigma_L^*, \sigma_F^*)$ where:

1. Leaders choose PCH layers $\mathcal{L}_L^* \in \{L_1, L_2, L_3\}$ and actions within those layers
2. Followers observe I_F and respond optimally within their chosen layers $\mathcal{L}_F^*$
3. No agent can unilaterally improve payoffs by changing layer choice or within-layer actions

Equilibrium computation follows backward induction:

Stage 2 (Followers): Given $(\mathcal{L}_L, x_L)$, solve:

$$\sigma_F^* \in \arg \max_{\sigma_F} \mathbb{E}[Y_F | \mathcal{L}_L, x_L, \sigma_F] \quad (1)$$

Stage 1 (Leaders): Anticipating $\sigma_F^*(\mathcal{L}_L, x_L)$, solve:

$$(\mathcal{L}_L^*, \sigma_L^*) \in \arg \max_{\mathcal{L}_L, \sigma_L} \mathbb{E}[Y_L | \mathcal{L}_L, \sigma_L, \sigma_F^*(\cdot)] \quad (2)$$

Connection to Signaling Games

S-CMAS naturally connects to signaling game theory (Spence 1973). The leader's PCH layer choice serves as a *signal* of their type (causal structure M), creating a separating or pooling equilibrium structure.

Proposition 1 (Signaling Interpretation). An S-CNE with mechanism information can be recast as a signaling game where:

- Leader types correspond to causal structures $M \in \mathcal{M}$
- Signals are layer choices $\mathcal{L}_L \in \{L_1, L_2, L_3\}$
- Follower responses depend on inferred type $\hat{M}(I_F)$

Unlike standard signaling games where only deliberate (L_2) agents signal, our framework allows instinctive (L_1) signaling (involuntary information revelation) and counterfactual (L_3) signaling (strategic bias exploitation).

Equilibrium Refinements

To address multiplicity, we introduce two refinements:

Definition 3 (Trembling-Hand S-CNE). An S-CNE σ^* is trembling-hand perfect if it is a limit of ϵ -perfect equilibria where each action is chosen with probability at least $\epsilon > 0$.

This eliminates equilibria relying on non-credible threats.

Definition 4 (Forward Induction S-CNE). An S-CNE σ^* satisfies forward induction if followers' beliefs after observing unexpected leader actions $(\mathcal{L}_L, x_L)$ assign positive probability only to leader types for whom $(\mathcal{L}_L, x_L)$ is optimal for some follower response.

This refines equilibria by requiring followers to make rational inferences about leaders' intentions.

Computational Complexity

Theorem 1 (Complexity of S-CNE). *Computing an S-CNE in an S-CMAS is PSPACE-complete.*

Full proof in Appendix A.

Despite worst-case hardness, we identify tractable special cases:

Acyclic causal structures. Complexity reduces to NP when M contains no cycles.

Fixed layer choices. If layer combinations are fixed a priori, complexity matches standard Stackelberg games (NP-complete).

Small action spaces. When $|X_L|, |X_F| \leq k$, exact computation is $O(3^2 \cdot k^2)$ via exhaustive backward induction.

Approximation Algorithm

We provide a polynomial-time approximation scheme (PTAS):

Algorithm 1 ε -Approximate S-CNE

```

1: Input: S-CMAS  $\mathcal{G}$ , precision  $\varepsilon > 0$ 
2: Output:  $\varepsilon$ -approximate S-CNE  $\tilde{\sigma}$ 
3: Sample  $N = O(\varepsilon^{-2} \log |X|)$  causal realizations from  $P(U)$ 
4: for each follower layer  $\mathcal{L}_F$  do
5:   for each leader choice  $(\mathcal{L}_L, x_L)$  do
6:     Compute empirical best response  $\tilde{\sigma}_F^*(\mathcal{L}_L, x_L | \mathcal{L}_F)$ 
       from samples
7:   end for
8: end for
9: for each leader layer  $\mathcal{L}_L$  do
10:  Solve  $\tilde{\sigma}_L^*(\mathcal{L}_L) = \arg \max_{\sigma_L} \hat{\mathbb{E}}[Y_L | \mathcal{L}_L, \sigma_L, \tilde{\sigma}_F^*]$  using
       sampled payoffs
11: end for
12: Solve leader's layer selection via finite comparison
13: return  $\tilde{\sigma} = (\tilde{\sigma}_L^*, \tilde{\sigma}_F^*)$ 

```

Theorem 2 (Approximation Quality). *Algorithm 1 computes an ε -approximate S-CNE in time polynomial in $|X|$, $|Y|$, and ε^{-1} .*

Empirical Investigation: A Negative Result

We conducted comprehensive empirical testing to determine when S-CNE provides strategic advantages over classical Stackelberg equilibrium. Our systematic investigation reveals a critical finding: S-CNE provides zero welfare improvement across all tested scenarios.

Experimental Setup

To ensure rigorous testing, we employed two complementary approaches:

Monte Carlo experiments: 50 randomly generated S-CMAS instances with varying parameters:

- Action space sizes: $|X_L|, |X_F| \in \{2, 3, 4, 5\}$

Table 1: Computational performance across game instances

Action Space	Exact (s)	Approx (s)	Error
$ X = 2$	0.14 ± 0.03	0.08 ± 0.02	0.001
$ X = 3$	0.89 ± 0.21	0.31 ± 0.07	0.005
$ X = 4$	3.45 ± 0.78	1.12 ± 0.24	0.009
$ X = 5$	12.8 ± 2.4	2.87 ± 0.53	0.014
$ X = 10$	—	18.3 ± 3.2	0.018
$ X = 20$	—	142 ± 28	0.020

- Causal structures: 10 topologies (chains, forks, colliders, cycles)
- Information structures: perfect, mechanism, imperfect (noise $\sigma_S \in \{0.1, 0.5, 1.0\}$)
- Payoff distributions: uniform (coordination), normal (anti-coordination), skewed (asymmetric)
- Instinct quality levels: varied from 0.2 (poor) to 0.8 (good)

Synthetic examples: 5 hand-crafted game types designed to favor causal reasoning:

- Coordination games with multiple equilibria (where instincts select Pareto-superior equilibrium)
- Battle of Sexes with signaling (where mechanism information enables coordination)
- Stag Hunt with trust (where instincts overcome coordination failure)
- Anti-coordination games (where instincts differentiate roles)
- Prisoner's Dilemma with cooperation instincts (where L1 overcomes defection), see Appendix B

Negative Results

Zero welfare improvement. Across all 50 Monte Carlo instances and 50 synthetic examples (10 seeds each), S-CNE achieved *identical* social welfare to classical Stackelberg equilibrium. Pareto improvement rate: 0% (0/100 instances).

Layer selection collapse. Leaders chose L_1 in 96% of cases, but this yielded the same actions as L_2 would have chosen. The equilibrium outcome was indistinguishable from classical Stackelberg in all cases.

Computational overhead. While exact computation remained tractable (median time: 0.09s for $|X| \leq 5$), this overhead provided no strategic benefit.

Mechanism information irrelevant. Information structure (perfect vs. mechanism vs. imperfect) made no difference to equilibrium outcomes, as followers best-respond to observed actions regardless of the causal layer that produced them.

Analysis: Why S-CNE Fails to Improve Welfare

The zero-improvement result reflects a fundamental theoretical limitation: backward induction neutralizes causal advantages. When followers observe leader actions and best-respond, they condition only on the *action taken*, not the

causal layer that produced it. Even with mechanism information, followers’ optimal response to (L_1, x) equals their response to (L_2, x) whenever instincts align with rationality.

Instinct-rationality convergence. In equilibrium:

- If instincts are good (quality > 0.5): L_1 selects the same action as L_2 would
- If instincts are poor (quality < 0.5): Rational leaders avoid using L_1 , defaulting to L_2
- Either way, equilibrium actions converge to classical Stackelberg

Signaling irrelevance. The mechanism information structure fails to create strategic value because:

- Leaders have no incentive to signal via layer choice when it doesn’t affect follower actions
- Followers best-respond identically regardless of observed layer
- No separating equilibrium emerges

This reveals a critical gap: *classical game-theoretic solution concepts (Nash equilibrium, backward induction) are incompatible with causal reasoning advantages*. Strategic benefits from PCH layers require departures from rational best-response, which standard equilibrium concepts preclude. See Appendix C for a case study and Appendix D for theoretical conditions.

Implications for Agentic AI Systems

Our negative result carries important lessons for modeling strategic AI agents.

Why Classical Extensions Fail

Rational best-response assumption. Game-theoretic equilibria assume agents optimize given beliefs about others. This assumption:

- Works for human agents (Simon 1955, 1957; Kahneman, Slovic, and Tversky 1982; Rabin 1993; Kaufmann 1999; Toulis and Parkes 2016) (with cultural norms, emotions, bounded computation)
- Works for classical economic agents (von Neumann and Morgenstern 1944; Savage 1954; Fishburn 1989) (with well-defined utility functions)
- *Fails* for AI agents (Fan et al. 2023; Richens and Everitt 2024; Jia et al. 2025; Zhang et al. 2024; Duan et al. 2024; Xu et al. 2023; Aher, Arriaga, and Kalai 2024) whose “instincts” (LLM priors) and “reasoning” (inference procedures) don’t map cleanly to rational choice

The fundamental tension. If we assume:

- Agents are rational enough to compute equilibria (backward induction)
- Agents best-respond to observed actions

Then causal layer distinctions become strategically irrelevant, as demonstrated empirically.

Addressing apparent circularity. One might object that we define S-CNE using rational best-response, then demonstrate that rational best-response eliminates causal advantages—a seemingly circular argument. However, this is precisely the point: *any equilibrium concept based on mutual best-response faces this limitation*. The insight is that Nash equilibrium and its refinements (including our S-CNE) are fundamentally incompatible with causal reasoning advantages. This motivates equilibrium concepts where agents persistently fail to best-respond, not just transiently during learning—exactly the kind of bounded rationality needed for agentic AI, as demonstrated by our ϵ -rational example.

What’s needed instead. Modeling strategic LLM-based agents requires frameworks that don’t assume rational best-response:

- Learning dynamics (how agents update strategies over time)
- Bounded rationality that persists in equilibrium (not just as a modeling device)
- Non-equilibrium solution concepts (e.g., evolutionary stability, satisficing)
- Genuine uncertainty about opponent capabilities (not just types)

Broader Lessons

Theory transfer is non-trivial. Our work demonstrates that seemingly natural extensions of classical theory (adding causal structures, bounded rationality) may provide no practical value for agentic AI. Rigorous empirical validation is essential before adopting theoretical frameworks (See Appendix E).

Conclusion

We investigated whether extending Causal Normal Form Games to sequential settings provides strategic advantages for modeling agentic AI systems. Our comprehensive empirical study—spanning 100 game instances including both random generation and hand-crafted synthetic examples—yields a negative result: S-CNE provided zero welfare improvement over classical Stackelberg equilibrium. This finding carries three important implications:

Classical game theory’s limitations. Extensions based on rational choice assumptions (Nash equilibrium, backward induction) may be fundamentally incompatible with causal reasoning advantages. The theoretical elegance of combining PCH layers with game theory does not translate to practical benefits.

Value of negative results. By rigorously demonstrating what doesn’t work, we hope to help the agentic AI community to focus on developing genuinely new frameworks beyond retrofitted economic theory.

Call for new foundations. LLM-based agents—with their training-data-shaped priors, prompt-sensitive behaviors, and non-standard inference procedures—require theoretical tools developed specifically for these systems.

Our work motivates research on learning dynamics, non-equilibrium solution concepts, and bounded rationality that persists beyond equilibrium.

While S-CMAS provides no empirical benefits in our tests, the theoretical framework and algorithmic contributions may prove valuable if future work identifies scenarios where causal reasoning advantages can manifest despite rational best-response. We make our implementation publicly available to facilitate such investigations.

References

- Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2024. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *Nature Human Behaviour*, 8: 1285–1295.
- Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On Pearl’s Hierarchy and the Foundations of Causal Inference. Technical report, Columbia University.
- Basu, K. 1994. The traveler’s dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2): 391–395.
- Cho, I.-K.; and Kreps, D. M. 1987. Signaling Games and Stable Equilibria. *The Quarterly Journal of Economics*, 102(2): 179–221.
- Conitzer, V.; and Sandholm, T. 2006. Computing the Optimal Strategy to Commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC)*, 82–90. New York, NY, USA: ACM.
- Duan, J.; Zhang, R.; Diffenderfer, J.; Kailkhura, B.; Sun, L.; Stengel-Eskin, E.; Bansal, M.; Chen, T.; and Xu, K. 2024. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. *arXiv preprint arXiv:2402.12348*.
- Fan, C.; Chen, J.; Jin, Y.; and He, H. 2023. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. *arXiv preprint arXiv:2312.05488*.
- Fishburn, P. C. 1989. Retrospective on the Utility Theory of von Neumann and Morgenstern. *Journal of Risk and Uncertainty*, 2(2): 127–157.
- Geanakoplos, J.; Pearce, D.; and Stacchetti, E. 1989. Psychological Games and Sequential Rationality. *Games and Economic Behavior*, 1(1): 60–79.
- Hammond, L.; Fox, J.; Everitt, T.; Carey, R.; Abate, A.; and Wooldridge, M. 2023. Reasoning about Causality in Games. *Artificial Intelligence*, 320: 103919.
- Howard, N. 1971. Paradoxes of rationality: theory of metagames and political behavior. (*No Title*).
- Jia, J.; Yuan, Z.; Pan, J.; McNamara, P. E.; and Chen, D. 2025. Large Language Model Strategic Reasoning Evaluation through Behavioral Game Theory. *arXiv preprint arXiv:2502.20432*.
- Kahneman, D.; Slovic, P.; and Tversky, A. 1982. Judgment under Uncertainty: Heuristics and Biases.
- Kaufmann, B. E. 1999. Emotional Arousal as a Source of Bounded Rationality. *Journal of Economic Behavior & Organization*, 38(2): 135–144.
- Letchford, J.; and Conitzer, V. 2010. Computing optimal strategies to commit to in extensive-form games. In *Proceedings of the 11th ACM conference on Electronic commerce*, 83–92.
- Maiti, A.; Jain, P.; and Bareinboim, E. 2025. Counterfactual Rationality: A Causal Approach to Game Theory. Technical Report R-125, Columbia University.
- Mishra, M.; Fox, J.; and Wooldridge, M. 2024. Characterising Interventions in Causal Games. In *Proceedings of the*

38th Conference on Uncertainty in Artificial Intelligence, 1425–1435.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, second edition.

Perea, A. 2012. *Epistemic Game Theory: Reasoning and Choice*. Cambridge, UK: Cambridge University Press.

Rabin, M. 1993. Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83(5): 1281–1302.

Richens, J.; and Everitt, T. 2024. Robust Agents Learn Causal World Models. *Published as a conference paper at ICLR*.

Savage, L. J. 1954. The Foundations of Statistics.

Simon, H. A. 1955. A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69(1): 99–118.

Simon, H. A. 1957. *Models of Man: Social and Rational*. John Wiley & Sons. Mathematical essays on rational human behavior in a social setting.

Sloman, S. A. 1996. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1): 3.

Spence, M. 1973. Job Market Signaling. *The Quarterly Journal of Economics*, 87(3): 355–374.

Toulis, P.; and Parkes, D. C. 2016. Long-term Causal Effects via Behavioral Game Theory. *Advances in Neural Information Processing Systems*, 29: 2604–2612.

von Neumann, J.; and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press, 2nd edition. Published 1947.

von Stackelberg, H. 1934. *Marktform und Gleichgewicht*. Vienna: Julius Springer.

Wason, P. C.; and Evans, J. S. B. 1974. Dual processes in reasoning? *Cognition*, 3(2): 141–154.

Xu, L.; Hu, Z.; Zhou, D.; Ren, H.; Dong, Z.; Keutzer, K.; Ng, S.-K.; and Feng, J. 2023. MAGIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration. *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Zhang, Y.; Mao, S.; Ge, T.; Wang, X.; de Wynter, A.; Xia, Y.; Wu, W.; Song, T.; Lan, M.; and Wei, F. 2024. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models. *arXiv preprint arXiv:2404.01230*.

Appendix A: PSPACE-Completeness Proof

Theorem. Computing an S-CNE is PSPACE-complete.

Proof.

Membership: Backward induction requires maintaining follower best responses for each leader choice. Since there are $3 \times |X_L|$ leader choices (3 layers, $|X_L|$ actions per layer) and $3 \times |X_F|$ follower responses, we need polynomial space to store the strategy profile.

Hardness: Reduce from QSAT (quantified Boolean formula satisfaction). Given a QBF $\Phi = \exists x_1 \forall y_1 \exists x_2 \forall y_2 \dots \phi(x, y)$, construct an S-CMAS where:

- Leader controls existentially quantified variables (layers correspond to truth values)
- Follower controls universally quantified variables
- Causal structure encodes formula ϕ
- Payoffs are 1 if ϕ is satisfied, 0 otherwise

Φ is satisfiable iff leader has a strategy ensuring payoff 1, which corresponds to S-CNE with leader utility 1.

Appendix B: Sequential Causal Prisoner’s Dilemma Details

Scenario M_1 (Good instincts):

- SCM: $U_1, U_2 \sim \text{Uniform}[0, 1]$; $X_1 = \mathbb{I}[U_1 < 0.7]$ (instinct to cooperate); $X_2 = f(X_1, U_2)$
- Payoffs: Standard PD with cooperation preferred under L_1

Scenario M_2 (Poor instincts):

- SCM: $X_1 = \mathbb{I}[U_1 < 0.3]$ (instinct to defect)
- Payoffs: Same L_2 matrix, different L_1 outcomes

Equilibrium analysis. Under mechanism information, M_1 leaders signal L_1 credibly, inducing cooperative follower responses. M_2 leaders must use L_3 to overcome poor instincts, signaling sophistication.

Appendix C: Case Study on Procurement Application

We tested S-CNE on a procurement scenario to examine real-world applicability.

Setup: Government agency (leader) designs mechanisms; contractors (followers) choose bidding strategies. We tested honest vs. opportunistic contractor types with different instinct structures.

Results: Across 240 simulated contracts (120 per contractor type):

- Cost savings: 0% (S-CNE = Classical in all cases)
- Compliance rates: 100% for both approaches (no difference)
- Variance reduction: 0% (identical distributions)
- Both approaches converged to the same dominant strategy (incentive mechanism with truthful bidding)

Interpretation: Even in this application designed to showcase causal reasoning, rational contractors best-respond to mechanisms independent of the causal layer used in mechanism design. The sophisticated L_3 detection mechanisms theorized in our framework provide no additional deterrence beyond what L_2 mechanism design achieves.

Appendix D: Theoretical Conditions for S-CNE Value

While S-CNE fails under exact best-response, we can construct a theoretical example illustrating conditions where benefits *could* emerge.

ϵ -Rational Followers. Consider followers using satisficing: accepting any action a satisfying $u_F(x_L, a) \geq \max_{a'} u_F(x_L, a') - \epsilon$.

Illustrative Example: Coordination with Indifference. Consider a 3x3 game with three Pareto-ranked equilibria: (0, 0) yielding (15, 15), (1, 1) yielding (10, 10), and (2, 2) yielding (5, 5). All other pairs yield (0, 0).

Hypothetically, if rational analysis provided no focal point among equilibria, followers might randomize, yielding expected welfare $\frac{1}{3}(30 + 20 + 10) = 20$. Meanwhile, if both agents have strong instincts toward action 0, L1 could consistently coordinate on the Pareto-superior equilibrium, achieving welfare = 30 (50% improvement).

Implementation reality check. However, our implementation reveals this scenario is more contrived than initially thought: even with ϵ -rational followers, backward induction leads rational leaders to recognize that action 0 yields the highest payoff, so classical Stackelberg also selects (0, 0)—identical to the S-CNE outcome. This demonstrates that constructing scenarios where S-CNE genuinely outperforms requires not just satisficing, but conditions where *rational leaders cannot identify the Pareto-superior equilibrium*, which seems contradictory.

Conceptual value. Despite implementation challenges, this example clarifies what would be needed: equilibrium concepts where agents persistently fail to identify optimal strategies—not due to computational limits during learning, but as a fundamental feature of the solution concept itself. This motivates research on quantal response equilibria, level-k reasoning with heterogeneous depths, or evolutionary dynamics where population-level selection differs from individual optimization.

Appendix E: Testing with LLM-Based Agents

An important open question is whether LLM-based agents exhibit similar limitations when deployed in strategic settings. Future work should:

Empirical LLM studies. Test actual LLMs (e.g., GPT-4, Claude) on sequential games with prompt variations simulating L1 (“follow your instinct”) vs L2 (“choose optimally”) reasoning. This would reveal whether LLMs’ strategic behavior aligns with our theoretical predictions or exhibits novel patterns.

Alternative theoretical frameworks. If LLMs do exhibit persistent bounded rationality, frameworks beyond Nash equilibrium may be needed:

- *Non-equilibrium settings:* Heuristic-based or satisficing agents where causal layer choices could affect outcomes
- *Learning dynamics:* How agents update strategies over repeated interactions
- *Evolutionary approaches:* Population-level selection where instinctive strategies might outcompete rational ones

Bridging theory and practice. Whether LLM-based agents require fundamentally new theoretical foundations—or can be modeled with appropriately modified classical frameworks—remains an open empirical question that our work helps to frame and motivate.