Responses to review of “The Curse of Shared Knowledge: Recursive Belief Reasoning in a Coordination Game with Imperfect Information”

October 26, 2021

Abstract

This document includes the received reviews and editor comments as well as our responses (in italics). Changes made in the new version of the paper are listed in the responses.

I have received comments from reviewers with expertise in this area and also read your paper carefully myself. The reviewers’ comments are appended below. As you can see, the reviewers are quite mixed in their assessment of your submission. Reviewer 2 is highly critical of the work and recommends rejection, Reviewer 3 is overall positive and recommends acceptance with minor revision, and Reviewer 1’s assessment is somewhere in between. Let me also highlight that Reviewer 2 comments that their review should probably be discounted; I trust that you can weigh their feedback accordingly (I do think they raise some important issues, more on this below).

I, myself, much enjoyed reading your paper.

Thanks a lot!

I agree with Reviewer 3 that your paper is well written, well organized and that the experimental design is clever (also noted by Reviewer 1). I much appreciate the (meta-)theoretical points raised in your introduction, analysis, and interpretation, as well as the rigour of the mathematical treatment/proofs, the open data, and supplementary materials. I believe that your observation that people behave as if they have common knowledge even when they only have shared knowledge (and seem unable to improve their strategy) is of theoretical interest and relevance (even if not surprising, as noted by Reviewer 1)

Response 1. *Yes, indeed it may not seem surprising. However, several papers come to the opposite conclusion: they claim that we* do *act differently depending on whether we have shared or common knowledge [14, 10, 17, 16, 18, 3, 2] as cited in section 2 in our manuscript (pages 4-5). The difference between their and our results, we argue, is a difference in the experimental design: While the previous studies rely on a detailed explanation of the level of common or shared knowlege among participants in their group, our study does not make any such explications. Instead, in our design, the level of knowledge needs to be* deduced *by the participants. This, we believe, is an ecologically more valid approach, because people in real life situations are rarely explicitly told what is known by whom. Our methodology of using arrival times thus opens up the possibility to investigate more rigorously the vast space of shared knowledge dynamics among coordinating humans. We have included a discussion of these points in the new conclusion of the paper (the conclusion of the paper has been completely rewritten).*

and the manuscript is in my opinion, in principle, well-suited for a multidisciplinary cognitive science audience. The work productively combines logic, philosophy, and psychology to bring new theoretical insights with potentially wider implications for social cognition in real-world contexts.

Yes, thanks, that's also what we hoped to achieve.

That all said, all three reviewers also raise very important critical issues that make the manuscript not yet suitable for publication. I won't repeat all the reviewers' points here and refer to their reviews for more details. The most pressing issue in my opinion is the following:

> (1) Ecological validity: the approach taken raises the question to what extent the results are due to the artificiality of the task instructions. For instance, in real world settings coordination in the canteen would naturally be the preferred outcome, and one could argue that the payoff structure is artificially constructed to make this ecologically relevant outcome get less payoff in the experimental set-up. Also, to what extent is part of the computational overload due to the complexity of the payoff structure?

In a revision, I would like to invite you to address this major concern. How you wish to satisfactorily address this point, I will leave up to you. This could include collecting new data (as some reviewers suggest), but possibly additional analyses or adding clarifications and argumentation could suffice. However, you choose to address it, keep in mind that your audience will include readers who are highly skeptical that these kinds of artificial scenarios can teach us anything important about reasoning and coordination as it occurs 'in the wild'. My advice is therefore to spell out your argumentation as clearly as possible.

Response 2. ***Ecological validity/Canteen preference outside game****: We take the first point to relate to the connotations about canteen and office participants might have outside the game, which might affect their behavior inside it and be a confounding factor. We discuss this in response 27 as well and have included a discussion of it in the new Ecological Validity section of the revised manuscript, immediately before the conclusion.*

Response 3. ***Ecological validity/Artificiality of the payoff structure****: We are not sure what is meant here. In our experimental set-up, coordinating in the canteen* does *give the highest payoff. Page 6 of the original submission: "Penalties are tiered in such a way that a small penalty is deducted for successful coordination into the canteen (achieving the highest payoff), which is doubled for coordination into the offices (achieving the second-highest payoff), while the penalty for miscoordination or forbidden choices, i.e. going to the canteen at 9 am or after, is much larger (up to 921 times larger, meaning a significantly lower payoff than the previous two)." See Response 12 for further discussion.*

Response 4. ***Ecological validity/Computational overload****: We aren't certain whether the "complexity of the payoff structure" refers to the use of logarithmic scoring rule or the fact that there are three different payoffs to keep track off: successful coordination in the canteen > successful coordination in office > miscoordination. Concerning the use of the logarithmic scoring rule, we address this in more detail below (in response to reviewer #1). Concerning the general question of computational overload from not understanding the rules of the game (including the payoff structure), we have the following comments. If computational overload was sufficient to distract our participants, we would expect more participants making either random or completely uniform choices. While we do see some nonsensically choosing canteen at 9:00, the vast majority choose office at 9:00 and 9:10, and canteen at earlier arrival times, with many still choosing office at 8:50. Besides this, we see participants be much less certain of their canteen choices at 8:50 or 8:40 than at later arrival times. Thus nothing suggests computational overload, though of course no experimental results would be able to prove that with certainty. We have included a clarification in the Experimental Design and Discussion sections of the revised manuscript.*

Response 5. ***Ecological validity/Reasoning in the wild****: Ecological validity of our game obviously depends crucially on how common it is to experience situations in which there is shared knowledge to some degree, but not common knowledge—and where falsely assuming common knowledge leads to a bad outcome. As we address in Response 21, due to unreliabilities in attention, communication and interpretation, many everyday situations lead to shared knowledge in the disguise of common knowledge. In most everyday situations, the penalty for confusing shared knowledge for common knowledge is of course much lower than what is represented in our payoff structure. However, our payoff structure should then lead humans to pay* more *attention to the difference between shared and common knowledge in the game than in the wild. In other words, we should expect the confusion to be even more prominent in the wild. In the revised submission, we discuss this in the Ecological Validity section.*

Thanks for addressing these issues in your feedback, we believe the paper has benefitted significantly from adding the explicit discussions concerning ecological validity.

A second point that I think is crucial to improve upon is the following:

> (2) Wider Implications: The current Discussion and Conclusion sections seem underdeveloped and rushed. These sections do not yet make the theoretical contributions sufficiently clear and fail to situate the findings in a larger context (this in strong contrast to the high-quality Introduction). For instance, what exactly are the implications of the experimental results for theories of social cognition and for social cognition in real-world situations? See the reviews for more suggestions.

Addressing point (2) may in part also help to address point (1).

Response 6. *We thank you for this observation which we agree upon completely after reading the reviews and re-reading the manuscript. We have revised and expanded the conclusion and included a discussion about the wider implications of our method and results. We have tried to make as clear as possible our* *contribution to the field(s) and situated our findings in a larger context. Please have a look at the conclusion of the manuscript.*

Lastly, to help you navigate the conflicting reviews I'd like to note that, while I understand Reviewer 1's reservations about what we can infer exactly from the findings, I read your manuscript as making a theoretical point and presenting a relevant empirical observation, not merely a methodological advance (though I do believe the game scenario could prove a methodological innovation in the study of social cognition).

I appreciated the conciseness of the paper and do not recommend expanding the manuscript with more experimental details, unless they bear on the point(s) you wish to make and/or as they serve to address points (1) and (2) above.

Accordingly, I am inviting you to revise your paper and resubmit it for further consideration. If you intend to resubmit your paper, please explain how you have decided to deal with reviewers' and my comments, and follow the guidelines in the 'Resubmission Checklist' below. This will ensure that your revised paper is processed as quickly as possible.

Thank you very much.

–Minor points:

1. "The fact that humans are able to adopt [adapt?] their actions to whether they are . . . "

   Response 7. *Corrected.*

2. "Because of this, coordinating species typically use heuristic shortcuts in order to work with nested knowledge states like common [shared?] knowledge, such as joint perceptual cues and broadcasted signals . . . "

   Response 8. *Changed to "Because of this, coordinating species typically use heuristic shortcuts in order to reduce complex shared knowledge states, such as..."*

4. ". . . common knowledge is the [normatively?] preferred informational state for all members of the group . . . "

   Response 9. *Included.*

5. "We find it interesting to dig deeper into how humans would play and reason about such games in practice." – Articulate why this is of interest, scientifically. Personal interests aren't that informative.

   Response 10. *The sentence was redundant and has been removed*

6. Throughout it seems the word "group" is used to refer to a pair of participants. Maybe it is clearer to just use "pair"?

   Response 11. *Changed in all appropriate places.*

7. "Penalties are tiered in such a way that a small penalty is deducted for successful coordination into the canteen . . . " – It seems unnatural that successful coordination would result in a 'penalty'.

Response 12. *We understand that our logarithmic scoring rule may confuse as to what is the preferred outcome (i.e. has the highest payoff). A logarithmic scoring uses penalties subtracted from an initial endowment of $10 instead of bonuses added to an initial endowment of $0. Therefore, a low penalty corresponds to a high payoff, while a high penalty corresponds to a smaller payoff. As it is clearly visible in actual player choices, any initial confusion among players wanes off quickly after a few rounds, after which the players have a clear understanding of the pecuniary consequences of their actions. In addition, as also noted in Section 2, players can see examples of actions and their payoffs on every single page of the experimental interface. On page 6-7 in the revised manuscript we have included the following clarification: "Penalties are calculated by a logarithmic scoring rule, which depends on the decision and the certainty estimate by each player. Using penalties instead of bonuses may seem an unintuitive way to reward participants, but in the literature of scoring rules, logarithmic scoring rules are common and known to work well, due to their ability to ensure that loss minimization remains central, and that the player's actions represent their actual beliefs [7, 15, 13, 12]. In addition, we note that logarithmic scoring is a "strictly propper scoring rule" and generally does not create a cognitive overload. As noted in appendix B, we accordingly find a good match between certainty estimates and choices at arrival times different from those that are prone to miscoordination. This corroborates that players tried to minimize their losses and that they made their choices and certainty estimates as honestly as possible."*

8. In Table 1, would the number of pairs not be more informative than N?

   Response 13. *Added.*

Reviewer #1

This paper explores people's behavior in a coordination game in which players have to decide which of two actions to take. The ideal strategy in the coordination game requires understanding that an Nth order recursive definition of common knowledge would indicate that participants can never be sure what their partner is going to do. Thus, the ideal strategy is to always opt for the option that yields the best joint payoff and accept that there will be a few cases in which they will not get that ideal payoff.

Response 14. *We are a bit uncertain about what is meant. Does "ideal strategy" in both places refer to the game-theoretical optimal strategy? Or to what you would expect to see if you take into account that players are in general not able to understand the required higher-order reasoning? We take it that it means the first, but are not completely certain. We also take it that "best joint payoff" refers to the payoff received when both go to the canteen before 9am (the highest payoff possible in any round of the game). Given those assumptions, the ideal strategy is to always go to the office, which is* not *the option that yields the best joint payoff. However, it removes the risk of miscoordination. The alternative strategy of always opting for the best joint payoff—and accepting a few cases where you don't receive it due to miscoordination—leads to very low expected* *payoff due to the high penalty of miscoordination (as also discussed in the paper). As noted in the manuscript, if both players were fortunate enough to never arrive at 9 am or later, the ideal strategy would indeed be always to go to the canteen as this yields the highest possible payoff of* $9.90 *in ten rounds (provided that the player gives maximum certainty for successful coordination into the canteen). However, an all-office strategy would yield an almost as high payoff of* $9.80 *for ten rounds (provided that the player gives maximum certainty for successful coordination into the office), while simultaneously removing the risk of costly miscoordination. Thus, the always-canteen strategy is a high risk strategy not worth the extra ten cents.*

The experimental paradigm is clever, but I am left feeling unsatisfied with the contribution of the manuscript overall for a few reasons.

The idea that people will not always reflect on the degree of shared knowledge emerges from some of the referenced studies. Keysar's "illusory transparency of intention" studies suggest that participants often fail to take into account what various characters in a narrative know. Gerrig's extensions of these studies suggests that people can sometimes be coerced into putting more effort into determining what other people know when the context and task require it. So, the mere fact that people are suboptimal in tasks like this is not surprising.

Response 15. *We thank the reviewer for pointing out to us to the paper by Keysar (1994) and the extensions by Gerrig. We agree that it is not surprising that people have difficulties in determining what others know. The argument we make in our paper is however stronger. In our game, the players actually* do *reflect on the degrees of shared knowledge, but they just fail to understand the implications. So, we believe, in our case it is actually not about "participants failing to take into account what others know", and that they would have to be coerced into putting more effort into it. They are already putting a lot of effort into it, but still fail to separate shared knowledge of some degree from common knowledge. Let's explain this in a bit more detail. Each game consists of many rounds. A player might ignore the issue of shared knowledge in the first rounds, but when miscoordination arises, a player certainly has to reflect on how and why that miscoordination could occur. So playing many rounds of our game ought to coerce players into putting more effort into determining what other people know, since indeed our context and task requires it. You might then argue that players potentially don't understand that it is degrees of shared knowledge they should be reasoning about in order to avoid miscoordination and high penalties. But the answers to our post-game question in Figure 6A clearly indicate that players* do *reflect on the degree of shared knowledge. If they didn't, nobody would have a reason to state a cutoff point earlier than 8:50.*

A task like this one is interesting, because it yields the possibility of trying to model how much information participants are taking into account or to do manipulations to see whether that affects people's strategy. This study demonstrates that the methods can be used, but I don't feel like any strong conclusions can be drawn just from this work. Cognitive Science is not a journal that focuses on purely methodological advances, so the existence of the experimental procedure alone doesn't seem sufficient to warrant publication.

Response 16. *We believe that our contribution goes significantly beyond a methodological advance. Our main theoretical contribution is that non-explicated* *higher-order shared knowledge is treated like common knowledge in realistic settings where humans can't communicate and need to* deduce *the level of knowledge of others. This means, informally, that bottom-up sensory stimuli ("the time is before 9 am")* wins the perception battle *over top-down conceptual knowledge ("she could think that I think that she thinks ..."). And this is the case despite repeated error signals. We think this is of theoretical importance for various subfields within the cognitive sciences such as predictive processing, theory of mind, and artificial intelligence (HCI etc.), and needs to be investigated further. We have included our arguments for why our paper is of theoretical interest into the new conclusion.*

I do think this method would need to be extended in some way to support stronger conclusions than just that people do not unpack the recursive game completely. It would be interesting to create variants in which different orders of shared knowledge would be required to succeed to determine both how far people typically unpack shared knowledge as well as what kinds of contextual manipulations might lead people to be more or less likely to consider this recursive structure more deeply.

Response 17. *Recursive games already exist which are designed to determine how far people typically unpack shared knowledge [6], if this ability is reliably used [9], or if such unpacking is relevant at all to coordination games [1]. Such studies often operate with small sample sizes (67, 38 and 37, respectively in the three studies, with the latter two consisting of American University students). The purpose of our paper is somewhat different. The purpose is not to determine whether players can for instance distinguish* n*th order and* $(n+1)$*th order shared knowledge, but whether they can in general distinguish shared knowledge (of some order) from common knowledge. In our revised submission, we clarify the relation to this other research that specifically looks at the depths of shared knowledge employed by human agents, see Section 2 and also the footnote in the new conclusion.*

Reviewer #2

This review is a thoroughly dyspeptic one. It seems to me that the game scenarios are not capturing the important motivations behind the decisions. I encounter social coordinations problems in my own work, but I think it is crucial to focus on issues which can be well captured in one's scenarios and formalisms.

So the review should probably be discounted, but might be a useful warning to the researchers that perhaps these iterated problems have little demonstrated ecological validity—we are evolved to design them out perhaps?

Response 18. *We might indeed have evolved to do our best to avoid situations with shared-but-not-common-knowledge (e.g. via co-presence) due to the complexity of making the distinction. This might explain why we don't identify them everywhere in our daily life. But it is unlikely that we can entirely avoid situations with such epistemic make-up, and when such situations arise, it is likely we are not aware of it (due to the complexity involved). And then we do not have a chance to limit the potential downsides, if we do not take care to* *establish actual common knowledge, in situations with potential harmful consequences. This is now discussed much more thoroughly in the new Ecological Validity and Conclusion sections.*

Reading notes:

"On the way to the top, they both observe a thunderstorm approaching, but are uncertain about whether the other person has seen it. At this point they both know the fact that "a thunderstorm approaches", but don't know whether the other knows. In this situation, we would say that Agnes and Bertram both have private knowledge that a thunderstorm approaches, and since they both know it, it is also shared knowledge between them." [Shared knowledge is exactly what it isn't. It may be knowledge they share (unknowingly), but that is different. The terminology seems peculiarly ill-chosen?]

Response 19. *Terminologies are always debated and we agree that they can be ill-suited for their task. In fact, definitions of private and shared knowledge in the literature are not consistent. After lengthy discussions among the authors, we decided to choose the definition of shared knowledge that is standard in both (epistemic) logic and many of the sources in economy that we have been looking at. We also already addressed this in the supplementary material of the submission, as referred to in the introduction: "The exact border between private and shared knowledge vary significantly between different papers. De Freitas et al. [6] consider the case Ep to still only be private knowledge, and for p to be considered shared knowledge furthermore requires that there is at least one agent i knowing Ep to be true (that is, requires K_iEp to be true for some $i \in G$). The point of De Freitas et al. is that if only Ep is true, it is not really shared knowledge, but only private knowledge held by everyone in G. In our paper, we have sought a compromise between the terminology by De Freitas et al. and the standard terminology in epistemic logic, and hence we have the overlap between private and shared knowledge." We have now included a brief version of this discussion in the main paper itself to avoid the potential confusion.*

"This argument can be generalised to prove that even nth-order shared knowledge for any arbitrary large number of confirmations is insufficient for safe coordination." [This depends what policy has been agreed (or intuited). If they agree that any observation of bad weather which has been confirmed both ways is acted upon without further messages (no changes of mind) then they both turn for the bottom after the first exchange (with double +1 confirmation). And it would be reasonable to call it shared knowledge. They both know that they have agreed to abandon the climb, and they know they both know it. If the thunderstorm, suddenly gives way to bright sun, and they decide to try again, with double +1 confirmation, then they could do that. The knowledge would be different (of sun rather than thunder and at a different time) but it could be shared. There maybe needs to be an agreement about how long to wait for completing confirmation? 10 Minutes? Depending on the volatility of the contact? I don't see the 'little did he know about the little that she knew' iterations as adding anything to the decision making. I'm not sure that we should regard it as changing anything we care about. It sounds to me like paranoia. Or formalphilia?]

Response 20. *There is a misunderstanding here, the argument above seems to rely on a 100% certain transmission and reception of messages, which is* *different from what we assume in our scenario. Assume we try to solve the mountaineering example by the pre-agreement that observation of bad weather, with confirmation both ways, warrants going to the bottom. Assume we take the place of Bertram, communicating with Agnes. Bertram tells Agnes that he sees the bad weather, and receives a confirmation message. He also receives the message that Agnes sees the bad weather, and sends a confirmation message. Now he turns for the bottom. But given the unreliability of signals, Agnes did not get the confirmation message and stays alone at the top, not surviving. This cannot be solved by adding the requirement for another layer of confirmation messages, since the uncertainty of the second message could result in the same issue. Adding a timeout, like 10 minutes, also doesn't solve it, since the sender of the last of the required confirmation messages will after the timeout still not know whether that last message was successfully received.*

So, the issue lies in the unreliability of communication in the example. But this only shows that the example warrants a problematic epistemic structure, and does indeed not show that this is realistic communication, which we touch upon below, in Response 21. As our submission explains, the example is analogous to the coordinated attack problem (Byzantine generals problem), a famous example from distributed systems within computer science. There is a wealth of literature studying the formal properties and proving that coordination is not possible, see e.g. [5]. It has been widely studied exactly because it has real, practical implications for distributed computer systems: that certain properties cannot be guaranteed when communication channels between computers are unreliable (messages might not arrive, and communication is not synchronous).

I'm not saying that could never be important but it seems to me they need to work harder to show a case where it is. Yes, it's all good academic fun, but leaves me unconvinced. Maybe there are some nice lights to be thrown about what is realistically part of communication? Communication doesn't seem to be transmission of knowledge of the truth values of arbitrary propositions: if anyone ever thought that? With each of their iterations, the propositions concerned change, and it seems to me we are not interested in them after the first round of confirmations. There might be some other purpose for them, but it remains to be shown.

Response 21. *We can perhaps in most real cases assume that our signal or message (whether face-to-face or online) has reached the recipient. But, while the signal itself might reliably reach them, this is not sufficient for establishing common knowledge about it. They must also reliably interpret it* and *this reliableness must be common knowledge. So, depending on the context and complexity of the intentionality involved, I might not be sure that my communicated intention becomes common knowledge.*

Public broadcasting of messages, e.g. instructions given in a classroom to a class of students or an email sent by the department head to everybody in a university department, will also in practice not necessarily lead to common knowledge. Some of the students might not be paying attention at the moment, and some of the department employees might not have checked their email, or it might have ended up in their spam filter. Suppose that the message was regarding to all meet at a certain place at a certain time. Can we then be sure that everyone will meet at that place at that time? Obviously, no. Then the reply might be that we can just ask everybody to confirm that they received the *message. Well, yes, but then the same argument repeats for the confirmation messages: the recipient of the confirmations might not receive all of them or might briefly not pay attention. We now discuss this and its consequences in the new Ecological Validity section.*

footnote 1 ”in the sense of guaranteeing that the other person will also go there” [it will take a lot more than proof to guarantee any such thing.]

Response 22. *We take it that if the agents have common knowledge about the decision to meet at the base, they* will *both go there. That might still of course not be guaranteed in real life. But it just means that in real life coordination is even harder, as not even common knowledge can be guaranteed to be sufficient (if the agents are not perfectly rational).*

Top para 3: they seem to have met me before, but I remain unconvinced. As I said, there may be *some* knowledge and some situations where it matters to communication, but I don’t. see it here. It seems that the framework ought to start with that as a first question.
‘threa*d*s, bribes and . . . ’ → threats?

Response 23. *Corrected.*

Section 3: I don’t see why we’re doing an experiment on a so called coordination game, which has no messaging. It doesn’t seem accurately analogous to the intro example.

Response 24. *It is true that the canteen dillema is not “accurately analogous” to the intro example. We also don’t claim that it is. But the intro example is very well-known (at least in the coordinated attack version), so that’s why we started out with it. If formalising the two examples in (dynamic) epistemic logic, the difference becomes this: In the intro example, the initial model is simple, but a complex nested model is achieved via a number of message passings. In the canteen dillema, the complex nested model is already present in the initial state. So, indeed, the difference comes down to whether we have messaging or not, but our paper is not about the role of messaging for coordination, it is about the role of degrees of shared knowledge for coordination. And those degrees of shared knowledge can then be achieved in different ways. We don’t necessarily regard message passing as any particularly priviliged way to achieve shared knowledge of a certain degree. As we stated in the paper: “The electronic mail game and the mountain trekking example are complicated in terms of the dynamics of iterated message passing. In this paper, we devise a novel game in which the higher orders of shared knowledge are not achieved dynamically via actions, but are already present at the beginning of the game, using uncertainty about arrival times. This, we believe, makes the game easier to understand.”*

There is no consideration of the fact there is presumably motivation towards the canteen (other than monetary)—whether constant or not. Perhaps the ‘all office’ solution is boring—or a victory for the North Korean bosses?

Response 25. *This is a good point, also made by reviewer 3. Propensities towards going to the canteen may confound the results showing large fractions of players who choose the canteen. We refer to Response 27 below.*

Even without cell phones it should be possible to arrange for a signal to be placed in the entrance about arrival time, and an agreement on a cut-off point.

Response 26. *This is counterfactual. We look at situations where no reliable public broadcasting of signals is possible. The agreement on a cut-off point is insufficient, as noted in the paper. As described above, the experimental results are important in cases where shared knowledge exists but not common knowledge. And such situations can occur if we have unreliable signalling such as in the mountaineering example. Real-world examples of unreliable signalling is now further discussed in the Ecological Validity section.*

1 Reviewer #3

In this paper, the authors investigate empirically whether human participants are sensitive to the distinction between private, shared and common knowledge when making decisions in a context where this distinction has not been pointed out to them explicitly. They find that while participants appear to make choices that are consistent with them being aware of the difference between private and shared knowledge, they seem to conflate shared and common knowledge, or at least to make decisions that overestimate their shared knowledge as being common knowledge. The paper also offers, in the appendix, a formalization of these concepts and a formal demonstration of the optimal strategies for the setting used in the experiment.

I found the paper to be overall well written and well organized (the theoretical discussion, in particular, was presented in an engaging way). The experimental design was clever and clearly described, and I believe the results will be of interest to a diverse audience. Nonetheless, I shall raise in what follows a few issues (mainly regarding the interpretation of the experiment) which I believe it would be beneficial for the authors to address.

⇒ Regarding the experiment and its interpretation:

- One possible interpretation of the results, which was not discussed in the paper, is that "meeting for coffee" was seen by participants as the desired goal (both because it's intuitively more pleasurable and because the cover story suggests that the colleagues actually want to meet for coffee). This could be confounding their apparent disregard for the fact that cooperation under shared knowledge is unsafe: they could be accepting the risks associated with unsafe cooperation because it seems to be outweighed by the benefits of the social and enjoyable coffee meeting.

  Response 27. *The results were repeated as strongly when initially tested on people using an abstract version of the game with integers instead of arrival times, and having them coordinate an answer if both got a number below some fixed integer. In this abstract version of the game, there could be no "intuitively more pleasurable" choice, only choices based on the stated payoffs. We now discuss this in the new Ecological Validity section. We decided against the abstract version of the game to make the game easier to understand and increase the ecological validity.*

- The authors discuss (and demonstrate in the appendix) that the two best strategies are the "all office" and "cafeteria before 9am" strategies. They also point out that none of these strategies were adopted by players. However, it was not clear to me, and not discussed in enough detail, how the strategy adopted by participants was different from the "cafeteria before 9am" strategy. It seemed to me that participants' behavior was fairly close to such a cutoff strategy, so I would appreciate some further discussion in the paper to ward off any misinterpretations.

Response 28. *The "cafeteria before 9am" strategy is equivalent to the cutoff strategy with cutoff at 8:55 (or any time strictly between 8:50 and 9:00). This is not the strategy that the players follow, as our experimental results show. If the players were playing "fairly close" to that strategy, we shouldn't expect that the frequency of canteen choices at 8:50 is close to* $\frac{1}{2}$*. Also, canteen choices at 8:40 should then be expected to be extremely rare. It of course depends on what we consider as "fairly close", but if the players had actually understood that the "cafeteria before 9am" strategy is optimal (and that it would be reasonable to expect the other player to have formed the same insight), one should expect the actual player choices to be significantly closer to that strategy. As we show (see e.g. Figure 2), there is a lot of uncertainty regarding the choice at 8:50, which is not consistent with a cutoff strategy. Or, at least it would imply that there is uncertainty about the correct cutoff time, but then it still doesn't match the "cafeteria before 9am" strategy.*

Furthermore, Figure 6A shows that only 20% *of the participants believe that it is safe to go to the cafeteria at any time before* 8:55*, since only* 20% *believe* 8:50 *to be a safe cutoff point (where the question is formulated such that* 8:50 *is included as being safe).*

⇒ A few comments regarding the statistical analyses and visualisation of the data:

- given that the data included repeated measures, a multilevel (aka random effects) model would be better indicated.

  Response 29. *We agree with the reviewers suggestion and have changed the model to a mixed effects logistic regression with the arrival time and experimental condition(MTurk, DTU1, DTU2) as fixed effects including interactions and random effects for each pair of players, because the reviewer rightly expects pairs to behave differently, depending on which levels of shared knowledge they individually take into account in their reasoning and collectively try to agree upon. However, the new result, which is incorporated into the revised manuscript, is only marginally different from the simpler previous version. Confidence intervals are somewhat larger, indicating that the DTU-experiments indeed could be put together into one group (more specifically: the p-value is* 0.012 *showing weak significance. Additional ANOVA tests of various model formulas indicate that either the intercept or the slope could be reduced, however, but not both).*

- I would also raise the possibility of including in the main model (the logistic regression as a function of arrival pairs), as an interaction term, the experiment (MTurk/DUT1 and 3), which would allow a more quantitative comparison of the different trends in the responses. However, given that the conditions had unequal samples and that there were different numbers of trials, this could raise additional issues, and so I leave this suggestion to the authors' discretion.

Response 30. *Yes, the experiment matters, see answer above. Exploratory analysis including the round number as an additional fixed effect or interaction term did not show any significant contributions, meaning that players did not really change their strategy during the game. This again corroborates the conclusion that no behavioral learning took place during the game.*

- I appreciated that the code was made available in a repository and that the repository had some structure. However, it was not clear enough to where the models themselves were being fit as most of the code shared seemed to be geared towards generating the figures (but this could be due to my lack of Python fluency).

Response 31. *Yes, we intend to include more detailed comments in the scripts and codes when the paper is about to be finalized for publishing. A new R script for the mixed effects model has been included.*

- I would recommend qualifying somewhat the discussion of figure 2, as the trends in the figure are much more subtle than the textual description suggests. (Or, alternatively, it could be better supported by a quantitative analysis).

Response 32. *The discussion of Figure 2 has been moderated, and the figure has been slightly revised (adding information about the median choice by a white dot).*

- The readability of Fig. 3 could be improved, maybe by normalizing all bars to 100%, and maybe rethinking the order of the stacks (or the stacking altogether), because the red section are difficult to compare. The readability of Fig. 4 also could be improved somewhat, maybe by using a gradient color scale?

Response 33. *These are good suggestions and the plots have been changed accordingly in the resubmitted version of the paper.*

Jeg kan næsten ikke se forskel på 8:50 og 9:00 i figur 4 på min computer. Kan man ikke vælge nogen tydeligere farvegradienter?

⇒ A few final points:

- There is on page 3 a reference to Johnson-Laird's theory of mental models that is somewhat misleading. To my knowledge, Johson-Laird's theory uses the notion of mental models not to refer to models we might construct of other people's minds but instead to refer to the mental representation of possible worlds. As such, I don't think the reference is relevant to the discussion of higher order beliefs.

Response 34. *Thank you for pointing out a misrepresentation of Johnson-Laird's theory. We have removed the reference.*

- In the formal analysis in the appendix, I was not able to follow at the end of page 20 why the demonstration required the assumption of two subsets of arrival pairs, T1 and T2.

  Response 35. *It didn't* require *it, but it was made to simplify some of the reasoning in the proofs. The intuition is that the game splits into two disjoint subgames, one subgame with arrival times in T_1 and another with arrival times in T_2. We have now made this explicit already in the beginning of the section, and we now already from the beginning restrict attention to one of the subgames. This actually simplifies the technical discussion further, so thanks for raising the issue!*

- I found the end of the paper (the discussion/conclusion section) a little weak and I think it would benefit from a more in-depth or concrete (although not necessarily very long) discussion of the implications of the experimental results for existing theories of private/shared/common knowledge. Do they suggest any revisions to existing models? How are they compatible with previous results? In short, how should these results be situated, by the reader, within the current body of research?

  Response 36. *We agree and have improved upon the concluding section of the paper with discussions about related work, implications, limits, and perspectives.*

2 RESUBMISSION CHECKLIST:

- The deadline for resubmissions is 185 days from the notification of decision. If you will need more time, please contact us for an extension. If you do not intend to resubmit, please let us know.
- Please submit your revised paper via the Journal's online peer review system, Editorial Manager: http://www.editorialmanager.com/cogsci
- Include a separate document listing the changes made in the new version of the paper. This document should not reveal your identity.
- Avoid sending PDF files if possible as these can create problems in the final stages of the peer review process (PDF figures and tables are acceptable, however). Most word processing formats are supported.
- Visit the Journal's website (https://onlinelibrary.wiley.com/journal/15516709) if you would like more detailed information on the Journal's submission guidelines for new and revised papers.

Please be aware that if you ask to have your user record removed, we will retain your name in the records concerning manuscripts for which you were an author, reviewer, or editor.

References

[1] Oliver Curry and Matthew Jones Chesters. 'putting ourselves in the other fellow's shoes': The role of 'theory of mind'in solving coordination problems. *Journal of Cognition and Culture*, 12(1-2):147–159, 2012.

[2] Julian De Freitas, Peter DeScioli, Kyle A Thomas, and Steven Pinker. Maimonides' ladder: States of mutual knowledge and the perception of charitability. *Journal of Experimental Psychology: General*, 148(1):158, 2019.

[3] Julian De Freitas, Kyle Thomas, Peter DeScioli, and Steven Pinker. Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences*, 116(28):13751–13758, 2019.

[4] Harmen de Weerd, Denny Diepgrond, and Rineke Verbrugge. Estimating the use of higher-order theory of mind using computational agents. *The BE Journal of Theoretical Economics*, 18(2), 2018.

[5] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge.* MIT Press, 1995.

[6] Liesbeth Flobbe, Rineke Verbrugge, Petra Hendriks, and Irene Krämer. Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17(4):417–442, 2008.

[7] Irving John Good. Rational decisions. In *Breakthroughs in statistics*, pages 365–377. Springer, 1992.

[8] Trey Hedden and Jun Zhang. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36, 2002.

[9] Boaz Keysar, Shuhong Lin, and Dale J Barr. Limits on theory of mind use in adults. *Cognition*, 89(1):25–41, 2003.

[10] James J Lee and Steven Pinker. Rationales for indirect speech: The theory of the strategic speaker. *Psychological review*, 117(3):785, 2010.

[11] John Edensor Littlewood. *A mathematician's miscellany.* Meuthen and Company, 1953.

[12] Randall G McCutcheon. In favor of logarithmic scoring. *Philosophy of Science*, 86(2):286–303, 2019.

[13] Thomas R Palfrey and Stephanie W Wang. On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization*, 71(2):98–109, 2009.

[14] Ariel Rubinstein. The electronic mail game: Strategic behavior under" almost common knowledge". *The American Economic Review*, pages 385–391, 1989.

[15] Teddy Seidenfeld. Calibration, coherence, and scoring rules. *Philosophy of Science*, 52(2):274–294, 1985.

[16] Kyle A Thomas, Julian De Freitas, Peter DeScioli, and Steven Pinker. Recursive mentalizing and common knowledge in the bystander effect. *Journal of Experimental Psychology: General*, 145(5):621, 2016.

[17] Kyle A Thomas, Peter DeScioli, Omar Sultan Haque, and Steven Pinker. The psychology of coordination and common knowledge. *Journal of personality and social psychology*, 107(4):657, 2014.

[18] Kyle A Thomas, Peter DeScioli, and Steven Pinker. Common knowledge, coordination, and the logic of self-conscious emotions. *Evol Hum Behav*, 39:179–190, 2018.

[19] Hans van Ditmarsch and Barteld Kooi. One hundred prisoners and a light bulb. In *One Hundred Prisoners and a Light Bulb*, pages 83–94. Springer, 2015.

[20] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, Dordecht, 2008.

[21] Peter van Emde Boas, Jeroen Groenendijk, and Martin Stokhof. The conway paradox: Its solution in an epistemic framework. In *Proceedings of the third Amsterdam Montague Symposion*, pages 159–182, 1980.

[22] Rineke Verbrugge, Ben Meijering, Stefan Wierda, Hedderik van Rijn, Niels Taatgen, et al. Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment and decision making*, 13(1):79–98, 2018.