Responses to review of @@

@@

May 31, 2021

Abstract

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

Answer: 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 [6, 3, 9, 8, 10, 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 @@@

PLAN: Robin makes the first attempt to address this in the conclusion: That you might have reasons to say that it is *not* surprising (since we generally struggle with higher-order reasoning, and we are not always taking it into account even if we could), but then on the other hand a lot of literature suggests that it *is* surprising.

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.

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. 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. 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.
>
> 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)."
>
> Also, to what extent is part of the computational overload due to the complexity of the payoff structure? 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.
>
> PLAN: Nicolet integrates discussion of the possibility of "computational overload" in the paper where relevant (in discussion, before conclusion). Perhaps integrate this sentence: "This, however, does not change any of the results nor conclusions in our paper relating to the player's willingness to choose the canteen at early arrival times, and thus being pray to a 'curse of shared knowledge."
>
> START HERE ON MONDAY 31 MAY 2021!!!!

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.

Answer:

3) Og måske her og i artiklen: Vi har tidligere eksperimenteret med andre scoring rules som leder til samme udfald.

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 this with an additional paragraph at the end of section 2. (@@@) Selvom det vi gør måske er bedre end hvad mange andre gør, så tror jeg stadig mange vil have det svært med vores ecological validity. Og vi kan nok desværre heller ikke argumentere for vores ecological validity bare ved at sige at vi er bedre end visse andre. Jeg mener stadig vi har et problem her med at argumenere for at de fænomener vi studerer i eksemplet kan overføres til nogen som helst situationer i den virkelige verden.

PLAN: The plan here is to address everything related to ecological validity. We here refer to a discussion of ecological validity in the conclusion/discussion of the paper.

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).

Answer: @@@@

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).

We agree.

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 . . . " 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 . . . " Changed to "Because of this, coordinating species typically use heuristic shortcuts in order to reduce complex shared knowledge states, such as..." Den nye formulering er mindre klar. Hvad betyder det at reducere en kompleks shared knowledge state? Var den oprindelige formulering ikke korrekt?

4. ". . . common knowledge is the [normatively?] preferred informational state for all members of the group . . . " 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. Changed to "However, it is important to understand how humans in fact reason in such a situation as it may provide profound insights into human abilities to coordinate without communication."

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

We understand that our logarithmic scoring rule may confuse as to what is the preferred outcome (i.e. has the highest payoff). Our 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. 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.

PLAN: Here and in the paper itself we add a lot of references to succesful uses of logarithmic scoring. Robin does this. Here we then say: This is standard and usually works. It doesn't usually create a cognitive overload. Robin also tries to integrate the following formulation above, depending on relevance: "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 [7, 5]. 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."

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

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 ideal 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 (see pages @@@). As noted on page 8 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.

Answer: We agree that it is not surprising that people have difficulties in determining what others know. The fact that our game is recursive compounds this problem. We thank the reviewer for pointing out to us to the paper by Keysar (1994) and the extensions by Gerrig.

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 could 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.

Answer: We believe that our contribution goes far beyond a methodological advance. Our main theoretical contribution is that non-explicated higher order shared knowledge is treated like common knowledge in any realistic setting where humans can't communicate and need to *deduce* the level of knowledge of others. This means, informally, that induction beats deduction: 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 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.

Answer: We profoundly agree with the reviewer that the method needs to be investigated further and, if possible, extended to support additional conclusion about human abilities to navigate shared knowledge spaces. We are intrigued by the reviewer suggestion to create variants in which different orders of shared knowledge are required to reach a certain goal. This would be very interesting to look at in future work. We include a paragraph about these possibilities in our conclusion of the paper (page @@@@).

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?

Answer: We appreciate this honest assessment, and point to our previous answer regarding the ecological validity of our experiments.

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?]

Answer: 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. Therefore, we go to great length in the manuscript to tease apart the various understandings, and we also carefully define private, shared, and common knowledge in the introduction as well as in appendix C, and give a formal definition in the supplementary material. 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. According to these, we have

Jeg kan ikke se at vi definerer det i appendix C?

Alternative response: 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."

"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?]

Answer: This argument relies upon 100% certain transmission and reception of the message. Without this, the suggested protocol would be risky, e.g. "I go down if and only if I get a confirmation reply to my message". Then you get a confirmation, send yourself one back, and go down, hoping that the other gets your message. The other then survives with a probability $p \neq 1$ which is equal to the probability that the other gets the same message correctly.

There might have been a misunderstanding here. Assume we try and 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 Bob, communicating with Anne. Bob tells Anne that he sees the bad weather, and receives a confirmation message. He also receives the message that Anne sees the bad weather, and sends a confirmation message. Now he turns for the bottom. But giving the unreliability of signals, Anne 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. So, the issue lies in the uncertainty 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.

Jeg kan godt lide Nicolets variant og ville tilføje: As the paper 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. [Fagin, Halpern, Moses & Vardi, 1995]. It has been widely studied exactly because it has real, practical implications for distributed computer systems: that certain properties can not 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.

That communication is not transmission of truth values of propositions is a salient point in the interplay of logical analysis and cognitive science. Even if it was, it would not be a complete account of communication, since the lack of logical omniscience would still make real communication fall short of formalized analysis. That communication involves non-propositional embodied knowledge complicates this further. Continuing from this, we can look at the problem in the paragraphs above, which might have misunderstood that the communication in the mountaineering example involves unreliable signaling. The argument then goes that this is unrealistic, due to reliableness of real face-to-face communication or even modern cell signalling. This is true, that we can in most real cases assume that our signal (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. The complexity might consist in the fact that the entities making up signals can be multifaceted too, i.e. choice of words, tonality, body language et cetera, besides propositional content as the reviewer points out. In other words, much like we are not robots in terms of passing propositional truth values around, and deducing from these, we are not robots in terms of interpreting the complex signaling of others, and as such, even face to face communication is technically unreliable, and does not necessarily warrant common knowledge, outside stylized examples. 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. (kunne måske også bruges som eksempel i selve artiklen?) Note that, as we mention humans are specialized in navigating real social situations, in fact so good that it might be unrealistic for robots to emulate it.

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.]

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? 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.

Answer: Coordination without messaging is a serious field of research in, for instance, human-computer interaction, artificial intelligence, in animal behaviour (stigmergic coordination), and in other multi-agent system.

Robin's svar adresserer ikke direkte reviewerens kommentar. Alternativt forslag: The canteen dillema is not "accurately analogous" to the intro example, no. 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?

Answer: 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 our response to reviewer 3 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.

Answer: This is a counterfactual. We look at situations where no broadcasting of signals is possible. The agreement on a cut-off point is insufficient, as noted in the paper.

A coordination game without messaging allows us to test how humans try to coordinate with each other when they have some level of shared knowledge, but not common knowledge. It might be relevant to notice, if our conclusion holds, the relevance of the fact that shared knowledge is hard to discern form common knowledge, is obfuscated by this very fact. That is, the relevance of the difference can be hard to see, because the difference is so hard to make out in the first place. And our experiment shows that we do not intuitively see the difference, even when it makes an operative difference.

The point that there might be a motivation towards canteen rather than office due to word connotation could be a confounding issue, and will be discussed .

The results were repeated as strongly when initially tested on people using integers instead of arrival times, and having them coordinate an answer if both got a number below some integer. We decided against this formal version in favor of the canteen formulation, exactly for reasons of ecologoical validity. The question remains of of ecologoical validity: Can the difference between shared knowledge and common knowledge make a difference in the real world?
Jeg foreslår at vi i selve artiklen diskuterer de tidligere varianter af spillet, og så refererer vi her i stedet til disse nye tilføjede diskussioner. Altså: Det som Nicolet skriver ovenfor skal ekspanderes og gøres til en del af den reviderede submission. Jeg kan se at Robin længere nede faktisk har ekspanderet beskrivelsen af eksemplet. Så vi kan evt tage Robins version af eksemplet og proppe ind i artiklen og så både referere til det her og længere nede.

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. We argued above that reliable signaling is not enough for common knowledge (since after all, even if our signals reach their recipients reliably, people are not computers which our signals simply stream into to). That is, we both require reliable signaling and reliable interpretation (and in fact recall too, but that is generally not an issue in short time spans. In other words, common knowledge might fade away due to limited memory). And reliable (i.e perfect) interpretation is an idealized view of reality, just as much as logical omniscience is. The question still holds if this theoretical possibility is an actuality. Cases where this is especially pertinent are situations where parties communicate but dont "play with open cards" as it were, and could include situations such as diplomatic negotiations and romantic communication. I will describe such a situation below:

Hvis vi skal inkludere lange udregninger i vores svar, så skal det hellere være

tilføjelser til selve artiklen. Et langt svar på et review uden tilsvarende ændringer af artiklen er et dårligt tegn, for det betyder at det som man forklarer til revieweren åbenbart ikke er en del af selve artiklen, og så hjælper det jo ikke. Ergo: Hvis meget af dette skal beholdes, så skal det i stedet integreres i artiklen og vi skal herfra blot referere til den pågældende nye diskussion. Selve reviewer-svarene bør ikke være alt for lange.

Suppose the following case of romantic interaction between Anne and Bob. Suppose that either can make an advance towards the other, and we might categorize that both doing this is a 'succesfuld coordination' and that if only one does it, its a form of 'miscoordination'.

Måske skulle man gøre det mere konkret, så det bliver helt klart at det faktisk også er et koordineringsspil. Fx at Anne og Bob sidder og snakker og kigger hinanden i øjnene, og så er der en lille pause. Her kan de så hver især vælge at læne sig frem til et kys eller ej. Hvis kun én gør det er det i bedste fald pinligt, i værste fald vil de miste deres venskab, fordi den ene ny ved at den anden har romantiske følelser.

Der er forresten også et afsnit af Friends hvor de itererer på shared knowledge, jeg kan ikke huske om jeg tidligere har delt det:

https://www.youtube.com/watch?v=LUN2YN0bOi8

This is a somewhat high stakes game, in the sense that successful coordination can lead to lifelong relationships, and that miscoordination is akward at best, and sexual harashment at worst. Måske lidt voldsomt at kalde det "sexual harashment", da det lyder meget voldsomt. Måske skulle man i stedet skrive "kan opleves krænkende" el.lign? Suppose that Anne is not interested in Bob, but Bob is interested in Anne. This might be analogous to Anne arriving at 9:00 and Bob at 8:50 in our game. Bob might make a move on Anne here, which is deemed to be a miscoordination. Depending on the discomfort this causes Anne, we might generally classify miscoordination as a moral failure (i.e. Bob ought not to have done this). Now suppose we change the situation, such that Anne is interested in Bob as well. For Bob, the situation is the same, and if we said before that Bob ought not to make a move on Anne, the same applies here. And so we can still imagine it discomforts Anne, since from her perspective, Bob would have done so regardless of her interest.

Det er ikke sådan jeg ser analogien. Jeg tænker på det som et rent koordinationsspil: enten læner man sig frem til et kys eller man lader være. Hvis begge læner sig frem får man max belønning. Hvis man er miskoordinerede er det max straf, og hvis ingen læner sig frem er det neutralt. Altså den slags belønningsstruktur som vores resultater gælder for. Så problemet er ikke at Bob alene vælger at erklære sine romantiske interesser og at Anne synes det er upassende, problemet er at det er i forbindelse med et muligt kys, så det er et koordineringsspil. I din fortolkning er det kun problematisk hvis man mener det moralt er problematisk at erklære sine romantiske følelser for én som ikke er interesseret i én. Eller, det kommer selvfølgelig an på hvordan man lægger an på den anden, om man fx slikker vedkommende i øret under en dans til en julefrokost. Men det er under alle omstændigheder en farlig og upassende strategi, og jeg vil helst have at eksemplet er så analogt til kantinedilemmaet som muligt. I det asymmetriske tilfælde hvor én agent kan vælge at lægge an på en anden, er problemet kun det moralske. Hvis der er shared knowledge til degree 1 at Anne elsker Bob og Bob

elsker Anne, så er det tilstrækkeligt til at de hver især ville kunne erklære deres romantiske følelser (hvis vi ser bort fra det moralske). Altså opstår problemet med forskellen på shared og common knowledge ikke, eller i hvert fald kun hvis man køber premisen med det moralsk uanstændige, men så er det jo et spil som er umuligt at vinde. Og det med det moralske er jo også noget som kræver selvstændig retfærdiggørelse som du har herunder med lejlighedseksemplet.

This is similar to a friend of mine entering my apartment while I'm not home, without having read my message inviting him, and I would be right to be upset (I might not be, but it would not be unreasonable if I was). Now suppose we change the situation further, such that Bob comes to know that Anne is interested in Bob. Now, they are both interested in each other, and Bob knows this fact (it is shared knowledge for him). So Bob might make a move on Anne, but, again, from the perspective of Anne, nothing has changed, amd might think Bob is unjustified in doing so (thus being a moral failure again). This is analogous in our game to Anne arriving at 8:50 and Bob at 8:40, knowing that the best is for both to go to the canteen, attempting to do so, but resulting in miscoordination. I min variant ville jeg sige at vi argumenterer som i kantine-dilemmaet. Det er klart usikkert for dem at læne sig frem til et kys, hvis kun den ene er interesseret. Så det er heller ikke nok at begge er interesseret, for de ved ikke nødvendigvis at den anden er. Tag så første-ordens shared knowledge at de begge er interesseret. Hvis de *kun* har første-ordens shared knowledge, så vil de se det som muligt at den anden ikke har knowledge, og derfor kan den anden være i den oprindelige situation hvor den optimale strategi er *ikke* at kysse. Altså bør de heller ikke kysse på første-ordens shared knowledge. Argumentet kan så generalises til vilkårlig orden af shared knowledge, som i kantine-dilemmaet. Man kan selvfølgelig lave en probabilistisk vurdering så man ved hvad man risikerer ved at læne sig frem, men man kan kun være *garanteret* ikke at få den store straf ved at holde sig fra at forsøge at kysse den anden. Vi kan stadig sammenligne med tidspunkterne, selvfølgelig, det er bare et spørgsmål om at blive enige om hvad reglerne er for dette romantiske spil. It is especially in such situations, where it is realistically plausible that people do not communicate their intentions openly, such that some shared knowledge exists, but not common knowledge, like in the mountaineering example. If we think that we can establish a plan ensuring coordination, without common knowledge, we will succeed sometimes (when we both fall on the right side of the threshold), but fail when falling on either side of the cut-off point. In summary, regardless the level of shared knowledge, Bob cannot make a move on Anne without the possibility of miscoordination. This does not entail that Bob ought never make a move on Anne, but rather that in order to remove the possibility of miscoordination, intentions must be made common knowledge, and not simply shared knowledge to some depth. In conclusion, if we do not take seriously the fact that we do not properly distinguish between common knowledge and shared knowledge of some level, we cannot hope to mitigate potential harmful effects arising from it. Such remedy could be emphasizing open and honest communication in situations where it matters.

Ja, det er fedt. Det skal klart være med i artiklen, hvis vi kan blive enige om en fornuftig variant. Og så skal det måske mere tydeligt indgå som vores respons på ecological validity i vores reviewer response, allerede der hvor Iris første gang peger på dette.

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.

  Answer: This is an important point, also noted by the second reviewer. We believe that an 'intrinsic desire' to meet in the canteen is not a confounding factor to the observed behaviors. We see two main resons for this: 1) the logarithmic scoring (being a *proper* scoring rule [7, 5]) secures that loss minimization is the primary concern for the participants, which can be corroborated by the strong coupling between participant belief (measured via their certainty estimates) and their actions (their actual choices), see figures 1 and 2 in the paper. Jeg kan ikke se hvordan dette kan være et argument imod at de "accept the risks associated with unsafe cooperation because it seems to be outweighed by the benefits of the social and enjoyable coffee meeting." Jeg tænker reviewerens påstand her er at de ikke klart kan skelne deres scoring rule fra deres hverdagsforståelse for hvad der er et godt udfald. Ingen scoring rule kan være et argument imod den påstand. 2) Other versions of our game exist in the literature, and they have been shown to recreate the same tendency by players not to think deeply about what their partner thinks. More specifically (also shortly noted in the experimental design section), our game is inspired by the structure of the consecutive number riddle, also called the Conway paradox, see e.g. [13, 11]. It exists in many different formulations, going back at least to Littlewood [4]. Assume two players receive a card with a number between 1 and 10 (both included). They can only see their

own card, but are informed that the two numbers are consecutive, e.g. one player gets 3 and the other 4. Now it can either be a game concerning whether the two agents know if they have the highest card, it can be a game concerning guessing the card of the other player, or it can be a game to determine whether both cards have a value strictly below 10. In most versions of the game or riddle, the point is that there is a lack of common knowledge between the two agents. Independent of the numbers received, there is no common knowledge that both numbers are strictly below 10 [12]. If player a receives a 1, she knows that player b will have received a 2, but then player b needs to consider it possible that a received a 3, and if a received a 3, a will consider it possible that b received a 4, etc. This alternating chain of each agent reasoning about what the other considers possible is of course structurally similar to what we observed in the mount trekking example as well as in the canteen dilemma. There have not been any systematic experimental investigations of how normal people play the consecutive number riddle, but informal tests and games in classrooms (personal communication) clearly show that a vast majority of students only engage in the required recursive reasoning to the most shallow depths. Note that, in this formulation of the game, there is no intrinsic desire to reach one number rather than the other. In our game, the card numbers have been replaced by arrival times at a workplace, and instead of trying to assess whether both cards are below 10, the goal is to assess whether both colleagues have arrived before 9 am, in time for a cup of coffee.

- 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. The players' behaviour is indeed close to a cutoff strategy, just not a cutoff strategy with the cutoff at 8:55. Skal vi tilføje et par sætninger om det? Jeg synes det burde være klart i artiklen. Lad os holde øje med det når vi læser artiklen igennem, og så tilføje en sætning om nødvendigt.

⇒ 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 will look into that

- 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 will look into that

- 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).

  Answer: Yes, we intend to include more detailed comments in the scripts and codes when the paper is about to be finalized for publishing.

- 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). Jeg har ikke så meget erfaring med violin-plots, så jeg har svært ved at vurdere det. Vi skriver: "It clearly shows that it is exceedingly rare for any of the participants to consider it problematic to go to the canteen when arriving early. Arriving at 8:30 or earlier is deemed sufficiently early to visit the canteen with very high confidence." Men jeg ved fx ikke hvor "exceedingly rare" det er, det ved jeg ikke hvordan man aflæser af et violin-plot. Så spørgsmålet er her: Er vores konklusioner for dramatiske, eller er det fordi vi bør rapportere i et andet format end et violin-plot, hvor vores konklusioner bliver mere tydelige? Vi har sikkert haft noget andet end violin-plots på et tidligere tidspunkt.

- 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?

  Answer: This is a good suggestion and has been changed in the resubmitted version of the paper. (@@@@)

⇒ 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.

  Thank you for pointing out our 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. Jeg kigger på det. Det er for at kunne dele op i to uafhængige delspil, men om det er strengt nødvendigt må jeg lige tjekke igennem mine beviser for at afgøre. Jeg husker det som at det gjorde tingene simplere, man skal lige være sikker.

- 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?

  Answer: We agree and have improved substantially upon the concluding section of the paper with discussions about impications, limits, and perspectives. See also the answer given to the editor's point 2. (@@@@)

Kunne man overveje at gemme romantik-eksemplet til diskussion/konklusion som en del af diskussionerne af ecological validity? Hvis vi alligevel skal udvide afslutningen var det måske et passende sted at adresser både ecological validity og wider implications?

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.

Gershman, Harvard University, USA James A. Hampton, City University London, UK Janet H. Hsiao, University of Hong Kong, HK Fiona Jordan, University of Bristol, UK Yasmina Jraissati, Ronin Institute, USA Sangeet Khemlani, Naval Research Laboratory, USA Pia Knoeferle, Humbolt University of Berlin, Germany Max M. Louwerse, Tilburg University, The Netherlands Nicole M. McNeil, University of Notre Dame, USA Daniel Mirman, The University of Edinburgh, UK Padraic Monaghan, Lancaster University, UK Kinga Morsanyi, Loughborough University, UK David C. Noelle, University of California, Merced, USA Pamela Perniss, University of Cologne, Germany Veronica Ramenzoni, National Council for Science and Technology of Argentina (CONICET) and the Pontifical Catholic University Santa María de Buenos Aires (UCA), Argentina Oron Shagrir, The Hebrew University of Jerusalem, Israel L. James Smart Jr., Miami University, USA Michael J. Spivey, University of California, Merced, USA Sashank Varma, University of Minnesota, USA Jiaying Zhao, University of British Columbia, Canada

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

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

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.