Review of

The Curse of Shared Knowledge: Recursive Belief Reasoning in a Coordination Game with Imperfect Information

Paper summary

The authors designed an experiment to investigate two research questions: (1) whether individuals can distinguish between *n*-th order shared knowledge and common knowledge, and (2) whether their behavior differs in environments where only *n*-th order shared knowledge is available instead of common knowledge. They find evidence suggesting negative answers to both questions.

The experiment, referred to as the "Canteen Dilemma," involves two players who must coordinate their actions and indicate their certainty about the other player's decision. Before each round, Nature determines the payoff structure of the game. Each player's payoff depends on the combination of both players' choices and their stated belief about the other player's action. The authors demonstrate that both an "all-office" strategy and cutoff strategies are equilibrium strategies in this game.

The experimental results show that the probability of choosing the canteen drops to around 50% when players arrive at 8:50, a time at which players also exhibit the highest uncertainty. Miscoordination peaks between 8:40 and 9:00. Prior to 8:50, players are more likely to coordinate on the canteen, while after 9:00, coordination shifts toward the office. Post-experiment surveys reveal that most participants do not clearly differentiate between common knowledge and *n*-th order shared knowledge, and that a majority identify 8:30–8:50 as a natural cutoff window for choosing the canteen.

Although the authors aim to address an interesting and important question, I have serious reservations about their experimental design, the specific game they ask participants to play, and whether these are adequate to answer the research questions they pose. In particular, the paper would benefit from a clearer and more precise definition of *common knowledge*, a detailed account of what type of *n*-th order knowledge is being shared in the experimental environment, and a formal discussion of the expected behaviors under each informational condition. Additionally, the authors should elaborate on how and why this particular experimental setup allows them to meaningfully distinguish between common knowledge and *n*-th order shared knowledge, as this connection currently feels underdeveloped.

Major comments

Clear Description of the Game and Formal Payoff Structure

The paper would greatly benefit from a formal and precise definition of the game before proceeding to equilibrium analysis. In particular, the authors should explicitly define the players' action sets and the payoff matrix. As it currently stands, it is unclear what constitutes an action in the proposed game. If each player's action is denoted as a_i {go to office, go to canteen}, then the payoff function presented in Appendix B should not include e_i (the elicited belief), unless beliefs are formally part of the action space.

If beliefs are indeed incorporated via incentivized belief elicitation, this should be treated as a separate mechanism, with its own incentives and theoretical justification. Furthermore, the authors should be careful interpreting elicited beliefs, as existing literature has shown that incentive mechanisms for belief elicitation do not always induce truthful reporting (Andersen et al., 2009; Charness et al., 2021; Schlag et al., 2015; Schotter & Trevino, 2014). A discussion of potential elicitation biases and robustness checks would improve the credibility of the belief data.

Andersen, S., Fountain, J., Harrison, G. W., & Rutström, E. E. (2009). Eliciting beliefs: theory and experiments.

Charness, G., Gneezy, U., & Rasocha, V. (2021). Experimental methods: Eliciting beliefs. Journal of Economic Behavior & Organization, 189, 234-256.

Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. Experimental Economics, 18, 457-490.

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. Annu. Rev. Econ., 6(1), 103-128.

In addition, the paper should clarify how this game structurally differs from other well-known coordination games — for instance, the **attrition game**. When players' arrival times are restricted to within a 10-minute window (which requires justification), the payoff structure for early rounds (e.g. before 8:50) appears analogous to an attrition game with the following matrix:

	Canteen	Office
Canteen	(a,a)	(c,c)
Office	(c,c)	(b,b)

Where a>b>c.

However, since the authors state the utility/payoff function as U(a1, e1, e2), I believe a player's action in each round is properly represented by a pair (a_i, e_i). If so, then the authors need to specify the combined action-belief payoff matrix for each time window. For

example, for arrivals before 8:50, the experimental interface implies a payoff matrix such as:

Player 1's Payoff Matrix Based on Experimental Interface for arrival time before 8:50

	Canteen	Office
Canteen, Very Certain	-0.02	-9.21
Canteen, Somewhat Certain	-0.58	-2.77
Canteen, Very Uncertain	-1.39	-1.39
Office, Very Certain	-9.21	-0.01
Office, Somewhat Certain	-2.77	-0.29
Office, Very Uncertain	-1.39	-0.69

Given this structure, it is not immediately obvious why certain strategies (e.g. *Office, Very Certain*) should be optimal without a formal proof, nor why others (e.g. *Office, Very Uncertain* or *Canteen, Very Uncertain*) can be ruled out. The equilibrium analysis section should be revised to rigorously address the strategic role of belief reporting in determining payoffs and equilibrium strategies.

Clear Definition of *n*-th Order Shared Knowledge and Common Knowledge

The paper currently lacks a clear operational distinction between *n*-th order shared knowledge and common knowledge. Due to the unclear description of the game, it is difficult to evaluate how the authors attempt to differentiate between the two. The authors should clarify whether there are observable behavioral markers or strategic thresholds that can meaningfully distinguish these knowledge types in the experiment. If not, they need to specify the conceptual and empirical criteria for concluding that players treated *n*-th order shared knowledge as common knowledge.

The authors might also enrich their theoretical background by revisiting the extensive literature on this distinction, particularly from the late 1990s and early 2000s (e.g., Geanakoplos, 1992). Additionally, it would be helpful for the paper to engage with the level-kkk reasoning and cognitive hierarchy literature, as this experiment may be capturing differences in agents' reasoning depth rather than a clean distinction between *n*-th order shared and common knowledge. Importantly, it is not even a common knowledge assumption in the experiment that all players are rational — this needs to be acknowledged and addressed in the design and interpretation.

Geanakoplos, J. (1992). Common knowledge. Journal of Economic Perspectives, 6(4), 53-82.

Justification of Experimental Design Choices

Several key experimental design choices require clearer explanation:

1) **Arrival Time Window**: The restriction that both players' arrival times must fall within a 10-minute window is a strong assumption. The authors should explain the theoretical or practical motivation behind this choice and discuss whether their equilibrium analysis and findings would hold without it.
2) **Repeated Game Structure and Matching Protocol**: The paper mentions that the game is played repeatedly, but does not specify the matching mechanism. If players are rematched, the mechanism (e.g., stranger matching) should be explicitly stated, as this has implications for belief updating and learning over time.
3) **University Game Payout Structure**: In the university's version of the game, a "lost-it-all" payment system was used rather than per-round incentives. This alters the strategic environment substantially. The authors should formally compare the two versions of the game — perhaps by including a side-by-side table summarizing key design elements and payoff structures — and discuss how these differences might affect coordination and belief formation.

Minor Comments

1. **Terminology: "Safe Strategy"**
   In game theory, a *safe strategy* typically refers to a strategy that maximizes a player's minimum possible payoff (the maximin strategy). Based on the payoff matrix outlined earlier, both (O,Not certain) and (C, Not certain) could qualify as safe strategies under this definition. However, it appears the authors use the term to describe a strategy that ensures coordination if both players adopt it. It would be helpful to clarify this terminology to avoid confusion with standard game theoretic usage.

2. **Clarification on "Curse of the Shared Knowledge"**
   The concept of a *curse of shared knowledge* is intriguing but underexplained. The authors should clarify its operational definition in the experiment and empirically quantify it by comparing outcomes under this scenario against those under an *all-office* strategy. Specifically, it would be informative to report whether, conditional on the other player's actual choices, the observed strategies result in systematically higher costs or lower coordination payoffs.

3. **Evidence of Learning Across Rounds**
   It would strengthen the paper to include evidence on whether players exhibit learning or strategy updating across rounds. Reporting whether coordination rates, payoff outcomes, or elicited beliefs converge or stabilize over time would provide valuable insight into dynamic strategic adjustments in this setting.

4. **Figure Design for Print**
   I recommend adjusting the figures to be black-and-white printer friendly. When printed in grayscale, several figures in the current draft were difficult to interpret. Using distinct markers, line styles, or patterns would improve the paper's accessibility in print.

5. **Clarifications on Lines 252–254 and Related Experimental Details**
   In lines 252–254, the claim about the *curse* and higher penalties lacks context without explicitly referencing the experiment. Additionally, upon reviewing the instructions, it remains unclear what is meant by *how uncertain you were*, particularly in cases of miscoordination. Is this experiment using a stranger matching protocol? Do participants know their arrival times are within 10 minutes of each other? Furthermore:

   - The text mentions five certainty levels, while Figure 3 appears to depict only three — this discrepancy should be addressed.
   - It might be valuable to include examples showing the cumulative consequences of repeated miscoordination (e.g., incurring losses over two or three consecutive rounds), as this would concretely illustrate the cost structure participants face.