The Canteen Dilemma - An Experiment on Higher Order Social Reasoning

Thomas S. Nicolet

March 25, 2019

Abstract

Many complex social interactions require people to have mental model of those around them. Being able to reason about the beliefs, intentions and knowledge of others allows people to better understand the actions and utterances of others, as well as helping with possible predictions of others. While this ability is generally developed in young children, numerous studies show even adults failing to reason about the mental states of others, or consult existing relevant knowledge. The Canteen Dilemma tests participants ability to reason about the mental states of others in order to predict their actions. While participants generally do not exhibit any Theory of Mind when asked for spontaneous decisions, their certainty estimate of success somewhat reflects a ToM.

Contents

1 Introduction

1.1 Logic and the facts - from Frege and Carnap to Quine and Benthem.

1.2 Epistemic Logic - normative aspect

1.2.1 Consecutive numbers example

1.3 Real higher order social reasoning

1.3.1 Why Theory of Mind is important for social interaction

1.3.2 Adults difficulties with applying adequate higher order social reasoning

1.3.3 Theory of Mind: Reflective versus spontaneous

2 Canteen Dilemma experiment

The Canteen Dilemma is two-player co-ordination game modeled similarly to the Consecutive Numbers example. The game gives the Consecutive Numbers example a thematic context by creating a narrative of two colleagues arriving at work every morning. Instead of being assigned a number, they are assigned a time they arrive at work. There are eight possible arrival times ranging from 8:00 am to 9:10 am in ten minute intervals. Like the Consecutive Numbers example, participants know their own arrival time and that the other person always has an arrival time ten minutes before or after their own. The game consists of a series of rounds where participants have to decide between going to the canteen and the office. They also have to give an estimate of how certain they are that the other player made the same choice as them, but I'll ignore that for now and come back to it later. The payoff structure is based on a logarithmic scoring rule, so instead of getting payoffs, participants start with a bonus which is reduced every round depending on how well they do. The narrative of arriving at work is continued: if both players arrived before 9:00 am, they have time to go to the canteen, which also nets the smallest penalty. Both players going to the office together results in twice the penalty than canteen. Going to different places, or going to the canteen at 9:00 am or later is the worst option however and results in a much larger penalty.

To iterate, there are 8 possible arrival times, which entails 14 unique possible arrival-time pairs. The game puts players at a dilemma in the following way. Players prefer to meet in the canteen before 9:00 am, and there are plenty of chances for doing so, since 10 of the possible 14 possible arrival-time pairs allows this. Imagine a player want to go

for meeting the other player in the canteen as much as possible, i.e. in all the 10 cases. One of the 10 cases consists of the arrival pair (8:50, 8:40). But when a player arrives at 8:50, they do not know if the other player arrived at 8:40 (meaning canteen would be the lowest payoff) or 9:00 (meaning canteen would be the lowest payoff). Call this the set *late*: {(8:50, 9:00), (9:00, 8:50), (9:00, 9:10), (9:10, 9:00)}), where a player is punished if they choose canteen no matter what. A player might attempt to avoid this dilemma by deciding to not go to the canteen when their arrival time is contained in the *late* set, i.e. they'll choose the office whenever they arrive at 8:50. But if they do this, they might assume the other player would do the same. This repeats the dilemma at 8:40, since this means that the other player might have arrived at 8:50 and gone to the office. This line of backwards induction reasoning applies to all possible arrival times, which entails the unintuitive conclusion that the only safe strategy is for both players to go to the office no matter what time it is.

This is due to the lack of common knowledge about the arrival time of the players. However, it's highly unintuitive to decide to go to the office when arriving at 8:00 am, after having been told that the highest payoff is going to the canteen together when you arrive before 9:00 am, and you know the following: you both arrived before 9:00 am, the other player know you arrived before 9:00 am, the other player knows you know you both arrived before 9:00 am (and so on for two more iterations). See also [24] for empirical evidence that people are more likely to act in accordance with forward inductive reasoning rather than backward inductive reasoning. In general, reasoning by backwards induction ignores previous moves in a game, while forwards induction takes these into account.

We used logarithmic scoring as a proper scoring rule, in relation to asking players how certain they were that the other player made the same choice as them. Research literature on eliciting belief also show 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" [43]. 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 [38].

Asking participants for a certainty estimate also has other benefits. It helps alleviate some of what Verbrugge calls "the danger of formal systems that posit a fixed bound on social cognition: everyone can reason up to order n but not on order $n + 1$" [46]. This is because it allows differentiation between two players both choosing e.g. canteen. There might be a distinction in terms of applied ToM between those who choose canteen and

those who do not. But there might also be a distinction within the group that chooses e.g. canteen in terms of their estimated certainty. Now for the experiment.

In order to gather more information about the reasoning of the participants, they were asked a few post-game questions. If a game ended because a player had lost all of their bonus, participants were asked whose fault it was that the game ended. Participants where asked three other questions when the game ended, regardless the reason. These questions pertained to (i) what strategy the player had, (ii) what time they deemed it safe to go to the canteen and (iii) if it was common knowledge that they had arrived before 9:00 am, given an arrival time of 8:00 am, i.e. their intuition about the concept of common knowledge.

2.1 Method and design

2.1.1 Experimental environment

The experimental setup was implemented in oTree 2.1.35 software [13] and conducted on Amazon Mechanical Turk (AMT) which is an online crowdsourcing platform. AMT works as an online labor market where workers can perform human intelligence tasks (HITs) 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 [9][11] and to provide reliable, replicable and more diverse data than legacy methods using American college students [14][16][29][34][41].

2.1.2 Participants and AMT Settings

The experiment included 714 participants. All participants gave informed consent to participate before being included. AMT also allows the researchers to apply a few different settings. Only turkers from Canada or the United States were able to enter the hit. Further restraints were used in order to only include those with an approved HIT rate of 98% or above and at least 500 completed HITs. 41 of the 714 did not play the full game due not making a required choice in time. The experiment was run in batches of 50 participants at a time within a timespan of a maximum of one hour. Participants had access to our emails during the experiment for questions and feedback.

2.1.3 Payoff structure

Participants received a participation fee of $2. Besides this, the experiment utilizes a logarithmic scoring rule which means the payoff is calculated as the logarithm of the

probability estimate for the actual outcome, where a 99% probability estimate of an event occurring would mean assigning 1% estimate to the event not occurring. Participants were also given an initial $10 bonus which is reduced during the game. Every round results in a penalty which is taken out of their bonus depending on the choices they make and how certain they are that their colleague has made the same choice. Payoffs range from $0.01 to $9.21.

If both players go to the canteen, their penalty is $\ln(player - certainty)$. If they both go to the office it is doubled: $\ln(player - certainty) \cdot 2$. If they go to different places it is $\ln(1 - player - certainty) \cdot 2$. Players who dropped out due to time limit received no bonus or participation fee, while the player matched to them received their participation fee.

2.1.4 Materials and set-up

After a participant had provided informed consent, participants were directed to a waiting page until they were matched with another player. After the groups were formed, players were directed to an initial introduction page, detailing the rules of the game, quoted below (see Appendix A for screenshot). This page was shown for 90 seconds before automatically starting the first round (the initial intended time limit was 240 seconds, but it was not used due to an unintended problem). The instructions were also shown on every subsequent round which lasted 61 seconds, where players were given an arrival time and had to decide to go to the canteen or the office.

> 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

After clicking next or the timer running out, participants were directed to a new page (see Appendix B) with the corresponding round number. They were assigned an arrival time from the set {8:00, 8:10, 8:20, 8:30, 8:40, 8:50, 9:00, 9:10}. The page had a time limit at 61 seconds. The 'instructions' from the previous page was shown below as well. After stating their given arrival time, particiapnts where asked the following question, assuming they were assigned 8:50 am: "It is Monday morning and you arrive at 8:50 am. Where will you go?". The possible answers being "Canteen" and Office", which had to be selected before a "next" button could be clicked, prompting the participant to the next page.

After making the canteen/office decision, the next page (see Appendix C) asked them: "How certain are you that your colleague has made the same choice as you?" This was based on a 5 point Likert scale, with possbile answers being very uncertain, slightly certain, somewhat certain, quite certain and very certain. The possible answers were used for the logarithmic proper scoring by assigning the probabilities {0.5, 0.625, 0.75, 0.875, 0.99} to each answer, respectively, which was used for the logarithmic scoring rule.

When participants had made their choice between the canteen and office and had given a probabiltiy estimate, they were prompted to a page showing a table for running results (See Appendix D). What round number they had just played, what time they had arrived including their choice and certainty estimate. It also showed the arrival time of the other player in all previous rounds including their decision. The penalty for each round was also shown. They were explicitly told below the table if they had both chosen to go to the canteen or office together, if they had made different choices, or if it had been too late to go to the canteen (a player had arrived at 9:00 am). It also stated the penalty for that round, their total losses and their remanining bonus. This running results page had a time limit at 30 seconds, which would automatically prompt a new round when running out.

When the game was over, participants were asked a few questions.

Question 1. If either player lost their entire bonus, both players where asked the following: "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 "I do not know".

Question 2. When the game had ended either due to someone losing their bonus or after 10 rounds, players were asked the question: "What strategy did you use while playing this game?", where participants could answer in free text.

Question 3. When the game had ended either due to someone losing their bonus or after 10 rounds, players were asked the question: "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. When the game had ended either due to someone losing their bonus or after 10 rounds, players were asked the question: "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".

After answering the last questions, participants were thanked for participating in the Canteen Dilemma experiment and provided with one of the researchers email for further inquiries.

2.2 Results

2.2.1 Percent going to canteen

The most basic results were the choices participants made between canteen and office. We first show a plot for how many percent chose canteen at each arrival time-

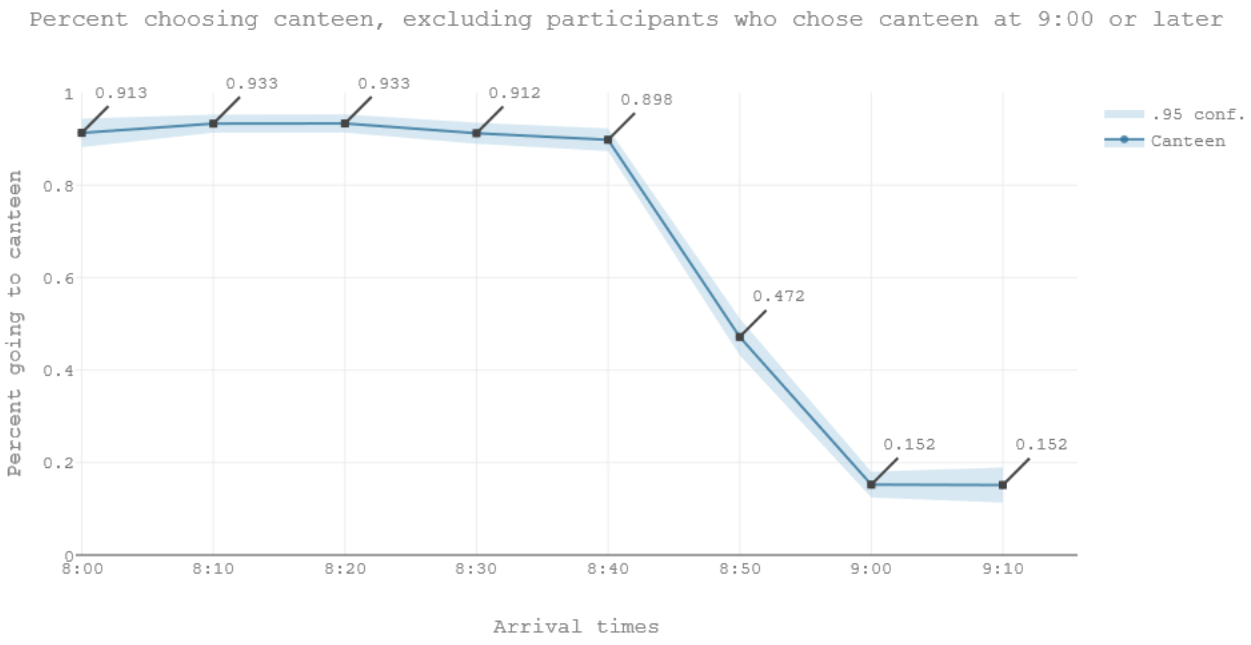


Figure 1. Percent going to canteen with confidence intervals and annotations

Figure 1 shows that participants generally (90-93% of the time) went to the canteen when arriving before 8:50. Slightly less than half (47.2%, .95 confidence interval 43.2% - 51.1%) of the participants chose the canteen at 8:50. More problematic however, is the fact that 15% (98 participants) chose the office at 9:00. This is not due to a lack of ToM or any reasoning process, since choosing the canteen at 9:00 always resulted in the biggest penalty. We will consider a few explanations for this. There is (1), the background 'noise' which is constituted by humans making mistakes, deliberately or not. No matter how simple a task you would set up on AMT stating *not* to do something, there would be someone doing it, either unintentionally clicking the wrong button, picking at random, or perhaps even some intentionally *trolling*. This is unlikely to account for the entire 15% choosing canteen however. (2) The possibly largest attributing factor could be the formulation of the rules not stating clearly enough to avoid the canteen when arriving at 9:00 or later. (3) Given the narrative of the game, going to the canteen or the office, it is possible some are biased towards the canteen due to their real-life preference of going to the canteen over the office. This is supported by some strategy answers, e.g. "I just went with what I normally do in real life". (4) Time constraints possibly compounded these problems.

The problematic upshot of this is that, if a player believe they can go to the canteen at 9:00 and get rewarded as long as the other player does the same, it might affect their choice at 8:50. That is, a player might believe that if they arrive at 8:50, they can be rewarded for going to the canteen even if the other player arrives at 9:00, as long as the other player chooses canteen as well. This possibly has the effect that those arriving at 8:40 are more inclined to chose the canteen as well, and so on. This is not strongly empirically supported however. If we filter out data-set to exclude all those participants who chose canteen at 9:00 or 9:10, we get the following plot.

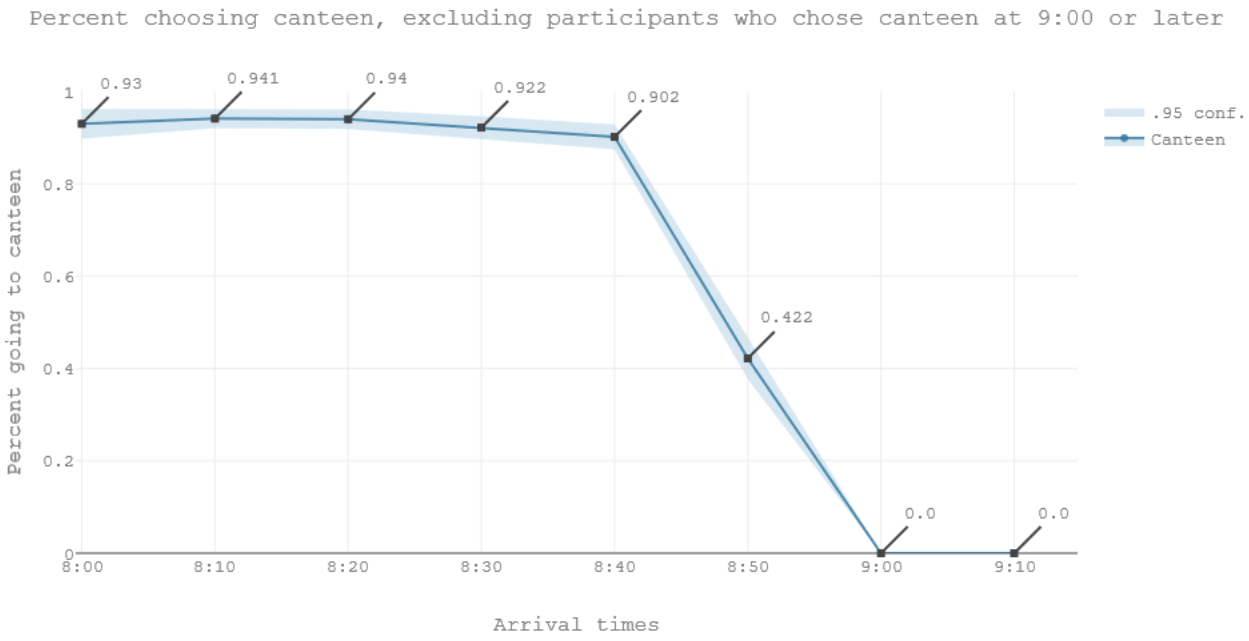


Figure 2. Line-chart showing percent choosing canteen at specific arrival times, filtered to exclude participants who chose canteen at 9:00 or later, with shaded error bars.

Instead of a confidence interval of 43.2 - 51.1 percent choosing the canteen at 8:50, we now have 37.7% - 46.7%. This is slightly lower (5%) but also with a 3% overlap. This means that even of the participants who knew they themselves would chose office if they arrived at 9:00, 42% of them still chose to go to the canteen at 8:50. Since they themselves always choose the office at 9:00, it is implausible that they choose canteen at 8:50 because they believe that the other player might choose canteen even when arriving at 9:00. This indicates that they either take a chance, hoping the other player would not arrive at 9:00, or that they were unaware of their preference of choosing the same as their colleague. Since players have the opportunity to hedge their decisions with an estimate of how certain they are, it's plausible that they simply took the chance, and then lowered their estimated certainty accordingly. We will now go through the results for the various post-game questions, before returning to estimated certainty.

2.2.2 Question pertaining to fault

We will first look at the data for the question pertaining to whose fault it was that the game ended prematurely.

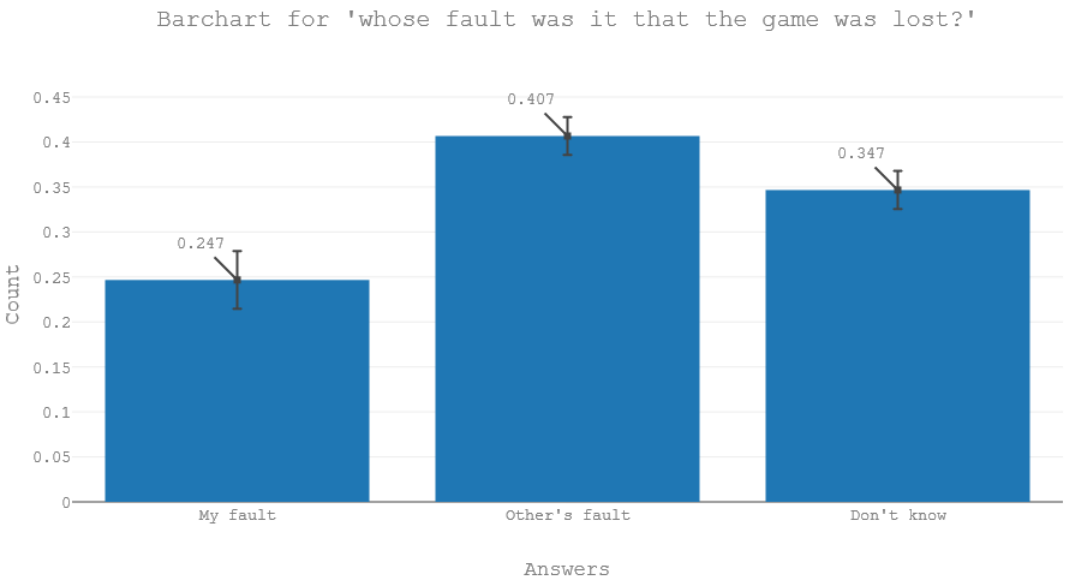


Figure 3. Bar-chart showing percent choosing different answers.

The question is inherently complex given the structure of the game. In most cases of dis-coordination, a player would not be certain to have cooperated with the other player, even if one of them had made a different choice. Imagine player a always goes to the office at 9:00, and meets player b who goes to the canteen at 8:50, leading to dis-coordination. However, if player b changes their strategy to go to the office at 8:50, she might miss-coordinate with a when a arrives at 8:40 and chooses canteen. There are however significantly more participants responding that it was the other players fault that the game ended, rather than their own fault or stating ignorance. This is curious since if there is an objective reason such that one player could be said to have made a mistake rather than the other, then the player who made the mistake would ideally answer that it was their own fault, while the other player would answer it was the others fault. In other words, there would be an equal amount of those answers under such scenarios.

Consider the the previous plots, where we see that there was a lot of dis-coordination at arrival time 8:40, 8:50 and 9:00. This is because players mostly chose canteen at 8:40 and mostly office at 9:00, while they were paired with those arriving at 8:50 who chose office only a bit more than half of the time. Those arriving at 9:00 had no choice but

choosing office, and if they dis-coordinate with a player arriving at 8:50, it is likely that they say it was the other players fault the game ended, since they could not have made a different choice. Those arriving at 8:40 and choosing canteen while the other player chooses office at 8:50 might be bewildered, assuming 8:40 was a safe time to go to the canteen, while the other thought it was clear that arriving at 8:50 would entail going to the office. In both cases they say it was the other's fault that the game ended. The majority answering 'other's fault' is arguably due to a difference in application of ToM. If one participant relies on her ToM and the other doesn't, the person not relying on ToM is likely not aware of its importance, and they therefore both claim it was the other's fault. However there is a caveat. Given the possibility of answering ignorance, it's possible that some of those that might truthfully believe it was their own fault that the game ended chose the option of ignorance rather than wanting to admit their own mistake.

2.2.3 Question pertaining everyday concept of common knowledge

The question concerning common knowledge asks participants about the everyday notion of common knowledge.

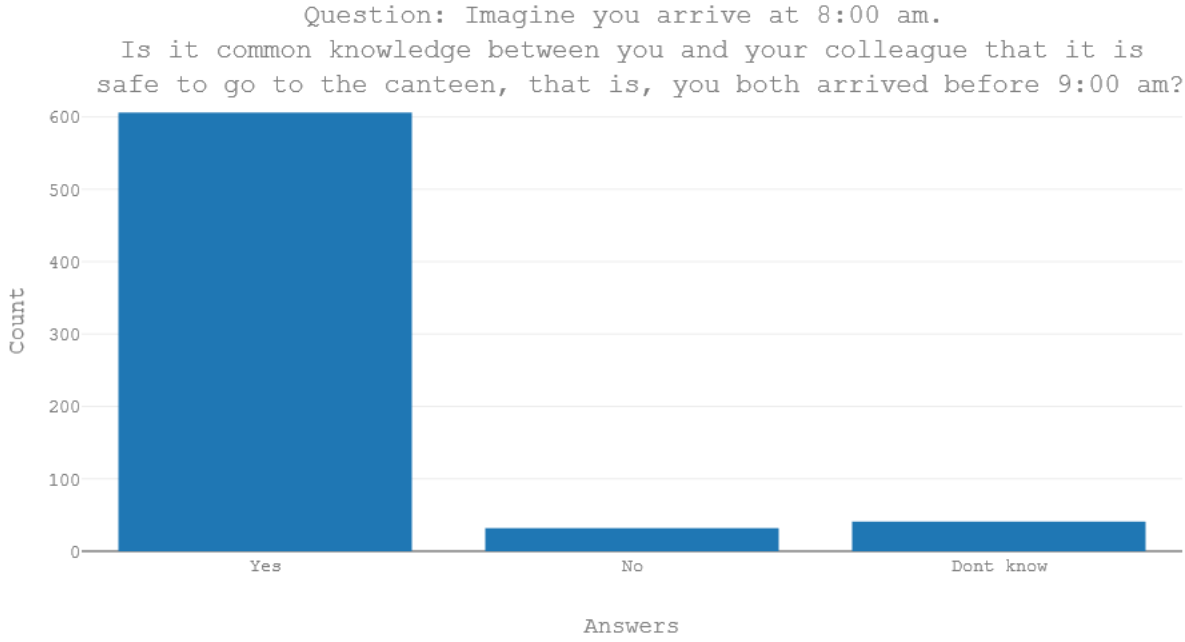


Figure 4. Bar-chart for the question concerning common knowledge.

This questions however likely ties into the everyday linguistic meaning of the term 'common knowledge', which is closer to what's called *mutual knowledge* in the research

literature, i.e. a fact known by everyone, which is different from the concept of common knowledge. In this sense, it would be right to answer that it is indeed mutual knowledge that both players arrived before 9:00 when one player arrive at 8:00. Unintuitively however, it is not common knowledge. The breaking point of the common sense meaning of 'common knowledge' could have been teased out of participants by asking them the question at randomly chosen arrival times. If participants truly understood common knowledge in the terms of mutual knowledge, most would answer 'yes' at 8:40 and earlier, and no after.

2.2.4 Cut-off question

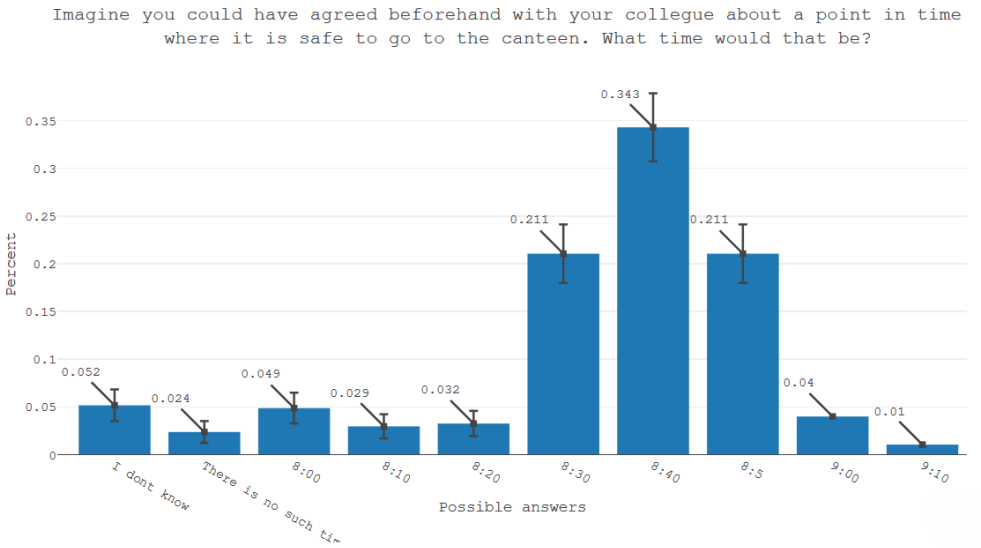


Figure 5. Bar-chart for the question concerning a safe time to go to the canteen.

This plot shows a significant cluster around 8:30, 8:40 and 8:50. There are two points for discussion. First, 21% deeming it safe to go to the canteen at 8:50 shows that people either might not consider the fact that the other player can arrive at 9:00, or that they believe that it's safe to go to the canteen at 8:50, even if the other player arrives at 9:00. This could either be due to (1) the belief that it's safe to go to the canteen as long as they individually arrived before 9:00, or (2) that they think it's safe to go to the canteen even when arriving at 9:00. The latter is implausible, since they could have answered '9:00' then. So, either they (a) simple don't consider the possibility that the other player would have arrived at 9:00 (this could be due to either (i) limit on observation powers meaning they made an unreasonable interpretation of the rules, (ii) limit on memory, i.e. they knew it at some point but forgot, or (iii) the rules were not stated explicitly enough,

meaning they simply acted according to a reasonable interpretation of the rules) or (b) they are aware of the possibility of the other arriving at 9:00, but simply believe their preference is choosing canteen at 8:50 even if the other arrives at 9:00. This could again be explained by (i), (ii) or (iii) stated above.

Secondly and more interesting, it might show a discrepancy between participants reflections over when it is safe to go to the canteen and their decisions in each round. This relates to a distinction between reflective and spontaneous use of Theory of Mind found in [Keysar et al. 2003]. Keysar et al. argue that while adults have the ability to use a sophisticated theory of mind both reflectively and deliberately, this ability is not incorporated thoroughly enough into the routine operation of their interpretation system to allow them to use it in spontaneous and non-reflective settings. They mention that the standard tests for assessing Theory of Mind in children are designed to tap into a meta-cognitive ability, i.e. the ability to evaluate and reflect, and not the ability to apply such reasoning in spontaneous decisions. The study from Keysar et al. focused on whether Theory of Mind was used in less reflective domains, when participants had to interpret the linguistic behavior of another person. The linguistic behavior consisted of a 'director' making to another participants such as 'Move the small candle to the right', which arguably triggers a less reflective decision-making procedure, compared to e.g. the quesiton 'What candle what I point out as the smallest candle?'. Their findings show that even adults who are perfectly capable of reflecting upon the difference between what they know and what another person know do not necessarily consult this knowledge when making spontaneous decisions, even in cases where it is crucial for successful cooperation with another person (see also [32] for evidence that adults application of ToM requires effortful attention)

This is highly relevant for any experiment concerning ToM. In relation to Figure 5, consider that 21% chose 8:30 as the safe time to go to the canteen. For those participants, 90% chose to go to the canteen at 8:30 (see Appendix E for plot showing a line-chart showing percent going to canteen, filtered by answers in Figure 4), which is unsurprising given their belief that it's safe to go to the canteen there. But for those same participants, 86% chose to go to the canteen at 8:40. This is possibly due to the decreased ability to consult an existing ToM when asked to make a spontaneous decision, while this knowledge is drawn upon to a larger extent when asked a reflective question. These two instances are not inconsistent however if we assume that participants choose to go to the canteen even when they don't think it's safe. This is possible due to the option of simply lowering their estimated certainty when they think it is less safe to go to the canteen of course, which we will look at in the next section. Focusing on Figure 5 however, our results show

that for those choosing 8:30 as the safe time to go to the canteen in Figure 5, 32% (±8%) choose to go to the canteen at 8:50 during the game. While this is lower than the average for our data, it is curious when compared to their answer in Figure 5. The plausible reasoning behind choosing 8:30 as the safe time to go to the canteen is due to backwards induction, i.e. 8:50 is not safe, and neither is 8:40 then. But even if being able to go through that reasoning, participants still chose canteen 32% of the time when they arrive at 8:50 and 86% of the time they arrived at 8:40.

Consider that participants interpreted the question in different ways. They might have interpreted the statement "Imagine you could have agreed beforehand with your colleague about a point in time where it is safe to go to the canteen", in the sense that answering '8:30' would not be inconsistent with also believing that 8:40 would be a safe time as well. While this is not logically inconsistent, there are two reasons for why participants have not made this interpretation. First, if some believed that they did not have to answer the latest possible safe time to go to the canteen, we would expect to see more answers at earlier times as well. But there are only 2-5% at times 8:00, 8:10 and 8:20. Secondly, given the pragmatics of the question, since the statement concerns agreement with the other player prior to the game, it would not make sense to agree on 8:30 being safe, if you thought you could just as well have agreed on 8:40.

In terms of ToM, interpreting the answer that it would be safe to go to the canteen at 8:30 as meaning it would not be safe at 8:40 amounts to first order-order Theory of Mind. This is because an answer that it is not safe to go to the canteen at 8:50 arguably depends on 0-order ToM, since choosing canteen at 8:50 involves the 50% chance the other player arrives at 9:00, and this is true regardless of the mental states of other players. Interestingly, the technically correct asnwer, 'There is no such time', is one of the least given answers at 2%. Since this is arguably below the background noise threshold in the data, it indicates that no participant made the full backwards induction.

Let us now turn to the certainty estimates in the data set. The discrepancy between the amount of participants choosing it would be safe to go to the canteen at 8:30 or 8:40 (i.e. not 8:50) and the amount of people choosing to go to the canteen at 8:50 is somewhat reflected in the certainty estimates provided by participants.

2.2.5 Differing certainty estimates per arrival time

First consider the absolute values contained in Figure 6 below.

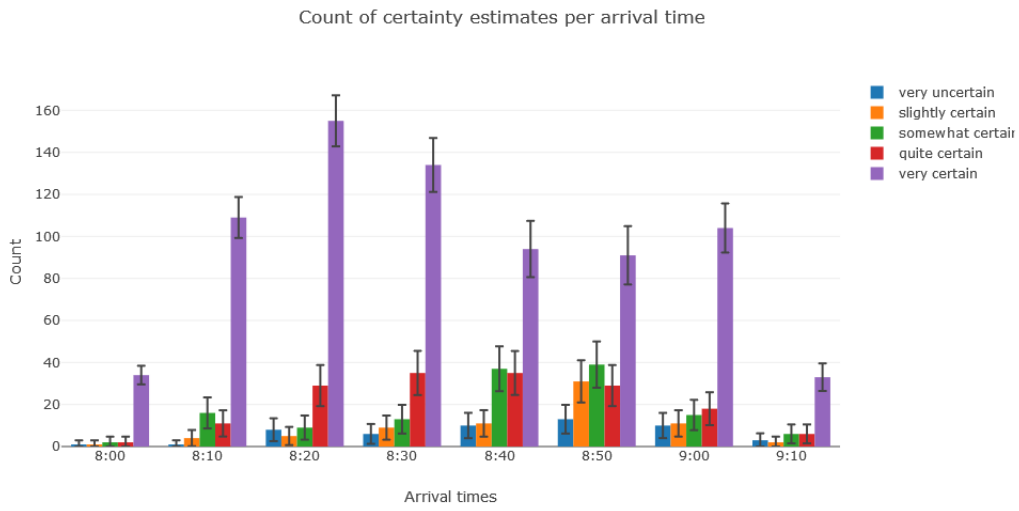


Figure 6. Grouped bar-chart showing counts of each certainty estimate for each arrival time with 95% conf. intervals.

Figure 6 shows a general picture of the amount of participants being 'very certain' in their choice is descending from 8:20 to 8:50. While participants quite generally (around 90%) went to the canteen when arriving before 8:50, there is trend of less participants being very certain about their choice as their arrival time approaches 9:00. Notice also the difference between amount being very certain at 8:50 and 9:00. The reason behind participants choosing 'very certain' less at 8:50 than at earlier times, might be due to them sometimes choosing canteen here, which they (should) know is penalized if the other arrives at 9:00, and them sometimes choosing office, which is they might believe is penalized if the other arrives at 8:40, due to them plausibly having the belief that they are penalized if the other arrives at 8:40 (since they themselves choose canteen there). But there are more participants who choose 'very certain' at 9:00 than at 8:50 or 8:40. Players quite generally went to the canteen at 8:30, went near randomly to the canteen at 8:50 and quite generally went to the office at 9:10, and suppose for the sake of argument that participants were somewhat aware of this (i.e. assume that most participants thought of others as similar to themselves). Then those choosing office at 9:00 are in a somewhat similar situation as those choosing canteen at 8:40. When choosing office at 9:00, participants arriving later will generally choose office as well, while it's near random for those arriving earlier. When choosing canteen at 8:40, something similar

applies, since they would believe those arriving at 8:30 to choose canteen most of the time, and those arriving at 8:500 to choose canteen nearly half the time. But while there are 53% (95% confidence intervals ±4) who are very certain when arriving at 8:40, there are 60% (similar confidence intervals) who are very certain of their choice when arriving at 9:00. However, the most telling result is likely the descending certainty across arrival times. To see this, consider Figure 7, a line-chart with both percent going to canteen per arrival time, and for those specific participants, the percent who were 'very certain'.

2.2.6 Percent being very certain about going to the canteen per arrival time

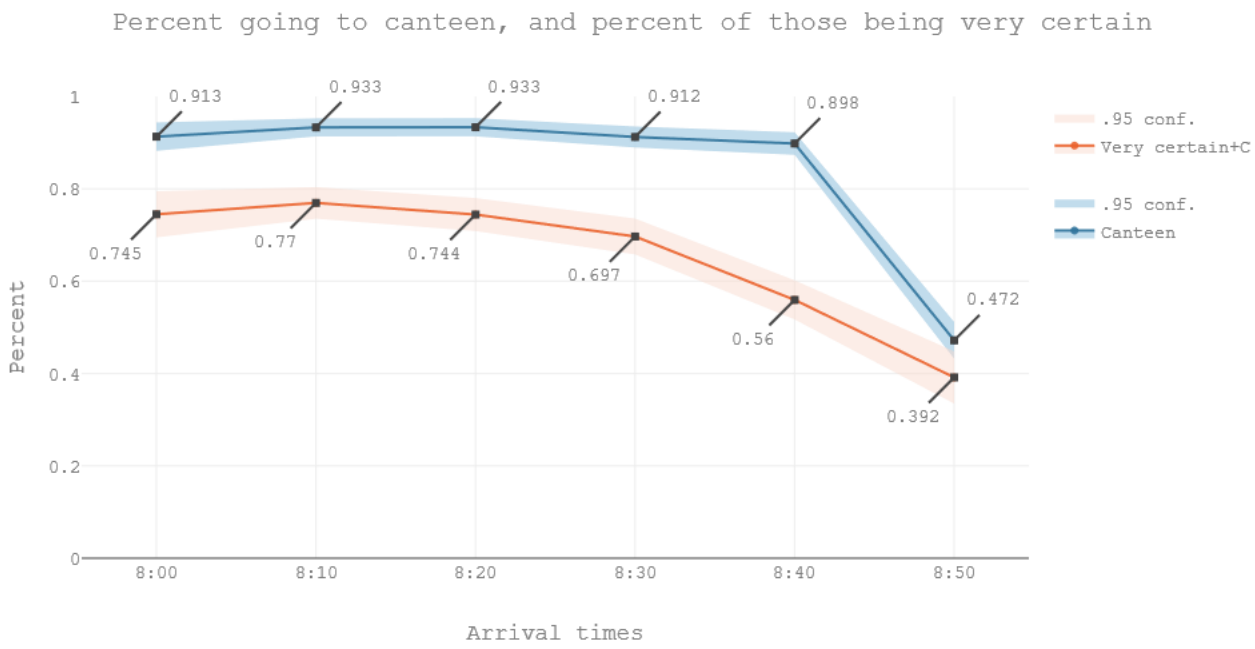


Figure 7. Line-chart showing percent who choose canteen, and percent of those who were 'very certain', excluding 9:00 and 9:10.

Figure 7 shows how many percent went to the canteen at different arrival times (blue line) and how many percent of those gave a 'very certain' estimate that the other player had done the same (orange line), both shaded error bars showing 0.95 confidence intervals. Note that times 9:00 and 9:10 were left out, since participants who went to the canteen at these times were few and less interesting, and their certainty estimates generally low with wide error margins due to low sample size (see Appendix F for full chart).

The blue line shows that people generally went to the canteen at the same rate (∼

91%) at times 8:00 to 8:40, before a significant drop to 47.2% at time 8:50. The orange line also shows a significant drop in certainty from 8:40 to 8:50, which is unsurprising, since participants likely went less to the canteen at 8:50 exactly because they were less certain about what the other player would do. The orange line has a different trend from the blue however, since there is also a significant drop from 8:30 to 8:40 (13.7 percent points), and it even drops 5 percentpoints from 8:10 to 8:20 and 8:20 to 8:30. In other words, participants seem to be less 'very certain' as their arrival time increase, while their willingness to go to the canteen does not change until 8:50. There is also an outlier at 8:00. There are slightly less participants choosing canteen at 8:00 and when they do, they're choose 'very certain' slightly less as well. A possible explanation for this is heuristic reasoning applied by participants due to the thematic narrative of the game. That is, given the propensity of office-jobs starting at 8:00, it is possible that some participants have been swayed by their real-world preferences when arriving at 8:00 to prefer the office over the canteen.

Percentwise, participants went to the canteen 47.2% of the time when arriving at 8:50 and chose 'very certain' 39.2% of the time they did so. It make sense that when participants don't chose the canteen that often (less than half the time), it is because they are uncertain if the other player arrived at 9:00, in which case that person would choose office, meaning they had to do the same. Given this belief, it makes sense that they often don't chose 'very certain' when choosing canteen at 8:50. Now consider 8:40, where 89.8% of the participants chose canteen, and 56% of those chose 'very certain'. Again, the reason for dominantly choosing canteen at 8:40 is likely due to the belief that the other person also chose canteen (which would result in a win regardless what time the other time the other had gotten). Therefore, it makes sense that the percent choosing 'very certan' is significantly higher than at 8:50. But now consider 8:30, where 91.2% chose canteen (a similar number to 8:40), likely due to the same reason as with 8:40, but where 69.7% of those chose 'very certain'. The surprising result here is that while participants were nearly equally predisposed to choose canteen at 8:40 as they were at 8:30, they are significantly less certain about it. The reason for being less certain at 8:40 than at 8:30 is clear in terms of ToM. Participants who arrive at 8:40, who themselves often choose office at 8:50, will not be very certain about their colleague making the same choice as them, given the possibility that the other player might have arrived at 8:50 and behaved like they would. But even if participants have this belief, they do not seem to act on it, since they still dominantly choose to go to the canteen at 8:40. To re-iterate, participants are less certain about going to the canteen at 8:50 than at 8:40 and they act accordingly, going to the canteen less at 8:50 than at 8:40. Participants are also less certain about going to the canteen at 8:40 than at 8:30, but they don't seem to act accordingly. While

they're significantly less certain about the other player choosing canteen when they arrive at 8:40 compared to 8:30, they choose canteen just the same.

2.3 Discussion

Let us focus on the previous sections 2.2.1 and 2.2.6 and more specifically Figure 1 and 7. Let us start analyzing Figure 1.

Figure 1 shows participants generally go to the canteen (˜91% of the time) before 8:50 and to the office after 8:50, while they go to the canteen a little less than half the times (47%) they arrive at 8:50. Going to the office at 9:00 or 9:10 does not indicate any ToM and choosing office at 8:50 does arguably not indicate any level of ToM either, since the rules state that "If you or your colleague arrive at 9:00 am or after, you should go straight to your offices. When a participant arrives at 8:50, they consider it possible that the other played arrived at 8:40 or 9:00. If the latter is the case, the rules state that they should go to the office, and this is true regardless of the other players decision and therefore modeling their mental state is irrelevant. This explains why slightly more than 50% go to the office at 8:50. The participants choosing the canteen at 8:50 likely assumes that if the other person arrives at 8:40, that person would go to the canteen, which makes 8:50 a random choice between office and canteen.

There is also a possibility that some thought that choosing canteen at 9:00 would be a winning move, as long as the other person arrived at 8:50 and also chose canteen. But even if some participants had an always-canteen strategy, it does not explain why those who chose office at 9:00 chose canteen at 8:50, or why some of those that chose office at 8:50 chose canteen at 8:40. If a participant believed the other person would choose always choose canteen, the only possibly winning move would be canteen as well. But if a participant would have chosen office at 9:00, it's unlikely that the participant would ascribe others an incorrect interpretation of the rules. Those choosing canteen at 8:50 might have read the first conjunct of the sentence: "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", thereby choosing to go to the canteen whenever they arrived before 9:00 am. But the amount of people choosing to go to the office at 8:50 is an indication of a correct interpretation of the last conjunct of the sentence, i.e. only go if your colleague goes to the canteen as well. In other words, the significantly larger amount of participants who went to the office at 8:50 than at 8:40 indicates that this subgroup G understood (i) they should not go to the canteen at 9:00, (ii) they should not go to the canteen if their colleague does not and (iii) their colleague's arrival time is within 10 minutes of their own. It's reasonable to claim that this leads to going to the office (at least sometimes) at

8:50, since they know that if they had arrived at 9:00, they would have chosen office, they assume the other person would have done the same which makes an office choice rational in at least half of the cases.

Let us focus on subgroup G above. Arriving at 8:50 they imagine that their colleague might have arrived at 9:00, in which case they know they should not choose canteen. If such a participant had a Theory of Mind, we might expect that they would have gone to the office at least sometimes at 8:40 as well, for the same reasons. If player a arrives at 8:40 and imagines the other player b arrive at 8:50, a might know that going to the canteen together would be best. But applying first-order ToM, a also knows that b does not know this. If a thinks b might behave like a herself, i.e. going to the office at 8:50, a would be in the same situation in 8:40 as in 8:50, which means a should go to the office at 8:40 to maintain a consistent strategy.

Now observe Figure 7, specifically the orange trace indicating the percent of those going to the canteen who were also 'very certain' that their colleague had made the same decision. To re-iterate the discussion in section 2.2.6, while we see a near flat development in the percent going to the canteen up to 8:50, we see a drop in the estimated certainty at arrival times before that. There is a drop of 13.7 percent points of 'very certain' estimates from 8:40 to 8:50, while there is only a drop of 1.4 percent point for going to the canteen from 8:40 to 8:50. This indicates that participants who choose canteen at 8:40 are aware that the other person might arrive at 8:50 and choose office, but this knowledge does not change the decisions for the participants. In other words, participants do not consult their theory of mind when asked where they would go (they go to the canteen at 8:40 since they know both players arrived before 9:00), but do consult their theory of mind when asked how certain they are that their participant made the same choice. It might be due to having more time to think about the particular situation, or because the question is framed in terms of what the other player did, which prompts players do model other players mental states.

There is a caveat here. Just because someone is less certain about going to the canteen at 8:40 than at 8:40 does not mean it would be rational for them to choose office, since they can still be more certain that canteen is the right answer than that office is. In fact, a canteen choice at 8:40 might be rational for the following reason. Imagine participants who choose office at 8:50 does it at random, i.e. they know they might have chosen office in another situation (ignore the fact that this would entail giving a 'very uncertain' estimate). When those participants arrive at 8:40, they might believe that if their colleague arrives 8:30, they would choose canteen, and if they arrive at 8:50, they would go to the office about half of the times. That would mean there would be a 75%

chance of the other player going to the canteen, which would warrant a canteen choice but with lowered certainty.

But let us look closer at the data. Let us expand subgroup G so it only includes those who chose office at 8:50, 9:00 and 9:10. We might then plausibly assume that these players understood the rules stated above, (i) don't go to the office at 9:00 or later (ii) only go to the canteen if your colleague does the same and (iii) you always arrive within 10 minutes of your colleague. Figure 8 is similar to Figure 7, besides excluding players who chose canteen at 8:50, 9:00 or 9:10. This filter includes 327 participants.

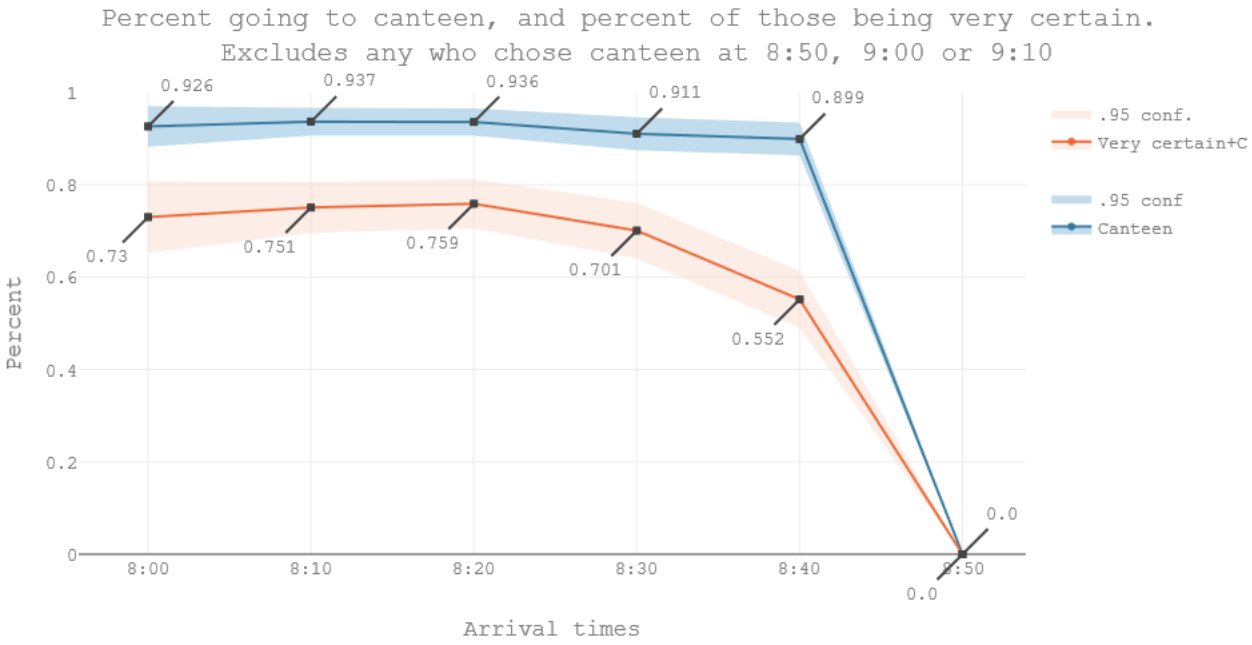


Figure 8. Percent going to the canteen (blue) and percent of those 'very certain' about their choice. Excluding any participants choosing canteen at 8:50 and later. Shaded error bars for 0.95 conf. intervals.

Figure 8. shows that even participants (N=327) who always go to the office at 8:50 repeat a similar pattern as found in Figure 7. That is, such participants still quite generally go to canteen from 8:00 to 8:40, while the percent being 'very certain' drops significantly. The percent going to the canteen drops 2.4% from 8:00 to 8:40, while the amount of participants choosing canteen and giving a 'very certain' estimate that the other player did the same drops 18.7% from 8:00 to 8:40. It's possible that these participants thought that while they themselves would always go to the office at 8:50, they do not trust that others would, thereby lowering their certainty estimate. It's also

possible that players have based their decision on going to the canteen whenever they knew they privately knew it was safe (ignoring ToM considerations), but when asked how certain they were about the other player doing the same, they are explicitly asked about what the other player would do, not what they themselves would do, which prompts ToM-related reasoning to a larger extend than simple decision-related questions. While we looked at subset G to see if the trend persisted when focusing only those choosing office at 8:50 and later, the same trend can be seen in the total data set. To summarize: Participants in the Canteen Dilemma did seemingly not exhibit ToM to large extend when given questions in the context of practical decisions, but when asked about their certainty about other participants decisions, they consulted their ToM to a larger extend. Let us now look at a few related results from the literature on Theory of Mind (and the lack thereof).

2.3.1 Limited use of ToM even in adults: reflexive versus spontaneous use

The possibly most applicable study is "Limits on theory of mind use in adults" by Keysar et al. (2003). Their results show a "clear dissociation between an ability that is firmly in place by adulthood, and the reliable use of this ability for the very purpose for which it is designed." [Keysar et al. 2003, 39]. It is specifically the ability to represent others' beliefs that is not reliably used to interpret others' behavior. Keysar et al. do not deny that the belief system of adults includes a ToM and that they are able to gain knowledge and reason about the mental states of others, but rather that even adults with this capacity do not necessarily draw on it, nor do they necessarily consult crucial knowledge about the mental states of others, even when it is required for successful interaction.

They argue that traditional tests for the development of ToM in children rests on reflective tasks, often meant to tap meta-cognitive abilities such as being able to evaluate and reflect on the knowledge of oneself and others. Keysar et al. results imply that there is a disassociation between reflective and spontaneous use of ToM. Their test focuses on adult's interpretation of another's linguistic behavior in order to test whether this reflective ability is used spontaneously. This relates to the Canteen Dilemma, since participants are told their arrival time before being asked 'where do you go?', which is in itself a practical question about a decision, rather than a reflective question about the mental states of others. The time limit on each round also plausibly adds to the spontaneity of the situation. They offer possible explanations for their findings:

> "Why might adults sometimes fail to deploy their fully-developed theory of mind? Perhaps in the "real world" perspectives tend to coincide such that what is present and salient to one person will tend to be salient to another.

> Under these circumstances, directly computing what another person knows or does not know at a given moment might be more trouble than it is worth. Furthermore, even when perspectives do not coincide, feedback from one's partner can obviate the need to compute that person's perspective for successful coordination. For example, although participants in our experiments often moved hidden objects, the director quickly corrected them and they eventually moved the correct object. Although participants could have pre-computed the director's perspective, they got away with being egocentric because they could count on the director's feedback. In short, the dynamic nature of face-to-face interaction gives people latitude to be egocentric by effectively distributing the burden of perspective taking across interlocutors." [Keysar et al. 2003, p. 39]

Their explanation points to the economics of applying ToM. If applying ToM is computationally resourceful, it might be more efficient to generally ignore one's ToM and accept the cost when errors occur, and then start incorporating ToM in that (and similar) situations. This is of course problematic in one-shot cases with high stakes.

2.3.2 Reflexively mindblind: modeling others as our selves

In "Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention" (2010) Lin et al. demonstrate that "people's ability to use their theory of mind depended on their capacity to expend effortful attentional resources, in particular on their working memory." [32, p. 555]. While adults might develop the capacity to reason about the mental states of others, Lin et al. write that "possessing a capacity and actually using it to interpret social action, however, are two different things" (ibid.)

Lin et al. gives a few explanations as to why do not always include their ToM in their reasoning, e.g. "People need not always employ their theory of mind to under stand others' behavior. In general, the more information a person shares with the actor, the less they need to consult their under standing of the actor's beliefs and instead can egocentrically use their own private mental states. In cases where information is perfectly shared, such egocentric" [32, p. 555]. The same explanation could be used for the Canteen Dilemma. In everyday scenario's it's likely that people even if people do not arrive at the same time, they have common knowledge that they both arrived close to a common arrival time.

The same study also state that "possessing a theory of mind is crucial for effective social functioning, but this does not mean that people automatically employ this capacity. Although it appears that people interpret others' mental states relatively effortlessly, what people do reflexively is use their own mental states. Using one's knowledge of another per

son's beliefs and mental states requires effortful attention." [32, p. 556]. I take 'reflexively modeling their own mental state' to mean that they imagine other's mental states as their own. This could either mean a mental model where you only ascribe beliefs, intentions and so on to the other person which would be justified from their perspective, or it might mean using your own mental state as a model for the other, including at least some of your own beliefs, intentions and so on. In that case, a participant in the Canteen Dilemma arriving at 8:40, would imagine the other player possibly arriving at 8:50, and thinking about what they themselves would do at 8:50 *assuming they know the other player arrived at 8:40*. This explains a high prevalence of canteen choices at 8:40. We'll focus on this as well in the next section.

2.3.3 Curse of knowledges - attributing our own mental state to others

The "The Curse of Knowledge in Reasoning About False Beliefs" by Birch and Bloom (2007) show something similar to Keysar et al. (2003), namely that adults fail false-belief tests much like younger children do if sensitive enough meassures are used. More specifically, they show that "adults' own knowledge of an event's outcome can compromise their ability to reason about another person's beliefs about that event. We also found that adults' perception of the plausibility of an event mediates the extent of this bias" [Birch & Bloom, p. 382].

Their focus was on what they called a *curse-of-knowledge* bias. In other words, when adults have to reason about the mental states of others, they are more likely to ascribe knowledge to the other person if they possess the knowledge themselves. Their second result show that the extend to which adults do this depend on how justified they believe their knowledge is.

To see how this relates to the Canteen Dilemma, consider participants who arrived at e.g. 8:40. Such a participant know that both players going to the canteen results in the biggest reward. When, or if, they start making a mental model of the other player and consider their reasoning if they arrived at 8:50, the 'curse of knowledge' entails that some are bound to ascribe their own knowledge to the other player. That is, because the player at 8:40 knows it's safe to go to the canteen, they possibly ascribe this knowledge to the other player as well.

The Canteen Dilemma shows this since there is a drastic drop from 8:50 to 8:40 in participants choosing office. Results from the Canteen Dilemma possibly shows that even when participants do not consult their ToM sufficiently to affect their choices, their certainty about their choices is still affected by their ToM.

2.3.4 Applying ToM requires knowledge of its benefits

The studies above show that higher order social reasoning presents a serious cognitive challenge for adults. Note for example the remark in "Learning to Apply Theory of Mind" by Verbrugge & Mol (2008): "Our results provide a warning against assuming that because some people (such as trained logicians) on some occasions apply higher-order Theory of Mind, such reasoning is at all widespread". That is, some research show that adults rely notable less on higher order ToM than researchers might intuitively think.

Verbrugge (2009) describes possible explanations for this apparent difficulty. The first is that there is a high processing cost associated with ToM, which causes failure to perform social reasoning on a sufficiently high level when the processing demands are high. The other Verbrugge describes is that higher order social reasoning does not necessarily transfer from one domain of application to the other. But the possibly most important reason as to why people fail to apply sufficient ToM, is that in order for a person incorporate higher order ToM in their decisions, it is not just necessary that they have the capacity for it, they must also be aware of the advantage of incorporating such knowledge into their decisions and actions [46, p. 661]. In other words, in order to use ToM, adults must be become aware of its benefits, but it might be difficult to notice its benefits without depending on ToM in the first place. We might therefore expect certain individuals to be proficient in using their ToM, not (only) if they have an increased capacity for it, but also if they more acutely pick up cues as to when it is required (e.g. recognizing patterns of interactions without common knowledge, where people are bound to have different beliefs about the same event).

Verbrugge's other point is more complex, i.e. that ToM might not be a uniform mental ability, but rather constitutes a broader family of mental abilities. It possibly depends on the extend to which ToM is task dependent, which is not well researched. However, the Canteen Dilemma is consistent with the hypothesis that ToM is task dependent in the sense that the everyday commonality of arriving at work and choosing between places like canteen and office does not trigger participants ToM considerations.

Griefing strategy comment

Lastly, consider the possible griefing strategy in the game, where a player does not think they can make it through 10 rounds with all their bonus intact (which is highly likely) and therefore intentionally makes the other player lose their bonus. If a player believes she knows what the other player would do, the player might still intentionally choose the opposite while lowering their certainty estimates. This causes the other player to lose a large amount, assuming they were very certain, while they themselves lose minimally.

After a few rounds, the other player loses their bonus and the game ends while the griefing player maintains a bonus. Consider the actual player*s* *a* and *b*. Griefing player *a* cooperated at 8:10 in round 1, giving a 'very certain' estimate, losing $0.01. In round 2 and 3, player *a* chose office at 8:10 and 8:20 respectively, resulting in miscoordination, while giving a 'slightly certain' estimate. This cost *a* $1.96 in round 2 and 3, while player *b* lost her entire bonus, allowing *a* to make it out with a bonus of $6.03 while giving the strategy response '*brain power*', while *b* gave the strategy response 'I only got to play three rounds because my partner clearly didn't understand the rules, and cost us our entire bonus.' However, those choosing office with low certainty might also plausibly employ ToM, knowing that while the other will likely choose canteen, it will not be a safe strategy throughout the game. Given its complexity it's also implausible that this was a prevalent strategy in the game.

3 General discussion

4 Conclusion

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.

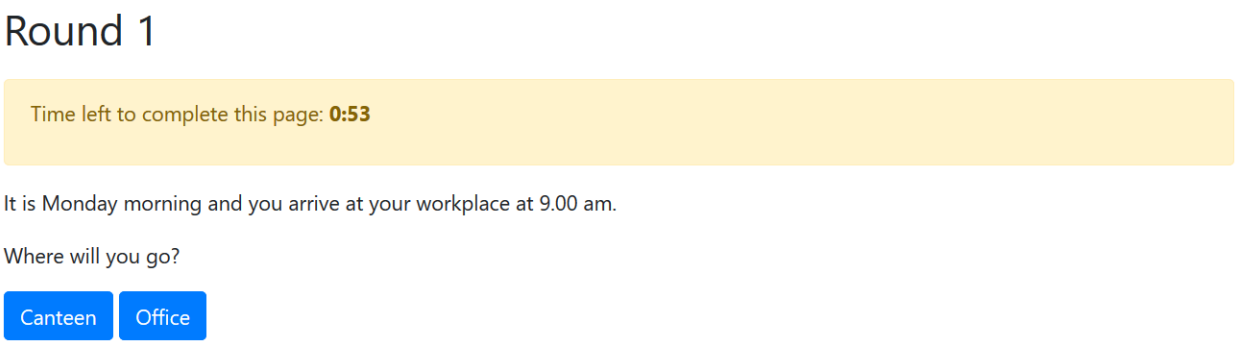
Payoff and penalties:

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.

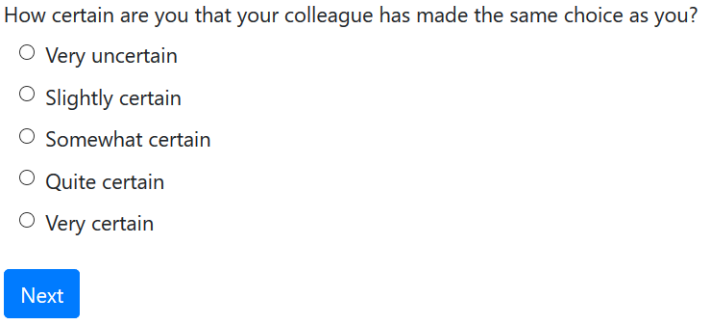
- **Both go to canteen**
  If you guessed correctly that both of you went to the canteen before 9:00 am, you pay a **small** penalty proportional to how **uncertain** you were, e.g.:
  - **-$0.69** if you were very uncertain.
  - **-$0.29** if you were somewhat certain.
  - **-$0.01** if you were very 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.



Appendix C



Appendix D

Results (after round 5)

Time left to complete this page: **0:15**

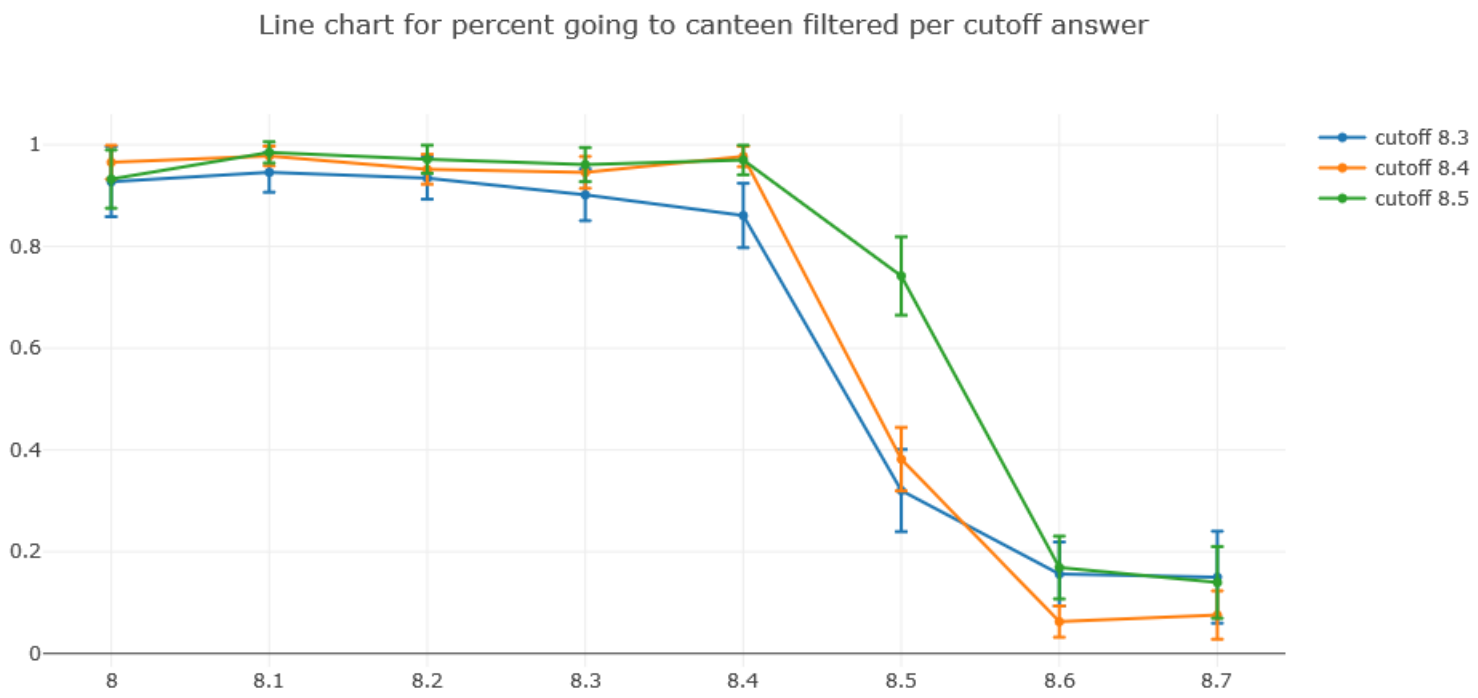
Round	You went to the	at	Your colleague went to the	at	Your certainty	**Penalty**
1	canteen	8.30	canteen	8.40	quite certain	**-$0.13**
2	office	9.00	office	8.50	quite certain	**-$0.27**
3	canteen	8.40	canteen	8.30	quite certain	**-$0.13**
4	office	9.00	canteen	8.50	quite certain	**-$4.16**
5	canteen	8.40	canteen	8.30	very certain	**-$0.01**

You both chose to go to the canteen. Good work!

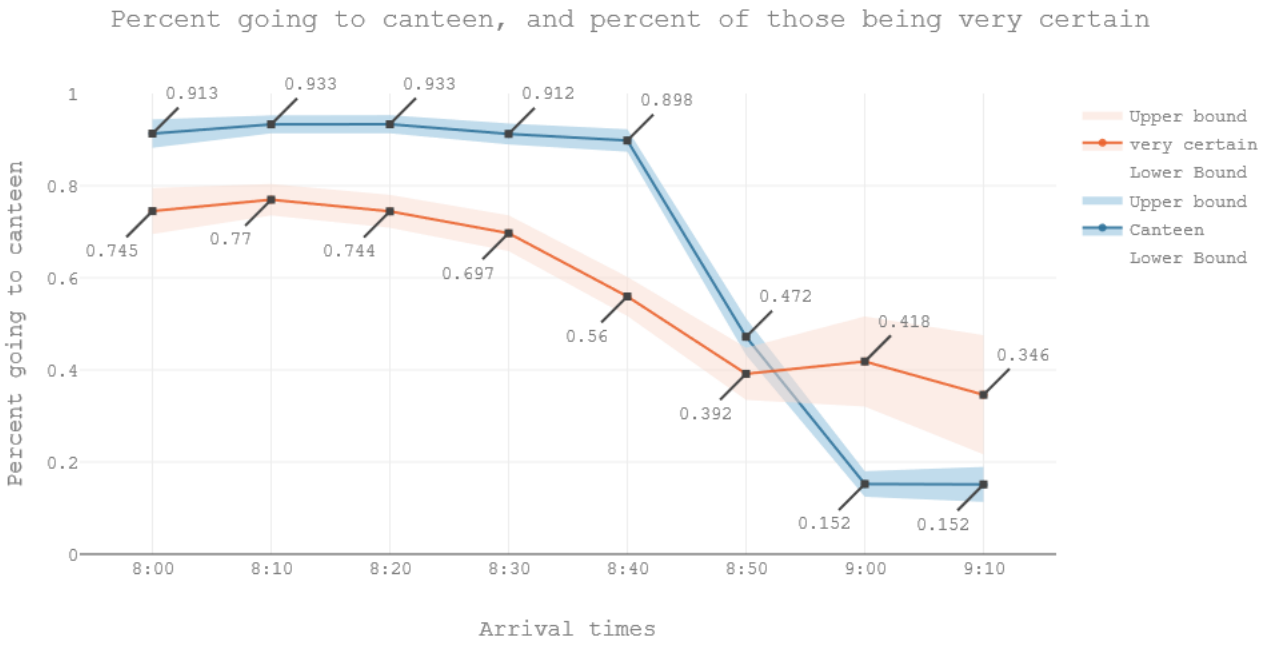
You lost **-$0.01** this round. Your total losses are **-$4.70**. You still have **$5.30**.

Next

Appendix E


Line chart for percent going to canteen filtered per cutoff answer
cutoff 8.3
cutoff 8.4
cutoff 8.5
1
0.8
0.6
0.4
0.2
0
8
8.1
8.2
8.3
8.4
8.5
8.6
8.7

Appendix F


Percent going to canteen, and percent of those being very certain
Percent going to canteen
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
0.913
0.933
0.933
0.912
0.898
0.745
0.77
0.744
0.697
0.56
0.472
0.392
0.418
0.346
0.152
0.152
Upper bound
very certain
Lower Bound
Upper bound
Canteen
Lower Bound

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

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

[5] 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.*

[6] 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.).*

[7] 141. 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)

[8] 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.)*

[9] 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

[Birch & Bloom] 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

[10] 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.

[11] 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

[12] 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.

[13] 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

[14] 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

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

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

[17] 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.*

[18] 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.*

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

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

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

[22] 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.)*

[23] 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.*

[24] 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

[25] 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.

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

[27] 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.*

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

[29] 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

[30] 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.*

[Keysar et al. 2003] 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.

[31] 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).

[32] 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.

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

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

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

[36] 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.*

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

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

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

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

[41] 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.

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

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

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

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

[Verbrugge 2008] 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.)

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

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

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

[49] http://www.glascherlab.org/social-decisionmaking/