The canteen dilemma: Higher-order social reasoning in a coordination game with imperfect information

Thomas Schrum Nicolet

May 30, 2019

Abstract

Social interaction often depends on the ability to correctly attribute mental states to others. This capacity is traditionally referred to as theory of mind and is a fundamental mechanic behind social dynamics. The canteen dilemma is a game aimed at testing how this ability is drawn upon in coordination problems without common knowledge. Findings[1] from the experiment suggest that adults are indeed limited to second-order reasoning at most. Most importantly, the results suggest that the higher-order beliefs of participants exist in various degrees not encapsulated by traditional descriptions of n or $n+1$ theory of mind. The present paper also introduce research on social cognition and discuss certain conceptual issues within the field as give a case example of the importance of social cognition.

Acknowledgments

The canteen dilemma was conducted as part of a research project at the Center for Information and Bubble Studies (CIBS) in collaboration with R. Engelhardt (CIBS) and T. Bolander (DTU Compute). I would like to thank both the center for making the experiment possible and R. Engelhardt and T. Bolander for co-designing and implementing the experiment as well as providing insightful discussion. I would also like to thank M. B. Andersen for help with technical issues and V. F. Hendricks for enabling me to work on the project.

[1] See https://github.com/thomasnicolet/canteen-dilemma-thesis for the Python code which makes up the data analysis of the experiment.

Contents

1 Introduction

We often have to attribute mental states to others in order to successfully predict and understand their behavior. This cognitive capacity plays an integral role in social interaction and is therefore a fundamental mechanic of social dynamics. The cognitive capacity in question has traditionally been referred to as 'theory of mind' and has seen immense scientific attention since Premack & Woodruff [70] asked the question "does the chimpanzee have a theory of mind" in their seminal paper of 1978. The canteen dilemma is a game aimed at testing how this cognitive capacity is drawn upon in coordination problems. The canteen dilemma is a two player coordination game with imperfect information devised and implemented as part of a research project at the Center for Information and Bubble Studies in collaboration with Robin Engelhardt and Thomas Bolander. The structure of the game emphasize reasoning about the possible beliefs of others, including their reasoning about one's own and so on. The term 'theory of mind' most often refers to the ability to represent mental states of others such as beliefs or intentions but has been used in various different research settings since its inception. The term has also been used somewhat interchangeably with 'mentalizing' [49], 'mind-reading' [50, 85], 'mind perception' [43], 'perspective taking' [73] and 'social intelligence' or '(higher-order) social cognition/reasoning' [4, 84]. The heterogeneity of the research studying this cognitive capacity has researchers like Schaafsma et al. [76] to call for a reformulation of theory of mind. Since terminological debates are outside the scope of this article, the various terms are viewed synonomously. Because this article is focused on making decisions depending on reasoning about the possible mental states of others, the favored terms will be higher-order social reasoning or cognition or even higher-order reasoning.

This paper attempts to focus on the nuanced complexities within social cognition. It specifically argues for a more nuanced picture of cognitive diversity than traditional descriptions allow, which often state that some demographic have a capacity to reason up to n order of social reasoning but not $n+1$. The possible lack of introspection into sociocognitive limits is also discussed as part of sociocognitive limitations.

The canteen dilemma experiment involved two players, each assigned an arrival time between 8:00 am and 9:10 am told their time is always 10 minutes apart. Participants then had to make matching decisions between the office and the canteen, while told that they have to go to the office if either arrive at 9:00 or later. Their decisions therefore involve considering the possible arrival time of the other player and possibly their considerations about the arrival times and so on. Our findings suggest that participants exhibited social reasoning in details which might not be encapsulated by n or $n+1$-orders of social reasoning. Our results also indicate a lack an introspection into limits of social reasoning.

1.1 Conceptual clarity

Some conceptual remarks are necessary due to research on social reasoning possibly being unclear in terminology (see also Section 1 of the accompanying introduction). Schaafsma et al. [76] have presented some healthy skepticism concerning studies on social cognition. Schaafsma et al. argue that the term 'theory of mind' has been used vaguely and inconsistently The authors call for a deconstruction of theory of mind into a comprehensive set of basic components, after which a scientifically tractable concept can be reconstructed. While such a project is outside the scope of this thesis, it is important to note that the current scientific picture of the mental capacity associated with theory of mind might be incomplete or lacking in precise terminology. Verbrugge [84] also mentions that

it is possible that theory of mind is not a uniform mental ability, but that different applications of social cognition constitutes distinct cognitive abilities.

The issues raised by Schaafsma et al. suggests that social cognition could be made up of a plethora of cognitive processes. This is already indicated by some of the debates within the field. Take for example discussion about *nativism* and *constructivism* regarding social cognition, arguing whether the capacity to represent the mental states of others is an innate genetically inherited capacity, or whether it is a socially developed and culturally inherited ability. This distinction is mirrored in discussion concerning differences in implicit/explicit and spontaneous/reflective applications of social reasoning.

Heyes & Frith [50] has argued that theory of mind is made up of implicit and explicit social reasoning. The implicit social cognition is a genetically inherited predisposition produced by natural selection, while the other is explicit representation of the mental states of others, much like print reading, a culturally inherited skill. Print reading does depend on cognitive predispositions but these were not developed for this purpose. Heyes & Frith argue that the implicit system develop early and works in a fast and efficient way, whereas the explicit system is slower, developed later and cognitively more demanding. For these reasons, referring to the cognitive capacity in question as 'social reasoning' might implicitly assume some explicit or active component. But given that the capacity often referred to as theory of mind consists of possibly distinct cognitive abilities, it is difficult to map a definite terminology onto it if the cognitive diversity is not well understood. None-the-less, there is no doubt that humans are able to understand and appreciate that others have minds of their own. Exactly how this process is constituted is still being investigated and it is the aim of the present thesis to help illuminate some of the intricate details of this ability which might also support development of a more rigid terminology.

1.2 Development and limitations of higher order social reasoning

The capacity for social reasoning develops through early childhood and is essential for recognizing that others have minds of their own with distinct desires, intentions and beliefs. As such, this capacity allows us to both interpret the behavior of others in terms of mental traits, but also to reason and predict how our behavior might affect their mental states. Mental states are broadly understood as intentions, desires, beliefs or emotions, even though the paradigmatic mental state most research focuses on is belief. Acknowledging that other organisms are sentient like ourselves is essential for distinguishing between them and physical constructs. So the cognitive capacity to represent the mental states of others have possible moral implications, which will be discussed later in section 2.3.3.

Much of the literature on social cognition have been on the developmental aspect and limitations in relation to having a higher-order theory of mind. Children are often thought to acquire a first-order theory of mind between ages 3 to 5 [90] and a capacity for second-order at ages 6 to 8 [69], some argue that this capacity is present even earlier [18, 58] and some argue that basic theory of mind aspects as goal perception are present already in 9 to 12-month-old infants [23]. As discussed earlier, it is important to notice that such studies mostly refer to implicit mental representation which happens automatically or with relative ease. In other words, what seems like higher-order social reasoning might not involve a deliberative reasoning process. This complicates the view of social cognition somewhat, since it makes it harder to determine whether the behavior in question is actually due to modeling the mental states of others or due to learned behavior which arguably does

not require the capacity to do so. It is also possible that such a demarcation is difficult to entertain at all.

Adults are usually limited at second-order social reasoning at most while there is evidence of third and fourth-order reasoning when playing games against computational theory of mind agents [82, 83]. There is evidence that social reasoning is not just limited in the sense of being restricted to second-order reasoning but also that any application of social reasoning is severely limited when it comes to deliberate and spontaneous use. See for example Lin et al. [59] for evidence that when interpreting the actions of others, the default is to rely on one's own mental state as representative of the mental states of others. This tendency have notable been reserved to children with autism disorder in a notable paper by Cohen, Leslie & Frith [3]. Frith & Happe [37] have subsequently argued that autism disorders cannot be characterized by theory of mind deficits and instead posit underlying cognitive impairments as explanatory forces. This shows both that deficits in theory of mind are not directly related to cognitive disorders, as well as how research on social cognition can affect our understanding of what normal cognitive development amounts to.

Lin et al. also argue that applying higher-order social reasoning requires effortful attention, meaning it is limited under cognitive load. Birch & Bloom [14] find evidence of what they call a *curse of knowledge* bias which refers to how adults' own knowledge about an event can make them worse at correctly attributing false beliefs about that event to others. Keysar et al. [56] show a disassociation between adult's ability to correctly attribute false beliefs to others and actually putting this ability to use when interpreting the instructions of others. Such studies indicate that there is a possible distinction between social reasoning applied in reflective settings and its use in guiding decisions in practice. As mentioned above, Heyes & Frith [50] argue it is possible that the fast, implicit social reasoning is a naturally developed cognitive ability, while the slower reflective use is culturally developed. The authors make the analogy between reading print and 'reading minds'. Following the analogy, it is possibly that profeciency in mind-reading, much like print-reading, is a lot more nuanced than just being literate or not. This leads us to the next section which argue that such diversity in proficiency might have been underappreciated in the literature.

1.3 Complexity higher-order social reasoning

Verbrugge [84] describes social cognition such that zero-order theory of mind concern 'world facts' while n-order theory of mind models $n+1$-order theory of mind of others. This is the commonly accepted view of higher-order theory of mind which I adhere to. But it involves a possible pitfall which Verbrugge calls a risky idealization, namely positing a fixed discontinuous bound on social cognition, or in other words, assuming each person can reason up to n-order of social cognition but not $n+1$. This assumes that sociocognitive diversity is exhaustively described by categories of n-order theory of mind for $n \in \mathbb{N}$. While possibly not intended, most studies seem to imply this. Such cases includes studies which show that subjects pass a first-order false belief task but not a second-order task. Such studies implicitly assumes that all those who are capable of n-order reasoning but not $n+1$ are cognitively on par at least in terms of social cognition. Much like capacity to read print comes in very nuanced degrees, I argue that capacities for higher-order social reasoning can be nuanced as well and in finer details than just n or $n+1$ orders of reasoning.

The traditional false belief task is intended to test for the *presence* of a belief and not its degree, but that does not mean that such a belief cannot be held in degrees. This is also an epistemological point (see for example Bayesian epistemology [93]). We can describe beliefs in terms of assigning

probabilities to something being true or false. Assigning probability p to some event being true is equivalent to assigning $1-p$ probability to the event being false. Now put this into perspective of a second-order false belief task like the one employed by Flobbe et al. [33]. It involves the following chocolate bar story: "John and Mary are in the living room when their mother returns home with a chocolate bar that she bought. Mother gives the chocolate to John, who puts it into the drawer. After John has left the room, Mary hides the chocolate in the toy chest. But John accidentally sees Mary putting the chocolate into the toy chest. Crucially, Mary does not see John. When John returns to the living room, he wants to get his chocolate." Participants are given the first-order false belief question "Does John know that Mary has hidden the chocolate in the toy chest?" and the second-order false belief question "Where does Mary think that John will look for the chocolate?" among other linguistic control questions. Those answering the first question correctly (John knows Mary placed the chocolate in the toy chest) but not the second one (that Mary thinks John will look for the chocolate in the drawer), will be said to have the capacity for first-order social reasoning but not second-order. Now assume that everyone in a group G passed the first test but answered incorrectly in the second. But suppose that some in G actually assigned a positive probability, for example 20%, to the correct answer being true. I argue that while this is not enough to warrant them answering it, it is indicative of more social reasoning than not considering the answer plausible at all.

A possible objection could be that you can either consider someone's mental state or not, because if even consider the correct answer as possibly being true, it must be due to attributing a mental state state to someone else. But this is inconsistent with saying that agents in G does not have a second-order theory of mind, which is what traditionally argued. I also accept that those failing the second-order test might not have a second-order theory of mind, but those assigning 20% to correct answer are closer to it than those assigning 0%. Humans might not explicitly view their beliefs as probabilities, but the current argument only requires that people can believe something to be true in non-binary degrees. It might also be argued that my argument is committed to the existence of 1.5-order social reasoning which is not well-defined. In response to this I argue that the terminology of theory of mind have already seen significant criticism, for example in terms of implicitly treating theory of mind as a monolithic capacity [76], so my point might just be another case of how theory of mind terminology is not up to date to the empirical reality.

To rephrase, describing orders of social reasoning only in integers such as first-order or second-order reasoning might be an approximation which does not provide the full picture of the socio-cognitive processes involved. Our findings in the canteen dilemma does suggest variation within those exhibiting n order theory of mind such that some are closer to $n+1$ than others. It indicates that (higher orders of) social cognition can be expressed in terms which are more continuous than simply n or $n+1$-orders of theory of mind. This leads to the canteen dilemma experiment below.

2 The canteen dilemma experiment

The canteen dilemma is a two-player coordination game with imperfect information. The game is structurally similar to the consecutive number example in Ditmarsch & Kooi [26]. It is framed in a thematic story as to make some of the zero-order logical reasoning easier [64, 87]. The story is the following. Each player is told that they and their colleague arrive for work every morning between 8:00 am and 9:10 am. They always arrive 10 minutes apart but only know their own arrival time. The payoff structure is ordered such that $(1 > 2 > 3)$ where (1) is going to the canteen together if

both arrive before 9:00 am, (2) is going to the office together at any time and (3) being all other configurations, that is, discoordination or either player going to the canteen at 9:00 or later. The game consists of a numbers of rounds where each player are given their own arrival time and have to decide between going to the canteen and the office. The game consisted of 10 rounds in the AMT trial and 30 rounds in the DTU trials. The game has the unintuitive structure such that there is no strategy which can guarantee canteen coordination. The Nash Equilibrium is therefore both players always playing office with maximal certainty.

The structure of the game involves reasoning about and predicting the actions of the other player and is as such facilitated by higher-order social reasoning. While canteen coordination has the highest payoff whenever both players arrive before 9:00, which is possibly true when arriving at 8:50, going to the canteen at 8:50 still involves the risk of the other player arriving at 9:00. I therefore stipulate that zero-order is sufficient for making an office choice at 8:50 since it requires players to reason about the possible arrival time of the other player which is a non-mental fact. I postulate the following: (1) participants choosing office 8:40 and earlier depend on first-order reasoning at 8:40, second-order at 8:30 and so on and (2) participants using only zero-order reasoning choose canteen at 8:40 and earlier, office at 9:00 and later while 8:50 is possibly chosen randomly. It is of course possible that a player who applies n-order social reasoning while not believing that the other is capable of $n-1$-order reasoning therefore chooses canteen regardless. That is, without common knowledge about rationality, even rational players would not be expected to play the rational strategy of only choosing office. So canteen choices at 8:40 and earlier do not necessarily imply a lack of social reasoning, even though a lack of social reasoning entails canteen choices at these times.

2.1 Method and design

2.1.1 Participants

Our experiment included 192 adults on Amazon's Mechanical Turk (AMT) platform. Another treatment performed on AMT is excluded because participants had an insufficient amount of time to read the instructions. Certain settings were applied in order to only include participants from Canada or the United States, participants with at least 500 approved Human Intelligence Tasks (HIT's) and a HIT approval rating of at least 98%. Participants were also given a unique ID such that they could only enter the experiment once and were awarded a $2 participation fee if they completed the HIT. We also conducted the experiment twice at the Technical University of Denmark (DTU) with 106 and 50 participants each during coursework. The courses in question were on artificial intelligence, multi-agent systems and logic, so the participants in the DTU experiments were likely primed for the experiment, especially the first of those trials. The second trial exceeded the duration of the allocated lecture, meaning some were eager to end the experiment.

2.1.2 Materials

The main experiment was conducted on Amazon Mechanical Turk (AMT) which is an online crowd-sourcing platform. The experimental setup was implemented in oTree 2.1.35 software [20]. AMT works as an online labor market where workers (also called turkers) can perform HIT's for monetary compensation. The platform has been used by social and economic researchers in lieu of typical lab experiments with local university students. Experiments on AMT have been shown to live up to the standards set by other data collection methods [13][16] and to provide reliable, replicable and more diverse data than legacy methods using university students [22][26][51][62][74].

After accepting our HIT and providing informed consent, participants were put in a 'waiting room' until they were paired up with another participant. After a group was formed, participants were directed to an initial introduction page which detailed the rules of the game with a time limit of 240 seconds (see Appendix A for screenshot). After reading the instructions, participants were directed to round 1 (of 10) where they were given their own arrival time and asked to make a decision between between going to the canteen or the office. Each round had a time limit of 61 seconds and also included the rules from the introduction. After making this decision they were prompted to estimate how certain they were that the other player made the same choice as them, ranging from very uncertain, slightly certain, somewhat certain, quite certain to very certain. After both players made their choices, they were prompted to a results page showing them the results of the previous rounds, including arrival times for both players, their choices, their own certainty estimate and resulting payoff.

Players payoff were implemented using logarithmic scoring as a proper scoring rule. Players were initially assigned a bonus of \$10 for the AMT trial, which was reduced by a penalty each round depending on well they did and their certainty estimate of success. AMT participants were paid their final bonus plus their participation fee while DTU students were told to maximize their payoff and awarded some playing cards for doing well. Palfrey & Wang [67] have shown that forecasts elicited from observers through proper scoring rules are significantly more accurate and calibrated then those elicited from observers using an improper scoring rule. Calibrated is defined as: "a set of probabilistic predictions are *calibrated* if p percent of all predictions reported at probability p are true" [77]. There is also evidence that forecasts elicited by the logarithmic scoring rule seem to have significantly less dispersion than quadratic scoring rules even though both are proper scoring rules [67], which among other reasons favored the logarithmic scoring rule over the quadratic. If both players went to the canteen, their bonus was reduced by $\ln(certainty)$. If they both go to the office the penalty is doubled: $\ln(certainty) \cdot 2$. If they go to different places, the penalty is $\ln(1 - certainty) \cdot 2$.

The experiments included four post-game questions and the DTU trials included 3 further questions (See appendix B). Results are shown below.

2.2 Results

All Python code for visualization of the canteen dilemma can be found at https://github.com/thomasnicolet/canteen-dilemma-thesis.

2.2.1 Canteen/office choices and certainty estimates

The first chart below shows the percentage of participants who chose canteen at specific arrival times (blue line) and how many percent chose canteen and giving a 'very certain' estimate of how certain they were that it was the right choice (orange line). Shaded areas represent 0.95 confidence intervals. Precise values are annotated.

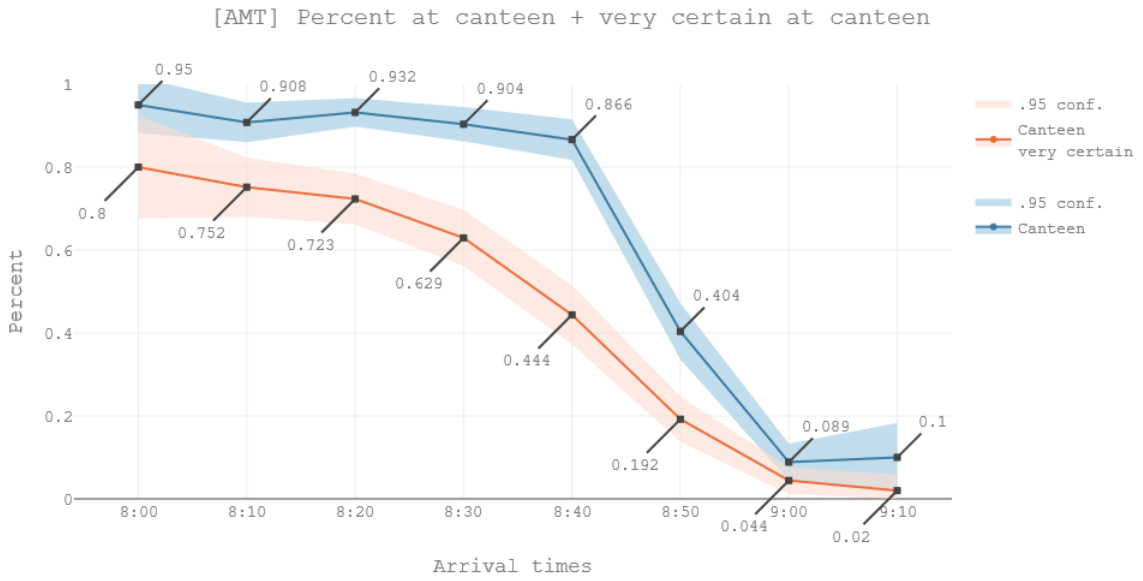


Figure 1. Choices and certainty estimates.

Figure 1 shows that participants generally went to the canteen from 8:00 to 8:40 and to the office at 9:00 at 9:10, while office was favored at 8:50. The orange line shows that even if participants went more or less to the canteen with the same propensity from 8:00 to 8:40, the percent who are very certain about it seems to start dropping already from 8:10 or 8:20. It was postulated above that going to the office does not necessarily rely on higher-order social reasoning, while office at 8:40 requires first-order social reasoning, 8:30 second-order and so on. While the general canteen choices at 8:00 to 8:40 do not necessitate a theory of mind, the different certainty estimates do. Ascribing p probability to canteen being the right choice is equivalent to ascribing $1-p$ probability to the office. So those who are less than maximally certain about canteen seem to consider office as a possibility, and I argue that some appropriate order of social reasoning is necessary to have this belief. The orange line shows that from 8:20 to 8:40, the percent being very certain about canteen choice drops from 72% to 44%, while the percent choosing canteen only drops from 93% to 87%. This result indicates that the certainty estimate given by participants indicated some theory of mind which was not shown as robustly in the binary canteen/office choices. It more specifically show that there is significant uncertainty about canteen choices at 8:30 and 8:40.

Figure 1 also implicitly show at which arrival times discoordination happened, which are the arrival-times around 8:50, that is, (8:40, 8:50) and (8:50, 9:00). See Appendix F for a plot of the percent of canteen and office coordination as well as discoordination for each arrival time pair. This

2.2.2 Question regarding cutoff

After the game had ended, participants were asked: “Imagine you could have agreed beforehand with your colleague about a point in time where it is safe to go to the canteen. What time would that be?". Error bars indicate 0.95 confidence intervals.

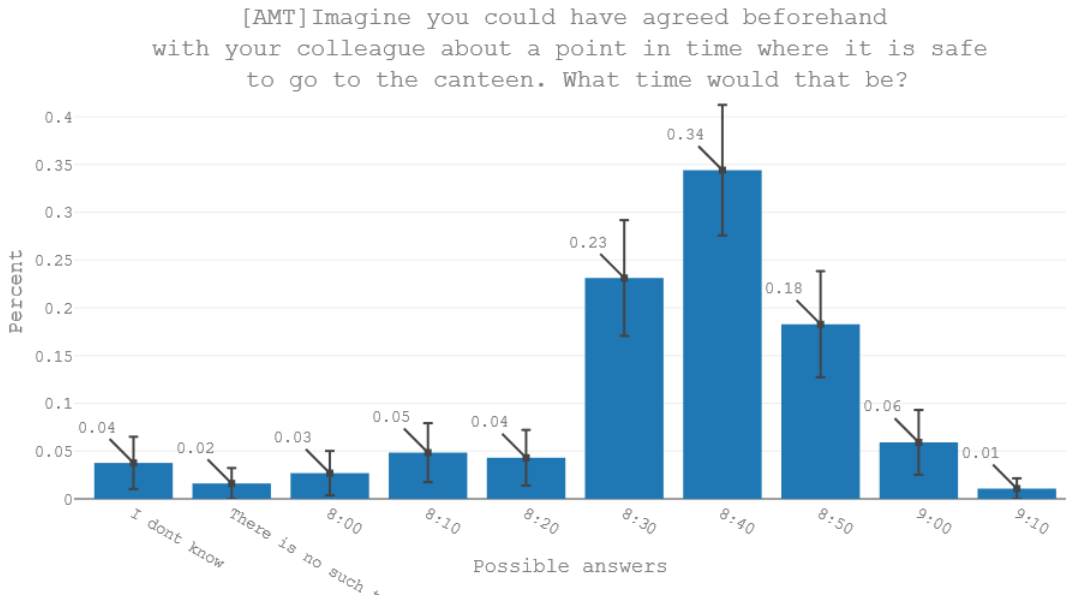


Figure 2. Bar-chart for question about safe canteen choice.

Figure 2 shows that a majority of participants (75%) chose 8:30 to 8:50 as the point in time which would be safe if they could have agreed upon it with their colleague before the game. The rest are somewhat evenly scattered among other answer possibilities, possibly due to random choices. The pragmatic nature of the question implies that answering for example 8:30 implies that all arrival times before that would be safe while those after would not be. If participants simply agree to go to the canteen at 8:50, then it would be safe to go to the canteen at 8:40. The pragmatic aspect of the question arguably imply a specific answer implies that they would not go to the canteen at later times. Notice that the answer "there is no such time" is only answered by 2% of participants possibly influenced by random decisions. This possibly means that participants believed they could use a strategy including canteen choices without risk of discoordination.

2.2.3 Whose fault was it the game was lost?

The game ended if either participant lost their entire bonus before the maximum number of rounds had been reached. In case of this happening, participants were asked the question "whose fault was it that the game was lost?".

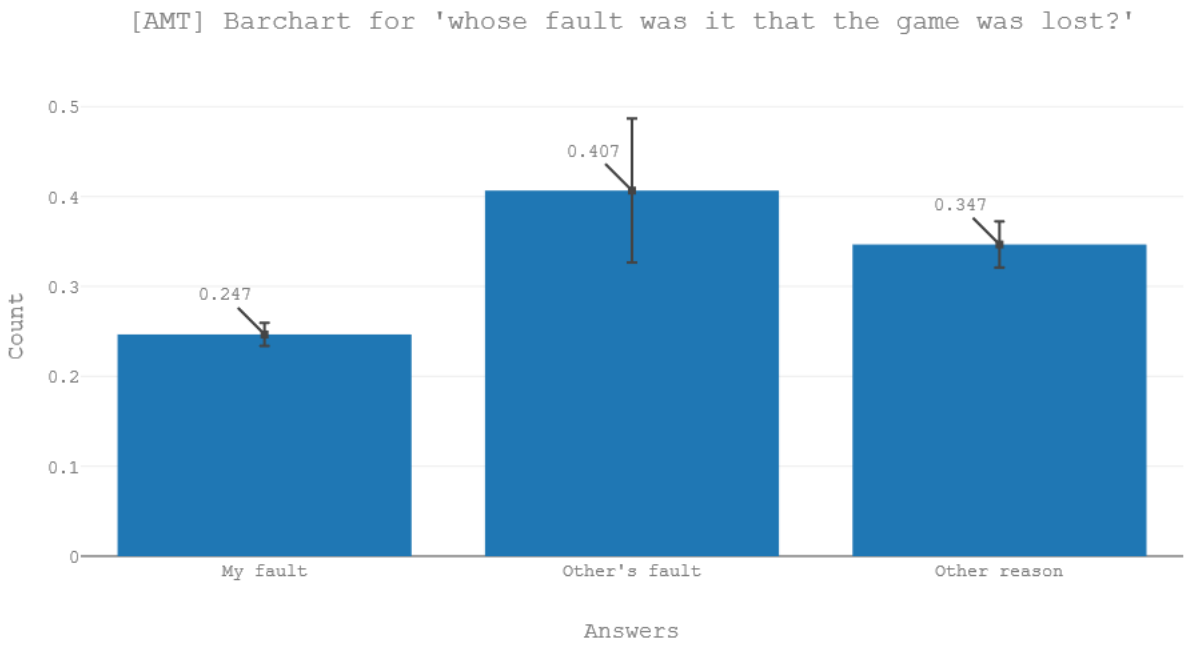


Figure 3. Bar-chart showing percent of different answers to post-game question.

Figure 3 shows that participants were least likely to answer that it was their own fault that the game had ended (25%) and most likely to answer that it was the other's fault (41%). This is a complicated question of course, since the nature of the game makes it difficult to say in most cases that either player made a wrong decision. The third category (other reason) is supposed to be a catch-all for other explanations. There are a few explanations for why participants might have blamed the other player over themselves. The first possible explanation is that even those aware of their mistakes might not want to admit it (even to themselves). A second and more interesting explanation is that those who engage in for example first-order social reasoning only attributes zero-order reasoning to others. When discoordination happens, this naturally leads to blaming the other player, since participants did not predict and nor can they understand the order of theory of mind used by the other player.

2.2.4 Question pertaining common knowledge

After the game had ended, for whatever reason, participants were asked about their beliefs concerning common knowledge.

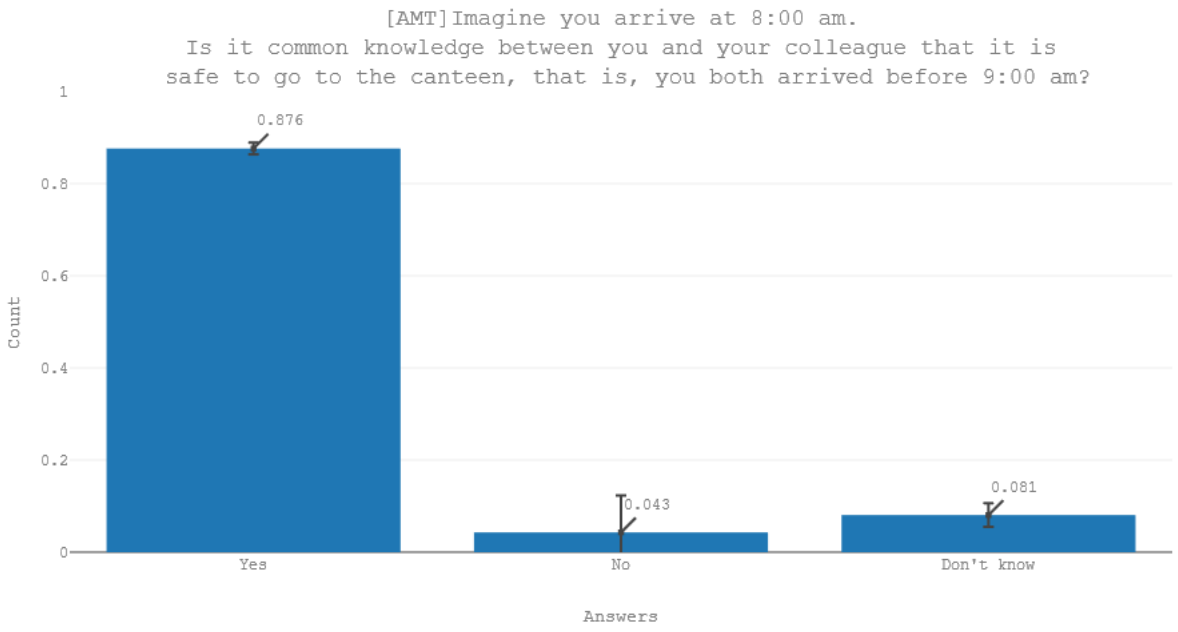


Figure 4. Bar-chart showing percent of different answers to question pertaining to common knowledge.

Figure 4 shows that 88% of participants responded that it was common knowledge that both players arrived before 9:00, when they themselves had arrived at 8:00. Due to the noise in the data and confidence intervals, the 4% answering negatively is possibly accounted for by random choices. The answers likely pertain to the everyday linguistic usage of the term 'common knowledge', which refer to something that either everyone knows (mutual knowledge, which is distinct from proper common knowledge) or even just something that most people know.

2.2.5 Question concerning curse of knowledge

The supplementary experiments at DTU included the question "Imagine you arrived at x and you have been secretly informed that your colleague's arrival time is 8:50. Where do you think your colleague will go?" Half of the participants were given 8:40 as their own arrival time while the other half were given 9:00. The question concerns whether player's own knowledge of the other's arrival time affect their prediction of the other player's decision. It relates to the curse of knowledge from Birch & Bloom [14] since participants might attribute their own belief (that it is early enough or too late to go to the canteen) to the other player. This is indeed the result we got in the first DTU trial but not in the second. See the plot for the first trial below and second trial in Appendix C.

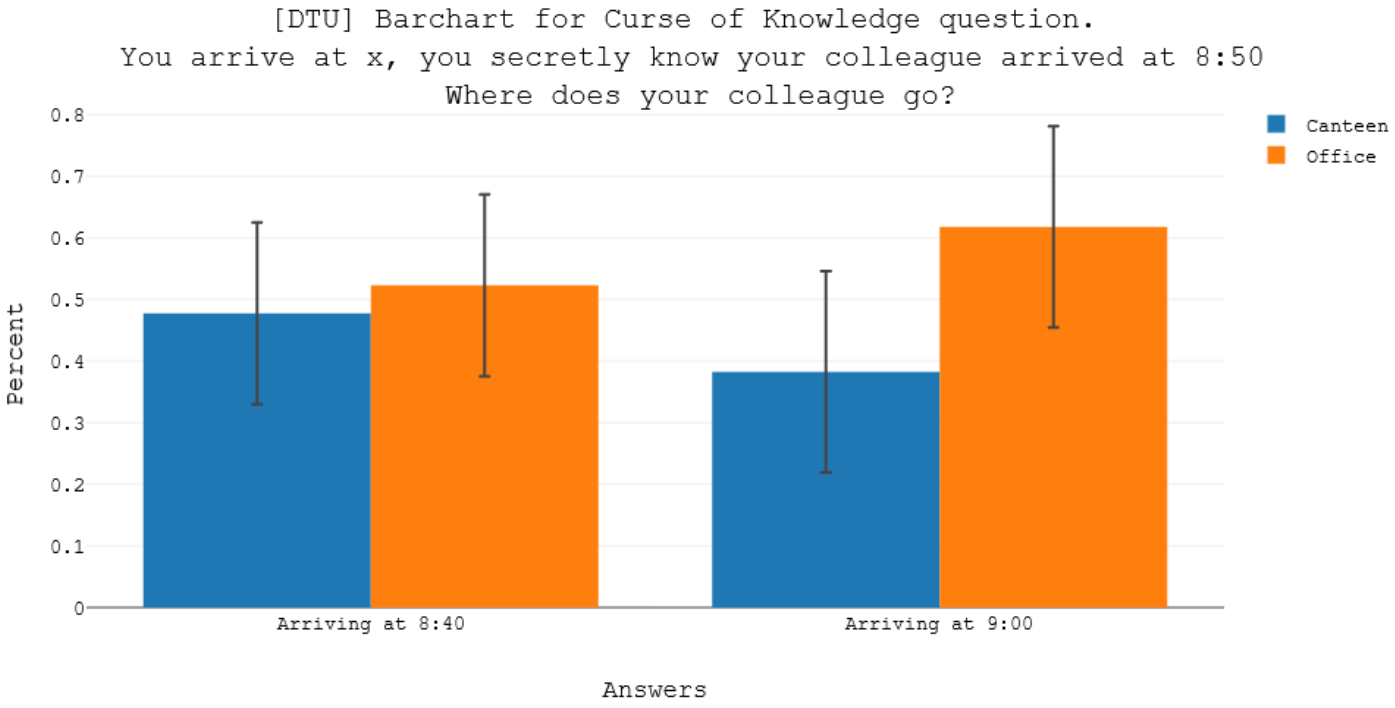


Figure 5. Grouped bar-chart for curse of knowledge question with 0.95 confidence intervals.

As figure 5 shows, participants who arrive at 9:00 are significantly more inclined to answer that the other player will go to the office than those arriving at 8:40. The confidence intervals are overlapping however and later results at at second DTU trial had differing results (see Appendix C). However the data from the first DTU trial overall indicated the participants utilized higher-order social reasoning more than both of the other trials. That is, the first DTU trial had results which implied that participants had both understood the rules of the games better and had higher-order theory of mind considerations, but none the less, the curse of knowledge was most prevalent in that case. The first DTU trial also included slightly more than twice the participants than in the second.

2.2.6 Categorizing free text answers

Participants were asked about what strategy they had used and there seemed to be some patterns in these answers. Some answer 'common sense' or 'intuition', while some simply stated that the earlier the arrival time, the more they went to the canteen, without further specification. A few answers were explicitly referring to first or second-order theory of mind considerations. Some answers were also indicative of forward inductive reasoning or reactive strategies, that is, decisions depending on the previous actions of the other player. Many of these answers indicate a belief that if some strategy, like the reactive strategy, was just deployed correctly by both players, it would lead to coordination with certainty (implying that once you know what the other player does at specific times, you can coordinate) However, regardless of knowing the other players strategy, if that strategy contains a

canteen choice, coordination is not certain across the whole game. This also includes players who employed first or second-order social reasoning, who were sure that if only the other player had done the same, their earlier canteen choices had been safe. See for example the following strategy answer from the first DTU trial:

> "It was easy for 8:00, 8:10 and 8:20. There I would go to the canteen based on the fact, that my friend would arrive at latest at 8:30, and then they would in worst case scenario think that I arrived at 8:40. Thus we would both go for coffee. In the case of 8:30, I would also go to the canteen, but I would not be very certain cause my friend might be there at 8:40 and think that I would be there at 8:50. They might think that IF I am there at 8:50, I would think that my friend is there at 9:00, and thus I might go to the office. Thus for 8:30 and above it is not certain. It would depend on the history of our decisions... "

While it explicitly refers to the participants using their second-order social reasoning, it indicates a lack of third-order social reasoning in relation to times 8:20 and earlier. There is no uncertainty about these arrival times, implying an awareness of the possible importance of third-order reasoning. This is important because it does not just show that people have different cognitive capacities which can lead to mistakes in games. It also indicates that people do not have sufficient introspection to predict the possibility of such mistakes.

2.3 Discussion

There are a few problematic aspects of Figure 1. Notice that there are 10% going to the canteen at 9:00 and 9:10. There are also 40% choosing canteen at 8:50, even though the rules state that if one or the other player arrives at 9:00, they should go straight to the office. There are also around 10% going to the office at times 8:00 and 8:10 and we do not expect participants to have a sufficiently higher-order theory of mind to make such choices informed. There are bound to be some random answers in such tests which accounts for at least some of the answers given. We might consider if all the canteen choices at 9:00 were due to random choices or if the rules were ambiguously stated such that these decisions were made deliberately.

Given a few of the free text strategy answers, it is possible that some participants were biased towards the canteens for reasons not stated in the rules. In other words, their real world preference for coffee in the canteen over work in the office might have persuaded them to choose the canteen as much as possible. This means that the positive facilitative effect of providing a thematic narrative to a logical reasoning test [64, 87] might not have been effective. Everyday narratives can help at explaining abstract structures, but it might do so by prompting heuristics which might sometimes be helpful and sometimes detrimental to logical reasoning. However if the rules were too ambiguous, we would expect the same results in the DTU trials. See the plot below for the first DTU trial which depicts the same as figure 1, percent choosing canteen and percent being very certain about their canteen choice.

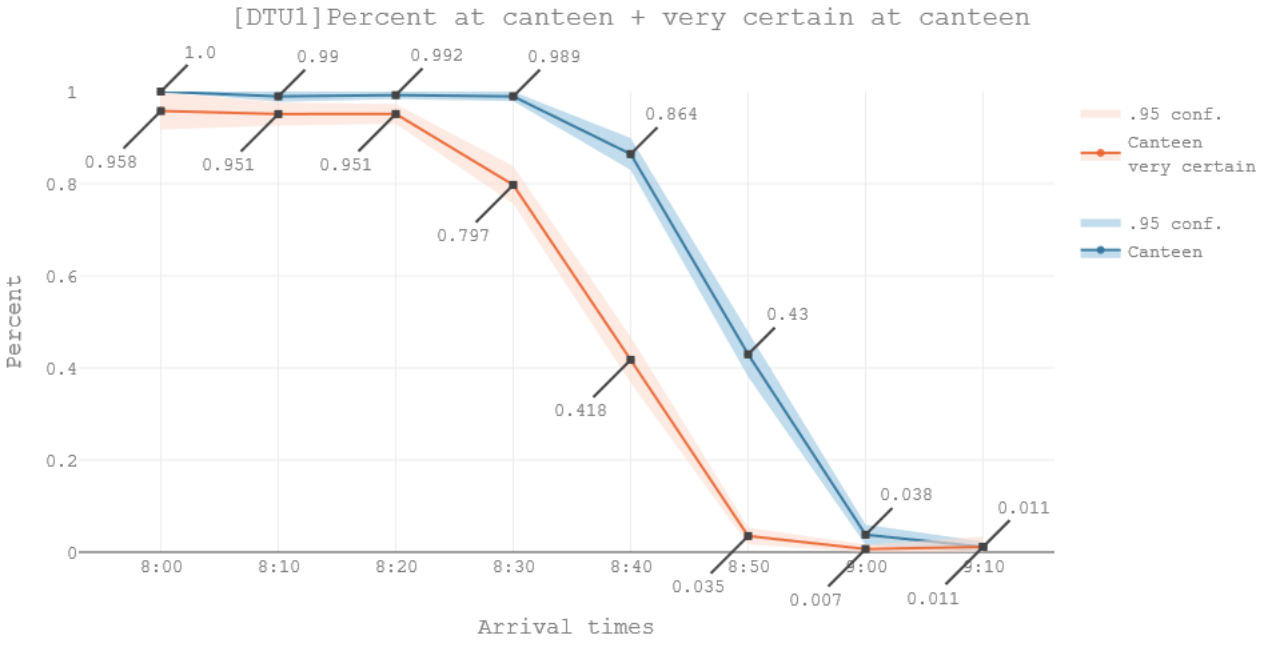


Figure 6. First DTU experiment.

Students in the first DTU trial made nearly exclusively canteen choces from 8:00 to 8:30 and nearly exclusively office choices at 9:00 and 9:10. A plausible explanation for the difference in results is that the DTU students were more motivated to participate and understood the rules better and as such made less random choices. Therefore, a significant amount of the noise in the AMT results are likely due to random decisions. In the DTU trial depicted in Figure 6, we might still consider 43% a high amount of participants choosing the canteen at 8:50. This was possibly due to the wording of the instructions. It was changed from "if you arrive before 9:00 am, you have time to go to the canteen" to "if you *both* arrive before 9:00 am ...", emphasizing that both players have to arrive before 9:00 am to warrant a canteen choice. See Appendix D for a plot similar to Figure 4 for the second DTU trial. Contrary to expectations, this lead to a significant increase of canteen choices at 8:50, as 63% chose canteen at 8:50 in the second DTU trial. This implies that it was at least not just the wording which lead to the behavior in question. Notice that the orange line shows the amount of people being very certain about their canteen choice. This means that every other choice was either office or less than maximally certain canteen choice. I have argued in section 1.3 that such answers are evidence of higher-order social reasoning, so the plot above shows 97% applying zero-order reasoning at 8:50, 58% applying some first-order reasoning at 8:40 and 20% applying some second-order reasoning at 8:30.

2.3.1 Continuity in orders of social reasoning

I have proposed that office choices at 8:40 and 8:30 are indicative of first and second-order social reasoning. Figure 1 and 4 shows that while participantss might have chosen canteen at most arrival

times leading up to 8:50, they are less certain at later arrival times. The reason for not being certain about your canteen choice is considering that the other might have chosen office and the reason for believing that at 8:40 and earlier is a result of higher-order social cognition. See Figure 7 below for all certainty estimates.

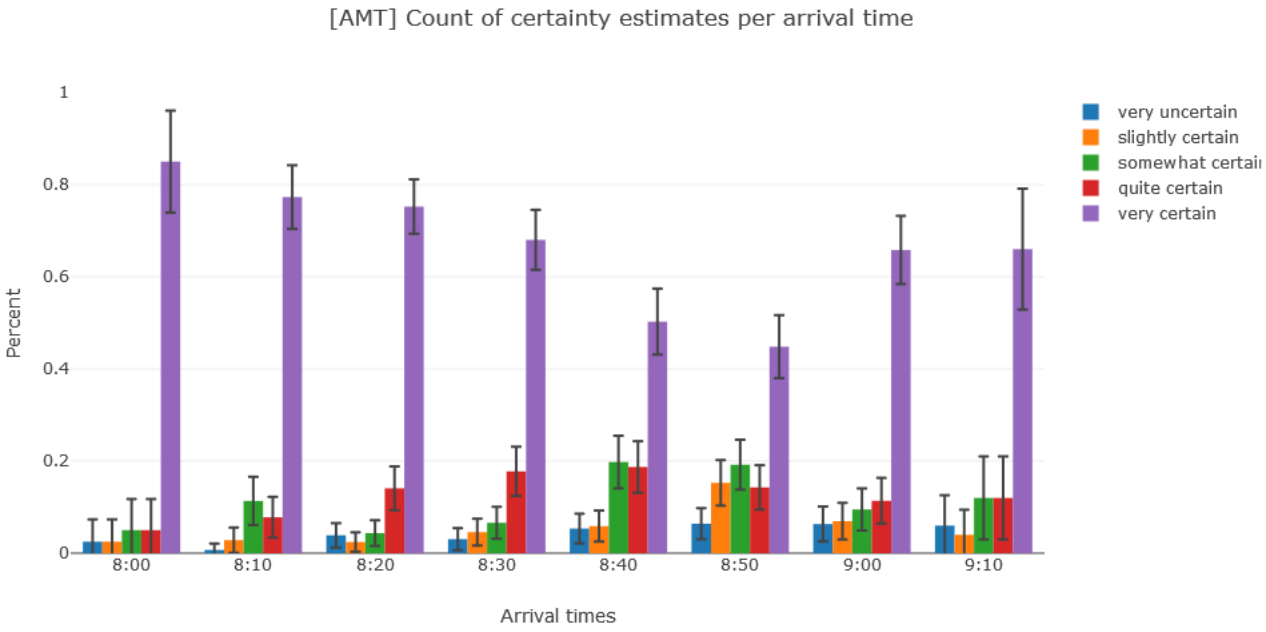


Figure 7. Certainty estimates

The certainty estimates of success given by the participants are important. The certainty estimates allowed for non-binary answers meaning participants could qualify their belief. We can also note that going to the canteen and ascribing p probability to the other player doing the same is equivalent to ascribing $1 - p$ probability to the other player choosing the office. Therefore, if this certainty estimate can be indicative of a higher order social reasoning, it is possible that higher order social cognition in general can be understood in degrees in between orders of social reasoning measured in integers. This deflects what Verbrugge calls the simplistic danger of positing fixed bounds on social cognition, such as 'everyone can reason up to n order of theory of mind, but not $n + 1$' [84]. It is possible that few actually believe that higher-order social cognition only exists in discrete degrees such that there is a sudden jump from first to second-order social reasoning and no middle ground. None the less, this is the picture often given in research on social cognition. This result arguably calls for a nuanced picture of the mental capacity related to representing the mental states of others which does not presume that it is an all or nothing capacity.

2.3.2 Lack of introspection

The certainty estimates in Figure 1 and 4 indicate that participants were very certain about going to the canteen at sufficiently early times. Due to the lack of uncertainty involved, these answers also seemed easier for participants than later arrival times as indicated in free text answers. See for example the following answer: "If it was close to 9:00AM, within 20 minutes or so, I would go to the office. Otherwise I would just chose to go to the canteen. Unfortunately it almost worked out for everyone except one, and I have no idea why they would have chosen to go to the office on said round because it was really early and we would have had plenty of time to go to the canteen." This answer implies that the participant chose canteen at 8:30 and earlier and office at 8:40 and later. Notice that not only does the reasoning stop at 8:30, the reasoning behind going to the office at 8:50 and 8:40 is not continued for earlier arrival times, even though both 8:40 and 8:50 include the possibility of canteen-coordination as well. The participant's co-player answered "I HAVE THINK OF MIND" and while grammatical unconventional, it likely refer to the mind of the other player, indicating that the importance of inferring the mental states of others had become apparent for the participant.

This means that because participants were unable to infer that they might have to choose office, they do not consider it a viable strategy. If simply arriving early enough, players seemed apt at believing that "there is plenty of time to go to the canteen". The statement is true in a zero-order sense, since both players could go to the canteen whenever one arrives at 8:40 or earlier. But the task of the canteen dilemma is not just to discern whether or not this is the case, it is also to think about if the other player knows this, if the other player knows you know it and so on. The participant choosing to go to the office at 8:40 means that even though the participant knew it was "early enough", the participant also knew that the other player possibly did not know this.

The importance of this is not just that higher orders of social reasoning were lacking, but also that participants did not have introspection in regards to this. The lack of introspection shown in the canteen dilemma entails that when participants arrive sufficiently early and are certain that the other player will go to the canteen, they often do not understand the other player and think they made a mistake.

This indicates that higher-orders of social cognition are not just epistemically unavailable to participants, they are implicitly regarded as irrelevant. We can infer this from the situations where such reasoning is required but also considered easy or obvious. This makes sense for two reasons. First, recognizing the relevance of higher orders of social reasoning requires consideration of these exact orders of social reasoning. Secondly, the seeming irrelevance of higher orders of social reasoning is reinforced by behavioral consequences of limited reasoning. Participants went to the office at 8:50 roughly half the times, while some went to the office at 8:40, but at 8:30 participants generally chose the canteen (although some were not 'very certain', likely due to second-order social reasoning). Of course participants could not have explicitly predicted this, but the general pattern of reasoning did indeed make 8:20 out to be a generally 'safe' canteen choice. This is sometimes referred to as a 'meta' within a game with multiple nash-equilibria, where one of them is adhered to through convention.

This is an interesting aspect of the problem of coordination without common knowledge. There is a risk of discoordination involved in any strategy involving canteen, no matter where the cut-off between canteen and office is put. However, if players uniformly choose to go to the canteen at 8:30 and office at 8:40 and later, this will only lead to discoordination at 8:30/8:40, while all other arrival times will be safe choices. It also explains that if both players accept they have to have a cut-off

point between the canteen and the office, the optimal strategy would be setting the cut-off point as late as possible (that is, canteen choices at all arrival times before 9:00), since this maximizes canteen coordination relative to office coordination. But this strategy requires accepting the risk of possible 8:50/9:00 discoordination, which can be inferred from zero-order reasoning. So, if one does not accept that risk, that is, one chooses the office at 8:50, the only explanation for starting to choose canteen at earlier arrival times is due to either (1) lack of sufficient order of social reasoning or (2) disbelief in whether the ability to perform sufficient order of social reasoning is common knowledge.

To summarize, the canteen dilemma indicated both that higher-order social reasoning can be continuous and that the participants lack introspection about the limitations of their social reasoning. The next section is an attempt to put the results and discussion into a practical context. I do this by going through a case structured similarly to the canteen dilemma but where we intuitively do not accept a risk of discoordination.

2.3.3 Think again

Higher-order social reasoning seem to present a sometimes insurmountable cognitive challenge. It is plausible that the complexity of this reasoning also bars off introspection into one's own limitation in regards to this cognitive capacity which makes the practical importance of such reasoning easy to ignore. This section aims at presenting a case where higher-order reasoning makes a practical and moral difference, showing that people cannot rely on their intuitions about higher-order social cognition. It is in other words intended to show that higher-order social reasoning is not simply intellectual gymnastics but makes a tangible moral difference. Numerous research papers on social cognition preface their work with homage to the quintessential role it plays in social interaction, [14, 21, 33, 49, 50, 82, 84]. Due to the centrality of the capacity in human life, the unreliable and limited use of it can have adverse consequences, but due to lack of introspection, we might not be able to properly predict or understand these consequences when they occur.

I will show this by going through an example with a structure similar to the canteen dilemma. The canteen dilemma can be abstracted as a situation where two people simultaneously have to coordinate on one of two decisions where they both prefer one coordinated decision in some situations (like canteen coordination) while such a decision is ruled out in others (like arriving at 9:00). The example will consists of a situation of two people with the possibility of establishing a sexual relation, depending on their mental states (either interested or not) and simultaneous choices (taking some action or defecting from doing so). Mental states and choices are mapped analogously to the canteen dilemma. The people in question have to decide whether they want to pursue a relation with the other (analogous to canteen) or defect, showing no interest (analogous to the office). Much like the canteen dilemma, there are some situations where both people in the example prefer pursuing a relation with the other, which is when both have the right mental state, which is some type of interest in each other. However, analogous to the canteen dilemma, there are situations where this is not possible, which is when at least one person is not interested in the other. So, whenever both persons are interested in each other, they both prefer pursuing some relation, but if either at least one is not interested, or if the other defects, none of them are interested in pursuing a relation to each other. I propose a situation where the act of pursuing a sexual relation with a person who is not interested is a violation of that person's boundaries. Due to deontological constraints, I assume that such a violation is impermissible. This is different form the canteen dilemma, where the risk of discoordination involved in choosing canteen at 8:50 can be outweighed by possible coordination. In the present example, the risk of discoordination is impermissible and as such cannot be outweighed

by other possible consequences.

Not only this, I assume much like the canteen dilemma that if one person defects (no matter their mental state), it is in the interest of both to defect. This can be argued for as an plausible assumption in different ways, but let us just assume that it is highly unpreferable even if both are interested due to social embarrassment involved. In other words, canteen/office choice is here made into taking some action $!P$ or refraining from it. Now for the case.

Imagine that a (Anne) and b (Bill) are in a social situation structured like above. Both Anne and Bill can either be interested in pursuing this relation or be disinterested (an internal mental state, call it m_i for i being interested, which is either true or false relative to a person). Both have to simultaneously choose to perform some action (colloquially 'making a move') in an attempt to pursue this relation or defect and turn their back to it (call this $!P_i$ indicating that i does $!P$, whereas $\neg!P_i$ refers to not doing $!P$). $!P$ simply refers to an action which in non-reciprocated would be morally problematic, which include speech acts. Assume that if a person is disinterested in this relation, they will always defect, that is $\neg m_i \rightarrow \neg!P_i$. As mentioned, if a person defects for whatever reason, it is in the interest of both that the other also defects. This depends on the assumption that the people in question have good intentions. If an action might violate the rights of another, it is in the interest of the person to not undertake that action. It is also easier to abstract a rights violation as a discoordination if it only involves good intentions. The following case therefore shows that well-intentioned people can partake in actions which has a chance of being morally impermissible, even if they want to avoid this, due to limitations of higher-order social cognition. Let us not start updating the situation.

Assume the default situation is s_0 where there is common knowledge that both Anne and Bill are not interested in pursuing any relation to each other and will therefore defect when having to make a simultaneous choice about pursuing a sexual relation with each other. Note that anyone choosing to pursue a relation generates common knowledge that they are interested, which is the converse of the implication stated above, that is $!P_i \rightarrow m_i$. We know that in s_0 both Anne and Bill will defect and we can state the counterfactual that if one of them did not defect, they would be morally blameworthy since they both know the other is not interested. The focus of the next few paragraphs is to keep track of what Bill will do and to consider the moral implications involved.

We can now start updating the situation, imagining s_0 changing over time. Let us update the default situation to s_1, where Bill tells Anne that he is interested in her (m_b).[2] It then becomes common knowledge that Bill is interested in Anne, while it remains common knowledge that Anne is not interested in him ($\neg m_a$). When having to make a choice, Anne will defect (she is not interested) and so will Bill because he assumes Anne is not interested per default. Bill knows that he should not pursue a sexual relation with Anne if she is not interested ($K_b \neg m_a$), since this would be a violation of her rights and he would be rightfully ostracized if he did so in this situation. This is common knowledge as well, so Bill knows Anne expects him to defect as well.

Suppose the situation is updated to s_2 such that Anne becomes interested in Bill which she secretly writes down on a note. So Anne is now interested in Bill (m_a), Anne knows Bill is interested in her ($K_a m_b$), but Bill still does not know Anne is interested in him and Anne knows this, since she knows that as far as Bill is concerned, he still believes he is in s_0. So Bill assumes Anne is still disinterested in him and he will defect like in s_1 and he would be morally blameworthy if he did not.

[2]This is strictly speaking a public announcement in the sense that is public to Anne and Bill, but it could still consist of Bill just writing it down in front of Anne with no one else around.

Anne knows Bill will defect and she chooses to do the same, as it is in both of their interest to not pursue a sexual relation without someone who does not reciprocate. Even though Anne is interested in Bill in s_2, if Bill where to pursue a relation to her, Anne would be right to berate Bill, since if Bill behaves consistently, this means would have done it in s_1 as well which we have established as wrong. So far so good, we accept that both Anne and Bill defects. We have also established that it would be wrong for Bill to do $!P$, because since he cannot tell s_1 and s_2 apart, it means that he would have done it in s_1 as well, which was clearly impermissible.

Now assume a further updated situation s_3 where Bill secretly reads Anne's note, learning that she is interested in him. While Bill has learned that Anne is interested in him, Anne does not know this and Bill is aware of Anne's ignorance. So the move from s_2 to s_3 is unnoticeable for Anne and Bill knows this. We have already established that Anne would not like Bill to pursue a sexual relation with her in s_2, so the same holds for Anne in s_3. Bill knows that nothing changed for Anne, so he chooses to defect. If he had done $!P$ in s_3, even after reading her secret diary that she is interested in him, it might seem like he did so in s_2 from Anne's point of view, at which point it is indistinguishable for Anne whether Bill knows he is in s_1 or s_2. Notice that for Anne to recognize this, she has to perform first-order reasoning, as she has to consider Bill's beliefs which from her point of view includes $\neg m_a$, enough to put blame on Bill if he chooses $!P_b$. For Bill to know this about Anne, he has to consider her first-order reasoning, which implies second-order reasoning from Bill. Both of these cases are empirically plausible in terms of cognitive capacities.

Suppose s_3 is updated to s_4 where Anne secretly notices that Bill reads her diary. So Bill is interested in Anne (established in s_1). Anne is interested in Bill (established ins_2). Anne knows Bill is interested in her (it is common knowledge in s_1). Bill knows Anne is interested in him (he read her secret note in s_3). Anne now knows that Bill knows that Anne is interested in Bill ($K_aK_bm_a$, as she secretly saw him read her note). But Bill does not know this epistemic fact! So alas, when Bill has to choose whether to make a move or not, he is still in the same position as s_3 and will choose to defect. Same as before, since Anne knows that Bill is in a situation similar to s_3 where she would be right to berate him had he pursued a sexual relation to her, she knows the same still holds and chooses to defect as well. As Anne has gained one more iteration of epistemic knowledge in s_4, Anne now also depend on second-order social reasoning.

Now introduce the last situation s_5 where Bill secretly realizes that Anne saw him reading her note. So we add that Bill knows that Anne knows that Bill knows that Anne is interested in Bill ($K_bK_aK_bm_a$). Like earlier, nothing changed for Anne. So due to second-order reasoning Anne will still expect Bill to not pursue a relation with her, and find him morally blameworthy if he does so anyway, and since Bill knows this, he ideally will defect. To summarize, Anne and Bill are interested in each other. It is common knowledge that Bill is interested in Anne. Bill knows Anne is interested in him so it is mutual knowledge that they are interested in each other. Anne knows that Bill knows this, and Bill knows that Anne knows this, so it is even mutual knowledge that it is mutual knowledge that they are interested in each other, $E_{\{a,b\}}E_{\{a,b\}}m_{\{a,b\}}$. But neither know this epistemic fact! Now for the crux of the argument. While Anne still relies on second-order social reasoning, Bill needs to apply third-order reasoning to reason sufficiently about her iterated beliefs in order to know she will berate him for pursuing a sexual relation with her. So since Bill is interested in Anne, knows she is interested in him and knows that she knows that he knows this, he might pursue a relation with Anne because the third-order reasoning required for Bill to correctly attribute second-order beliefs to Annie are cognitively out of reach. But not just this, the findings from the canteen dilemma concerning lack of introspection indicate that Bill is not aware of this

limitation. So Bill cannot predict nor understand why Anne defects or why his action was morally blameworthy. The possible reasoning for Bill to stop defecting in s_5 could be some default reasoning kicking in, which assumes common knowledge to exist whenever two or more iterations of mutual knowledge is established.

What does this mean? When Bill evaluates whether or not to do some action in situation above, he has to evaluate how it will be viewed by Anne. We assume that Bill is consistent in his actions. If Bill thinks that Anne will find $!P_b$ morally acceptable in s_5, where Anne is in an equivalent state as s_4, he has to believe it is acceptable for her in s_4 as well. Since s_3 and s_4 are indistinguishable for Bill, he therefore must think it is acceptable in s_3 as well (where he just read her note). But remember in s_3, Anne might think she is in s_2. So from Anne's perspective, Bill doing $!P$ in s_3 is judged as if Anne was in s_2. And in s_2, Anne does not know if Bill makes a does $!P$ from the perspective of s_2 or s_1(indistinguishable for Bill), so Bill must think it acceptable to make a move in s_1 as well, where Bill believes Anne is not interested in him, which has been established as morally impermissible.

This is just reverse route of inferring why making a move would be wrong for Bill in s_5, but notice that it shows that if Bill would do $!P$ in some s_n he has to consider how it is evaluated by Anne in the situation before it, which also depends on how it would be perceived in the situation before that. So without common knowledge of $m_{\{a,b\}}$, there is no strategy for Anne and Bill that involves $!P$ which does not involve the possibility of discoordination. One might argue that it is simply not morally blameworthy for Bill to choose $!P$ in s_5. But remember that whenever we say $!P$ is deemed morally permissible by some agent in a situation, we must also say that it is morally permissible in situations which are indistinguishable for that agent. It is also possible to hold a moral theory such that the permissibility of an agent doing an action depends only on the mental states of others and not on the agents knowledge of such mental states. This would amount to an error-theory, such that even in s_2, where Anne secretly writes down that she is interested in Bill, it would be permissible for Bill to pursue a relation with Anne and our intuitions about the situation are simply wrong. Arguing against this is outside the scope of this thesis, but it might suffice to say that it is not a practically viable moral theory. One might argue that while Bill should not do $!P$ after reading Anne's note, because from Anne's perspective it is extensively the same as when Bill believes Anne to be disinterested, but still argue that Bill should do $!P$ after Anne realizes Bill read her note. But this implies Bill will somehow behave differently given the same information (since the situations are indistinguishable for Bill).

Since humans are limited in terms of higher-order social reasoning, it is difficult for us to realize that there is no strategy which includes doing $!P$ in a situation s_0 to s_n where $n \in \mathbb{N}$ which does not involve the risk of discoordination. Agents might realize the implications in one situation, and therefore defect and but then do $!P$ in different situation where they have more information but still no common knowledge. The possible lack of introspection entails that they cannot foresee the negative consequences of social discoordination and they possibly cannot understand how it might have happened, meaning that neither the moral implications can be dealt with nor can future behavior be adjusted. This whole section can be stated explicitly in the terms of certain social interactions requiring consent. Even under the assumption of morally well-intentioned people, certain interactions requires common knowledge about consent, not just private or mutual knowledge of consent. Any finite iteration of mutual knowledge of consent is not sufficient if we want to avoid the risk of discoordination entirely. This section is intended as describing the practical consequences of being limited both in terms of higher-order social reasoning and in terms of proper introspection into this

very limitation. While the normative implications of this are not discussed further, this section is not meant to be defeatist, since the problem of lacking common knowledge can be solved by clear communication. This section is rather intended as describing how cognitive limits can have consequences which we want to avoid, and which we can only avoid if we become aware of the structures that lead to them. This implies reconsidering if the necessary common knowledge is established when dealing with complicated situations where successful social interaction is paramount.

3 Conclusion

Our findings indicate that the capacity for higher order social reasoning might exist in various degrees which are not captured by the approximation of assuming a capacity for n-order of social reasoning but not $n+1$. Due to the complex and still somewhat enigmatic nature of this otherwise well-researched cognitive ability, it is also important to develop and properly define terminology of the broad cognitive capacity originally referred to as 'theory of mind' by Premack and Woodruff [70]. This is difficult however before the diversity and complexity involved is better understood. Our results indicates that there is more cognitive diversity in terms of higher order social reasoning than what a traditional categorization of n and $n+1$ order reasoning implies. Like other studies have suggested, this calls for research investigating the possible diversity of processes involved in reasoning about and understanding the minds of others. Our own research aids in this, as accepting the possibility of continuous degrees of higher order social reasoning possibly indicates that some might be in different stages of gaining appreciation for higher orders of social reasoning. This and similar research on the diversity on this complex cognitive capacity can in turn help with reconstructing a better scientific terminology to describe it.

Appendix A

The Canteen Dilemma

Time left to complete this page: **0:51**

These instructions will also be shown on the following pages.

Instructions for the game:

This game is about trying to do the same as your colleague.

Every morning you arrive at work between 8:00 am and 9:10 am. You and your colleague will arrive by bus 10 minutes apart. Example: You arrive at **8:40 am**. Your colleague may arrive at **8:30 am**, or **8:50 am**.

Both of you like to meet in the canteen for a coffee. If you arrive before 9:00 am, you have time to go to the canteen, but you should only go if your colleague goes to the canteen as well. If you or your colleague arrive at 9:00 am or after, you should go straight to your offices.

At the beginning of each round you will know only your own arrival time. You will have to decide whether to go to the canteen or to the office. As a general rule, you will maximize your payoff by honestly choosing the option you think your colleague will also choose.

Payoff and penalties:

You start the game with $10.00 and will have to pay various amounts of penalties in each round, depending on how well you both do. Your challenge is to have as much money left as possible when the game ends, after which the remaining amount is paid out to you as a bonus. The game ends after 10 rounds or when you or your colleague has no money left.

- **Both go to canteen**
  If you guessed correctly that both of you went to the canteen before 9:00 am, you pay a **small** penalty proportional to how **uncertain** you were, e.g.:
  - **-$0.69** if you were very uncertain.
  - **-$0.29** if you were somewhat certain.
  - **-$0.01** if you were very certain.
- **Both go to office**
  If you guessed correctly that both of you went to your offices, no matter what time, your penalty is **doubled** and proportional to how **uncertain** you were, e.g.:
  - **-$1.39** if you were very uncertain.
  - **-$0.58** if you were somewhat certain.
  - **-$0.02** if you were very certain.
- **One goes to the canteen, the other to the office**
  If you guessed incorrectly and one of you went to the canteen while the other went to the office - or if any of you went to the canteen at 9:00 am or after, your penalty is **doubled** and proportional to how **certain** you were, e.g.:
  - **-$1.39** if you were very uncertain.
  - **-$2.77** if you were somewhat certain.
  - **-$9.21** if you were very certain.
- In summary, try to do your best doing the same as your colleague. As a general rule you will minimize your losses by giving an honest estimate of the chances of doing the same as your colleague

Next

Appendix B

Question 1: "The game is over. Do you think it was your fault it is over, your colleagues fault, or do you think it was because of some other reason?" with the possible answers being "Yes", "No" and "Other reason".

Question 2: "What strategy did you use while playing this game?", where participants could answer in free text.

Question 3: "Imagine you could have agreed beforehand with your colleague about a point in time where it is safe to go to the canteen. What time would that be?". The possible answers where: "I don't know", "There is no such time", "8:00", "8:10", "8:20", "8:30", "8:40", "8:50", "9:00" and "9:10".

Question 4: "Imagine you arrive at 8:00 am. Is it common knowledge between you and your colleague that it is safe to go to the canteen, that is, you both arrived before 9:00 am? ". The possible answers were: "Yes", "No", "I do not know".

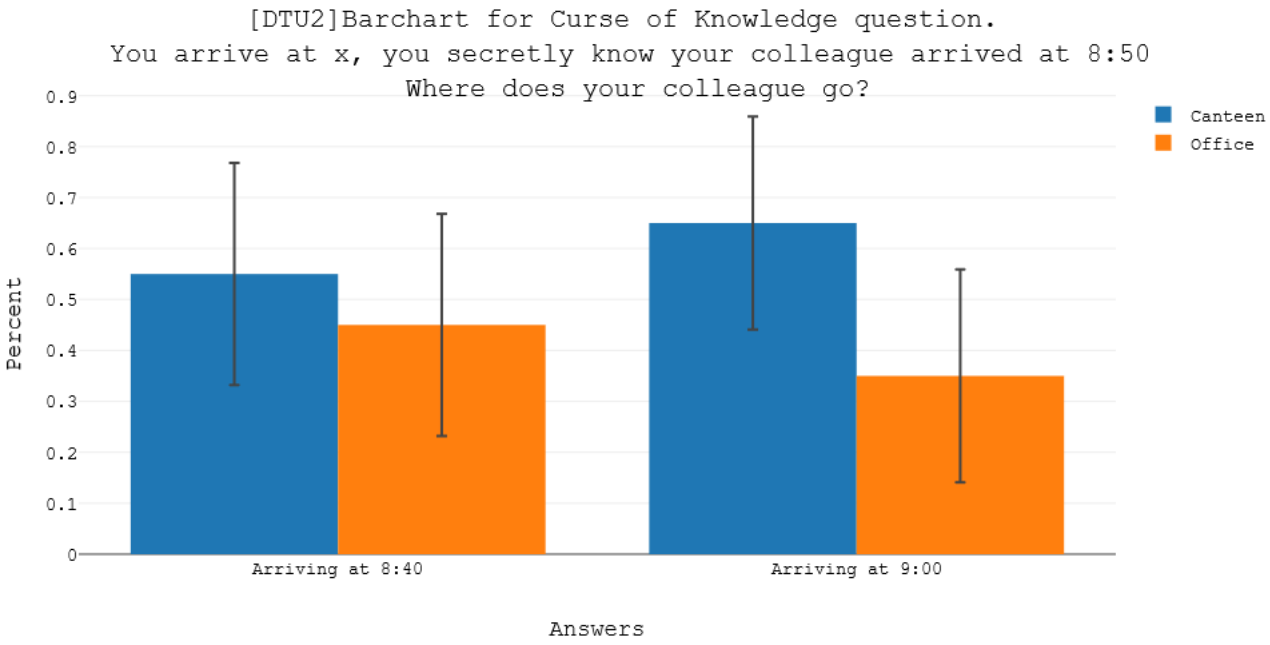
The DTU trials included the further questions:

Question 5: "Did you ever go to the canteen at an arrival time later than what was safe according to your previous answer? Why or why not?" Free text answer

Question 6: "Did you ever choose differently after seeing the same arrival time again at a later point in the game? Why or why not?" Free text answer

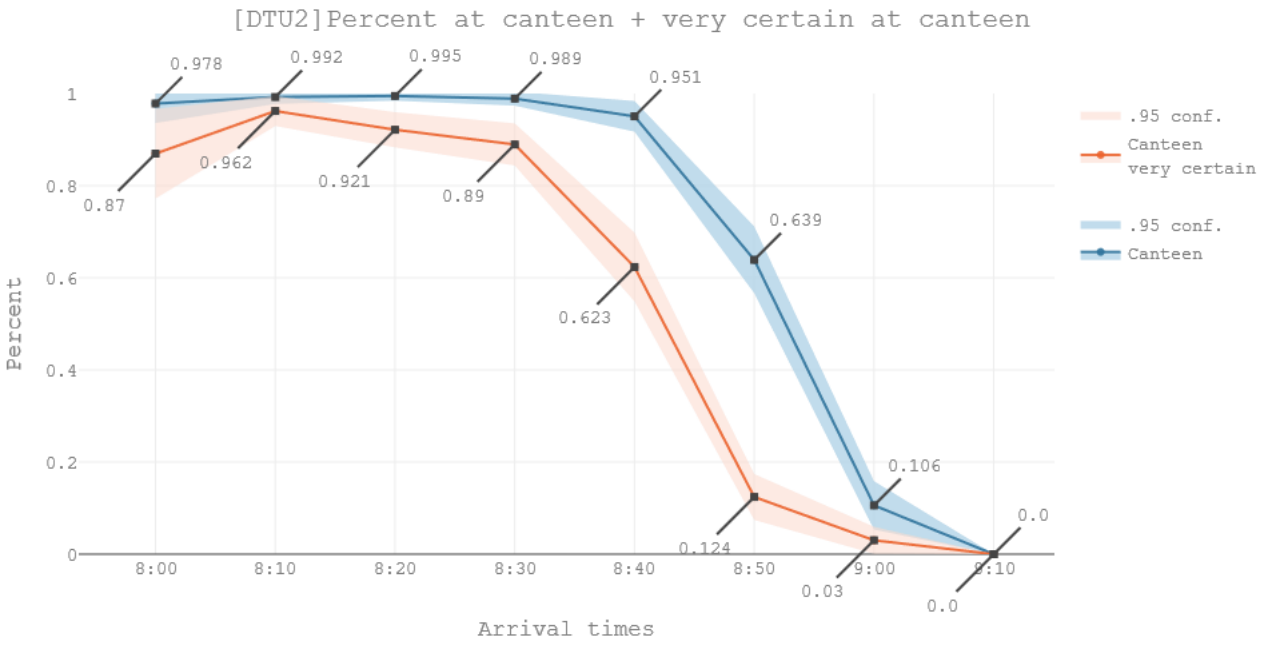
Question 7: "Imagine you arrived at 9:00//8:40 and you have been secretly informed that your colleague's arrival time is 8:50. Where do you think your colleague will go?" Possible answers canteen/office. Half the participants got 8:40 and the other got 9:00

Appendix C

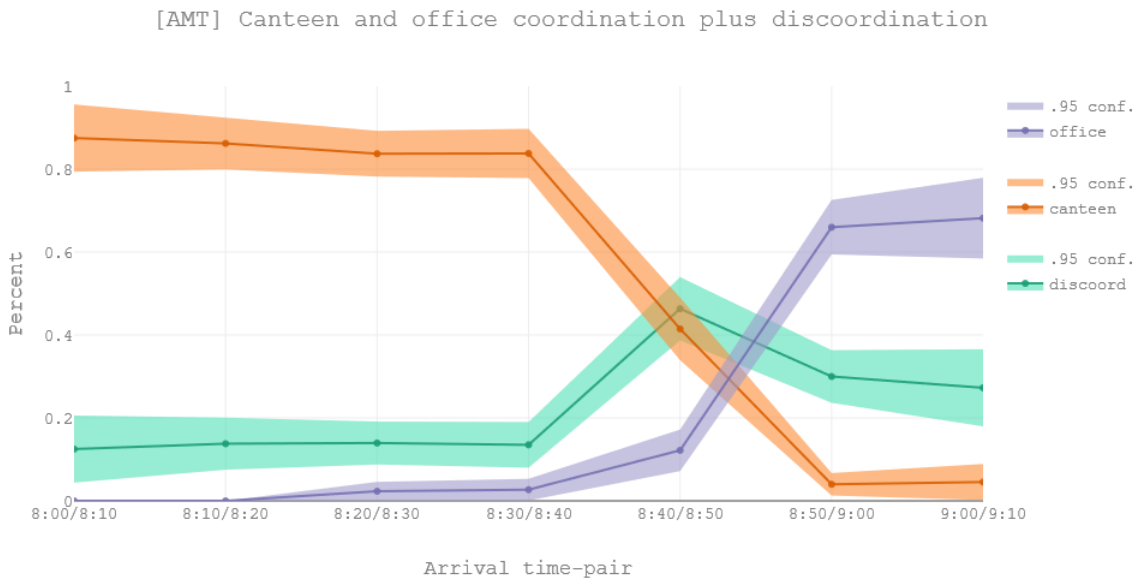


Grouped bar-chart for curse of knowledge question from second DTU trial.

Appendix D


[DTU2]Percent at canteen + very certain at canteen
Percent
Arrival times
1
0.8
0.6
0.4
0.2
0
8:00
8:10
8:20
8:30
8:40
8:50
9:00
9:10
0.978
0.992
0.995
0.989
0.951
0.639
0.106
0.0
0.87
0.962
0.921
0.89
0.623
0.124
0.03
0.0
.95 conf.
Canteen very certain
.95 conf.
Canteen

Appendix E



Line-plot for percent of coordination in the canteen (orange) and the office (blue) as well as discoordination (green) per arrival time-pair.

References

[1] Anderson, R. L. (2005). *Neo-Kantianism and the Roots of Anti-Psychologism*, British Journal for the History of Philosophy, 13:2, 287-323, DOI: 10.1080/09608780500069319

[2] Bacharach, M., & Stahl, D. O. (2000). *Variable-frame level-n theory.* Games and Economic Behavior, 32(2), 220–246.

[3] Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). *Does the autistic child have a 'theory of mind'?* Cognition, 21, 37–46.

[4] Baron-Cohen, S. et al. (1999). *Social intelligence in the normal and autistic brain: an fMRI study.* Eur. J. Neurosci. 11, 1891–1898

[5] Barwise, J. (1989). *On the model theory of common knowledge.* In: The situation in logic (pp. 201–221). Stanford: CSLI.

[6] van Benthem, J. F. A. K. (2003). *Logic and the Dynamics of Information.* Minds and Machines 13: 503–519, Kluwer Academic Publishers

[7] van Benthem, J. F. A. K. (2007a). *Cognition as interaction.* In Proceedings symposium on cognitive foundations of interpretation (pp. 27–38). Amsterdam: KNAW.

[8] van Benthem, J. F. A. K., Gerbrandy, J., & Pacuit, E. (2007). *Merging frameworks for interaction: DEL and ETL.* In D. Samet (Ed.), Theoretical aspects of rationality and knowledge: Proceedings of the eleventh conference, TARK 2007 (pp. 72–81). Louvain-la-Neuve: Presses Universitaires de Louvain.*

[9] van Benthem, J. F. A. K., Hodges, H., & Hodges, W. (2007b). *Introduction.* Topoi, 26(1), 1–2. (Special issue on logic and psychology, edited by J.F.A.K. van Benthem, H. Hodges, and W. Hodges.).*

[10] van Benthem, J. F. A. K. (2008). *Logic and reasoning: Do the facts matter?* Studia Logica, 88, 67–84. (Special issue on logic and the new psychologism, edited by H. Leitgeb)

[11] van Benthem, J. F. A. K (2010). *Modal logic for open minds.* CSLI Publications.

[12] Benz, A., & van Rooij, R. (2007). O*ptimal assertions, and what they implicate. A uniform game theoretic approach.* Topoi, 26(1), 63–78 (Special issue on logic and psychology, edited by J.F.A.K. van Benthem, H. Hodges, and W. Hodges.)*

[13] Berinsky, A., Huber, G., & Lenz, G. (2012). *Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. Political Analysis.* 20(3), 351-368. doi:10.1093/pan/mpr057

[14] Birch, S. A. J., Bloom, P. (2007). *The curse of knowledge in reasoning about false beliefs.* Psychol Sci. 2007 May; 18(5): 382–386. doi: 10.1111/j.1467-9280.2007.01909.x

[15] Buhrmester, Michael & Kwang, Tracy & Gosling, Samuel. (2011). *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?.* Perspectives on Psychological Science. 6. 3-5. 10.1177/1745691610393980.

[16] Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). *An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use.* Perspectives on Psychological Science, 13(2), 149–154. https://doi.org/10.1177/1745691617706516

[17] Castelfranchi, C. (2004). *Reasons to believe: cognitive models of belief change.* Ms. ISTC-CNR, Roma. Invited lecture, Workshop Changing Minds, ILLC Amsterdam, October 2004. Extended version. Castelfranchi, Cristiano and Emiliano Lorini, The cognitive structure of surprise.Costa-Gomes, M., Weizsäcker, G., (2008). Stated beliefs and play in normal form games. Review of Economic Studies 75, 729–762.

[18] Chandler, M., Fritz, A. S., & Hala, S. (1989). *Small-scale deceit: deception as a marker of 2-, 3-, and 4-year-olds' early theories of mind.* Child Development, 60, 1263–1277.

[19] Cheng P.W., Holyoak K.J, Nisbett R.E., Oliver L.M. (1986). *Pragmatic versus syntactic approaches to training deductive reasoning.* Cogn. Psychol. 18:293–328

[20] Chen, D.L., Schonger, M., Wickens, C., 2016. *oTree - An open-source platform for laboratory, online and field experiments.* Journal of Behavioral and Experimental Finance, vol 9: 88-97

[21] Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). *Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist.* Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362, 507–522.

[22] Crump M. J. C, McDonnell J. V., Gureckis T. M. (2013). *Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research.* PLoS ONE 8(3): e57410. https://doi.org/10.1371/journal.pone.0057410.

[23] Csibra, G., Gergely, G., Biro, S., Koos, O., & Brockbank, M. (1999). *Goal attribution without agency cues: the perception of 'pure reason' in infancy.* Cognition, 72, 237–267.*

[24] van Ditmarsch, H., & Labuschagne, W. (2007). *My beliefs about your beliefs: A case study in theory of mind and epistemic logic.* Synthese: Knowledge, Rationality and Action, 155, 191–209.

[25] van Ditmarsch, H., van der Hoek, W., Kooi, B. (2008). *Dynamic Epistemic Logic.* Synthese Library, Springer Netherlands.

[26] van Ditmarsch H., Kooi B. (2015) *Consecutive Numbers.* In O*ne Hundred Prisoners and a Light Bulb.* Copernicus, Cham.

[27] Donkers, H. H. L. M., Uiterwijk, J. W. H. M., & van den Herik, H. J. (2005). *Selecting evaluation functions in opponent-model search.* Theoretical Computer Science, 349, 245–267.*

[28] Dunin-Keplicz, B., & Verbrugge, R. (2006). *Awareness as a vital ingredient of teamwork.* In P. Stone, & G. Weiss (Eds.), Proceedings of the fifth international joint conference on autonomous agents and multiagent systems (AAMAS'06) (pp. 1017–1024). New York: IEEE / ACM.*

[29] van Eijck, J., & Verbrugge, R. (Eds.) (2009). *Discourses on social software.* Texts in games and logic (Vol. 5). Amsterdam: Amsterdam University Press.

[30] Erb, Benjamin. (2016). *Artificial Intelligence & Theory of Mind.* 10.13140/RG.2.2.27105.71526.

[31] Fagin, R., & Halpern, J. (1988). *Belief, awareness, and limited reasoning.* Artificial Intelligence, 34, 39–76.*

[32] Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge.* 2nd ed., 2003. Cambridge: MIT.

[33] Flobbe, L., Verbrugge, R., Hendriks, P.,&Krämer, I. (2008). *Children's application of theory of mind in reasoning and language.* Journal of Logic, Language and Information, 17, 417–442. (Special issue on formal models for real people, edited by M. Counihan.)

[34] Frege, G. (1964 [1893]). *The Basic Laws of Arithmetic: Exposition of the System*, M. Furth (trans.), Berkeley, CA: University of California Press.

[35] Frege, G. (1897). *Logic*, reprinted in Frege [1997], pp. 227–250.

[36] Frege, G. (1997). *The Frege reader* (M. Beaney, editor), Blackwell, Oxford.

[37] Frith, U. and Happe, F. (1994). *Autism: beyond 'theory of mind'.* Cognition 50, 115–132.

[38] Ghosh, S., Meijering, B., & Verbrugge, R. (2014). *Strategic reasoning: Building cognitive models from logical formulas.* Journal of Logic, Language and Information, 23(1), 1–29.

[39] Ghosh, S., Heifetz, A., & Verbrugge, R. (2015). *Do players reason by forward induction in dynamic perfect information games?.* TARK.

[40] Ghosh, S., Meijering, B. & Verbrugge, R. (2018). *Studying strategies and types of players: experiments, logics and cognitive models.* Synthese (2018) 195: 4265. https://doi.org/10.1007/s11229-017-1338-7

[41] Gierasimczuk, N., Hendricks, V. F., Jongh, D. d. (2014). *Logic and Learning.* In Johan van Benthem on Logic and Information Dynamics, Baltag, Alexandru, Smets, Sonja (Eds.). Outstanding Contributions to Logic, Vol. 5. Dordrecht: Springer.

[42] Gigerenzer, G., Todd, P., & The ABC Research Group. (1999). *Simple Heuristics that Make us Smart.* New York: Oxford University Press.

[43] Gray, K. et al. (2011). *Distortions of mind perception in psychopathology.* Proc. Natl. Acad. Sci. U.S.A. 108, 477–479

[44] Griggs R.A., Cox J.R. (1982). *The elusive thematic-materials effect in Wason's selection task.* Br J Psychol 73:407–420

[45] Halpern, J. Y., & Moses, Y. (1990). K*nowledge and common knowledge in a distributed environment.* Journal of the ACM, 37, 549–587.*

[46] Harbers, M., Verbrugge, R., Sierra, C., & Debenham, J. (2008). *The examination of an information-based approach to trust.* In P. Noriega, & J. Padget (Eds.), Coordination, organizatiosn, institutions and norms in agent systems III. Lecture notes in computer science (Vol. 4870, pp. 71–82). Berlin: Springer.*

[47] Hedden, T., & Zhang, J. (2002). *What do you think I think you think? Strategic reasoning in matrix games.* Cognition, 85, 1–36.

[48] Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). *Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis*. Science, 317, 1360–1366.

[49] Heyes, C. (2014). *Submentalizing: I am not really reading your mind*. Perspect. Psychol. Sci. 9, 131–143

[50] Heyes, C., & Frith, C. D. (2014b). *The cultural evolution of mind reading*. Science. Jun 20;344(6190):1243091. doi: 10.1126/science.1243091

[51] Horton, J.J., Rand, D.G. & Zeckhauser, R.J. (2011). *The online laboratory: conducting experiments in a real labor market*. Experimental Economics, Sep. 2014, Vol. 14: 399. https://doi.org/10.1007/s10683-011-9273-9

[52] Humphrey, N.K. (1980). *Nature's psychologists*. In Consciousness and the physical world (eds B. D. Josephson & V. S. Ramachandran), pp. 57–80. Oxford, UK: Pergamon Press.

[53] Husserl, E. (1970 [1900]). *Logical Investigations*. J. N. Findlay (trans.), London: Routledge & Kegan Paul.

[54] Isaac, A. M. C., Szymanik, J., & Verbrugge, R. (2014). *Logic and complexity in cognitive science*. In Johan van Benthem on Logic and Information Dynamics (pp. 787–824). Springer.

[55] Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge: MIT.

[56] Keysar, B. & Lin, S. & J Barr, D. (2003). *Limits on theory of mind use in adults*. Cognition. 89. 25-41. 10.1016/S0010-0277(03)00064-7.

[57] van Lambalgen, M., & Counihan, M. (2008). *Formal models for real people*. Journal of Logic, Language and Information, 17, 385–389. (Special issue on formal models for real people, edited by M. Counihan).

[58] Leslie, A. (2000). *How to acquire a 'representational theory of mind'*. In D. Sperber & S. Davies (Eds.), Metarepresentation, Oxford: Oxford University Press.

[59] Lin, S., Keysar, B., Nicholas, E. (2010). *Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention*. Journal of Experimental Social Psychology Volume 46, Issue 3, May 2010, Pages 551-556.

[60] Liu, F. (2008). *Diversity of Agents and Their Interaction*. Springer Netherlands.

[61] McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *Proposal for the Dartmouth summer research project on artificial intelligence*. Technical report, Dartmouth College.

[62] Mason, Winter & Watts, Duncan. (2009). *Financial incentives and the performance of crowds*. SIGKDD Explorations. 11. 100-108. 10.1145/1600150.1600175.

[63] Maddy, P. (2012). *The philosophy of logic*. Bulletin of Symbolic Logic 18 (4):481-504.

[64] Meijering, B.,Maanen, L. v., Rijn,H. v., & Verbrugge, R. (2010). T*he facilitative effect of context on secondorder social reasoning*. In Proceedings of the 32nd annual meeting of the cognitive science society, (pp. 1423–1428). Philadelphia, PA, Cognitive Science Society.

[65] Mol, L. (2004). *Learning to reason about other people's minds.* Technical report, Institute of Artificial Intelligence, University of Groningen, Groningen. Master's thesis.

[66] Pacuit, E., Parikh, R., & Cogan, E. (2006). *The logic of knowledge based obligation.* Synthese: Knowledge, Rationality and Action, 149, 57–87.

[67] Palfrey, T., & Wang, S. (2009). *On eliciting beliefs in strategic games.* Journal of Economic Behavior & Organization, 71(2), 98-109.

[68] Parikh, R. (2003). *Levels of knowledge, games, and group action.* Research in Economics, 57, 267–281.

[69] Perner, J. (1988). *Higher-order beliefs and intentions in children's understanding of social interaction.* In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), Developing theories of mind (pp. 271–294). Cambridge: Cambridge University Press.

[70] Premack, D., & Woodruff, G. (1978). *Does the chimpanzee have a theory of mind?* Behavioral & Brain Sciences, 1, 515–526.

[71] Putnam, H. (1978). *There is at least one a priori truth.* Erkenntnis 13 (1978) 153-170.

[72] Quine, W. V. O (1951). *Two dogmas of empiricism.* Reprinted in his From a logical point of view, second ed., Harvard University Press, Cambridge, MA, 1980, pp. 20–46.

[73] Qureshi, A. W., Apperly, I. A., Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. Cognition 117, 230–236 (2010). doi: 10.1016/j. cognition.2010.08.003; pmid: 20817158

[74] Rand, David. (2011). *The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments.* Journal of theoretical biology. 299. 172-9. 10.1016/j.jtbi.2011.03.004.

[75] Rosenthal, R. (1981). *Games of perfect information, predatory pricing, and the chain store.* Journal of Economic Theory, 25, 92–100.

[76] Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2014). *Deconstructing and reconstructing theory of mind.* Trends in cognitive sciences, 19(2), 65–72. doi:10.1016/j.tics.2014.11.007

[77] Seidenfeld, T., 1985. *Calibration, coherence, and scoring rules.* Philosophy of Science 52, 274–294.

[78] Stahl, D. O., & Wilson, P. W. (1995). *On players' models of other players: Theory and experimental evidence.* Games and Economic Behavior, 10, 218–254.

[79] Stenning K, van Lambalgen M. (2008).*Human reasoning and cognitive science.* MIT Press, Cambridge

[80] Stulp, F., & Verbrugge, R. (2002). *A knowledge-based algorithm for the internet protocol TCP.* Bulletin of Economic Research, 54(1), 69–94.

[81] Sycara, K. & Lewis, M. (2004). *Integrating intelligent agents into human teams.* In E. Salas,& S. Fiore (Eds.), Team cognition: Understanding the factors that drive process and performance (pp. 203–232). Washington, DC: American Psychological Association. 133.

[82] Veltman, K.H., Weerd, H.D., & Verbrugge, R. (2018). *Training the use of theory of mind using artificial agents.* Journal on Multimodal User Interfaces, 1-16.

[83] Verbrugge, R., & Mol, L. (2008). *Learning to apply theory of mind.* Journal of Logic, Language and Information, 17, 489–511. (Special issue on formal models for real people, edited by M. Counihan.)

[84] Verbrugge R. (2009): *Logic and Social Cognition.* Journal of Philosophical Logic.

[85] Vogeley, K. et al. (2001). *Mind reading: neural mechanisms of theory of mind and self-perspective.* Neuroimage 14, 170–181

[86] Wason, P. C. (1966). *Reasoning.* In B. M. Foss (Ed.), New Horizons in Psychology I, (pp. 135–151). Harmondsworth: Penguin.

[87] Wason P.C., Shapiro D. (1971). *Natural and contrived experience in a reasoning problem.* Q J Exp Psychol 23:63–71

[88] Wason, P., & Shapiro, D. (1971). *Natural and contrived experience in a reasoning problem.* The Quarterly Journal of Experimental Psychology, 23(1), 63–71.

[89] Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

[90] Wimmer, H., & Perner, J. (1983). *Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception.* Cognition, 13, 103–128.

[91] Wooldridge, M. J. (2002). *An introduction to multiagent systems.* Chichester: Wiley.

[92] http://www.glascherlab.org/social-decisionmaking/ (visited 05-05-2019)

[93] https://plato.stanford.edu/entries/epistemology-bayesian/ (visited 05-05-2019)