Responses to review of @@

@@

May 19, 2021

Abstract

This document includes the received reviews and editor comments (in black) as well as our responses (in red).

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)

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 [5, 3, 8, 7, 9, 2, 1] 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 seldomly have a clear understanding of 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 strengthened this point in the conclusion of the paper on page XX @@@

and the manuscript is in my opinion, in principle, well-suited for a multi-disciplinary 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.

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.

We thank the reviewers and the editor for stressing this important point and for making us work harder to be much more clear and precise about the aspect of ecological validity. We have done so via four improvements:

1) We have improved the description of the expermental design in section 3, stressing that coordination into the canteen indeed is the naturally prefered outcome and also is the outcome that has the *highest payoff* (i.e. lowest penalty) associated to it (see blue text in section 3 @@@). We understand that our logarithmic scoring rule which uses penalties instead of bonuses may confuse reviewers as well as players as to what is the prefered outcome (i.e. has the highest payoff). But as it is clearly visible in actual player choices, any remaining 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 & their payoffs on every single page of the experimental interface.

2) In regard to the question of a potential computational overload due to the complexity of the payoff structure, we found no such overload in our data.

As seen in figure 4, the mean frequencly of canteen choices in the first couple of rounds do not form outliers, indicating that players indeed did understand the payoff rules well enough even in the first couple of rounds. One exception to this observation can be made for the arrival times 9:00 and 9:10, which show a slightly larger fraction of players chosing the canteen in early rounds compared to later rounds, even though the instructions clearly stipulate that they always should go to their offices when arriving at 9 am or later. Therefore, we can conclude that some players in the beginning of the game indeed did not understand the rules about when they should go to the office. This, however, does not change any of the results nor conclusions in our paper relating to the player's willingness to chose the canteen at early arrival times, and thus being pray to a 'curse of shared knowledge'. Of course we clearly acknowledge that a small fraction of players did not understand the rules about when to go to their office, and accordingly we have included a sentence about this in the conclusion. (@@@)

3) Our choice of a logarithmic scoring rule may be unfamiliar to a lot of people, but in fact, it has been shown that logarithmic scoring rules are the best choice for ensuring that the player's actual actions represent their beliefs [6, 4]. As noted in appendix B, we accordingly find a good match between estimates of what their colleague may do and their actual choices at arrival times different from those that are prone to miscoordination. This corroborates that loss minimization remained a central concern for the players and that they made their choices and estimates as honestly as possible.

4) Still, we wish to emphasize that ecological validity was one our of main reasons to develop the game. We believe that the canteen dilemma demonstrates a novel mechanism by which researchers can experimentally investigate shared knowledge dynamics, omitting the tedious (and artifical) explications of who knows what when. Suspecting that this point was not made clear enough in our initial manuscript, we have clarified it with an additional paragraph at the end of section 2. (@@@)

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).

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.

–Minor points:

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

2. "Because of this, coordinating species typically use heuristic shortcuts in order to work with

3. nested knowledge states like common [shared?] knowledge, such as joint perceptual cues and broadcasted signals . . . "

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

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.

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

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'.

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

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. 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.

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.

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.

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?

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?]

"This argument can be generalised to prove that even nth-order shared knowledge for any arbitrary large number if 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?]

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.

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.]

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?

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. 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? 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.

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.

- 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.

⇒ 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.

- 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.

- 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).

- 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).

- 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?

⇒ 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.

- 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.

- 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?

References

[1] 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.

[2] 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.

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

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

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

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

[7] 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.

[8] 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.

[9] 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.