Introduction to the canteen dilemma and higher-order social reasoning

Thomas Schrum Nicolet

May 30, 2019

Abstract

It is often necessary to attribute appropriate mental states to others in order to be able to interpret and predict their behavior. The cognitive capacity to do so have often been referred to as 'theory of mind'. Although interpreting the behavior of others in terms of mental traits seems to occur naturally, the reasoning involved is notoriously difficult. The present thesis involves an experiment called the canteen dilemma aimed at investigating the limits of this capacity. The experiment was devised and implemented in conjunction with R. Engelhardt (CIBS) and T. Bolander (DTU Computer). The present paper is an introduction to an accompanying article on the experiment. It presents empirical findings from cognitive science which are relevant for understanding the canteen dilemma as well as motivate its importance and relation to logic and other fields. I motivate interdisciplinary work between logic and cognitive science by arguing that empirical insights are necessary in order for logic to make psychologically realistic models of agents, which benefits both cognitive science by providing insights into the computational complexity involved in human reasoning as well as allowing artificial intelligent technology to behave more human-like.

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.

Contents

1 Introduction

Making sense of observed behavior of others in terms of unobservable mental traits is cognitive capacity essential for successful social interaction. Inferring what others might think, believe or intend helpsus navigate our complex social world. This sociocognitive capacity allows us to not just predict and understand the behavior of others based on their mental states, but also predict and understand how our behavior might affect the mental states of others. It is this capacity which allow self-aware humans to acknowledge and appreciate that others have mental states just like themselves. It has understandably been described as the pinnacle of social cognition [25]. The capacity was originally referred to as 'theory of mind' in the seminal paper by Premack & Woodruff [83] in 1978. It has seen implicit philosophical attention for centuries and been of immense scientific interest in the last 40 years within studies such as psychology, biology, neuroscience and philosophy. This thesis consists of two parts: An article on an experiment, called the canteen dilemma, which aims at investigating the nature and limited use of theory of mind, as well as the present paper which works as an introduction to the article. The canteen dilemma experiment was conducted at the Center for Information and Bubble studies and was designed and implemented in conjunction with R. Engelhardt and T. Bolander. The experiment consists of a two-player strategic coordination game with imperfect information with a structure such that one's social reasoning affects one's behavior in the game. The present introduction is intended to supplement the accompanying article by presenting and discussing relevant research on social cognition, how and why it is relevant to research in logic, as well as providing supplementary results and discussion on the canteen dilemma experiment itself.

It is necessary to preface some of the discussion on social cognition with some reservations. The field has been criticized in terms of "theory of mind" being used vaguely and inconsistently between studies and by reference to studies (i) treating theory of mind as a monolithic process, (ii) referring to a single brain network for theory of mind and (iii) conflating varieties of theory of mind [90]. It has also been argued that certain studies which point to cases of theory of mind could be explained without reference to mental states [57]. This calls for a critical assessment of the scientific understanding of theory of mind which will be discussed later in section 3.1. Attributing mental states to others have also been referred to as 'mentalizing' [57], 'mind-reading' [58], 'perspective taking' [87] and 'higher-order social cognition/reasoning' [4, 99]. Given that the canteen dilemma's emphasis on reasoning about possible mental states of others in order to inform one's decisions, I will often use the term higher-order social reasoning or cognition, including the term 'social' to emphasize interpersonal aspect involved.

Despite methodological difficulties, the ability to attribute mental states to others is undeniably both real and important. This ability allows us to develop a more complete and accurate understanding of those around us and as such is an essential part of understanding what it means to be

human. The reasoning involved can be applied recursively, modeling the mental state of others, including their model of one's own mental state, and so on. The recursive modeling of the mental states of others is referred to as higher-order social reasoning or cognition and it is this cognitive activity that is the general focus of the canteen dilemma experiment and this thesis.

Epistemic logic, the modal logic of knowledge, is a logical language capable of representing the knowledge of agents in complex situations and as such is a natural tool for describing higher-order social reasoning. In general, zero-order reasoning concerns facts about the world, while $n+1$-order reason concerns facts about n-order reasoning of other people [99]. Since higher order social reasoning can be described in various epistemic logics, the interface between logic and cognitive science has recently received interest. There is a two-fold question in this interdisciplinary field: To what extent do people reason differently from logical prescriptions and how do we interpret possible divergence in the two areas of research? The first question will be discussed in section 9 while the latter question will be discussed in the section below. After the following section I present the epistemic logic $S5$ before I turn to relevant empirical insights from cognitive science, focusing on the possible limits of human reasoning. This leads to a description of the canteen dilemma. As the essential details and results of the experiment are fleshed out in the accompanying article, this introduction will keep to discussion of supplementary results and analysis of the experiment.

1.1 The cognitive turn in logic

Logic bears a long but intricate relation to human reasoning. The logic of argumentation has been studied since antiquity and is still frequently taught to first year philosophy students. The goal of this is presumably not just to teach students about valid inferences but also to improve the validity of inferences they make themselves. Modern logic has besides this traditionally been seen as unrelated to empirical facts about human reasoning except as a normative force of how we ought to make inferences. Modern logic has been undergoing a cognitive turn in recent years which has emphasized its relation to human reasoning, notably championed by logicians such as Johan van Benthem [11] and Rineke Verbrugge [99].

This cognitive turn is a step away from the anti-psychologism exhibited in the early days of modern logic, as Frege stated that "logic is concerned . . . not with the question of how people think, but with the question of how they must think if they are not to miss the truth" [41, p. 250]. Frege refers to the normativity of logic but his point rests on a more fundamental claim that logic is the study of objective mind-independent logical laws, which he states explicitly: "the laws of truth are not psychological laws: they are boundary stones set in an eternal foundation" [40, p. 13]. So logic is normative according to Frege because the norms for reasoning and beliefs are structured by what is truth-conducing and the laws of logic are objective truths, in fact necessary truths. Andersen [1] argues that Husserl followed Frege in this anti-psychologism as Husserl even argued that any

argument that makes logic independent from psychology by reference to the distinction between how we ought to think and how do do think is implicitly psychologistic, since any such argument still refers to human reasoning (see [63, §17-20]). So for Frege and Husserl, the normativity of logic is a side-effect of its objectivity.

The objectivity of logic has been part of a fundamental philosophical debate about what makes logical truths true, what logical facts are about and whether they are necessarily or contingently true. There is a rich historical dialogue on this from Frege, Husserl, Wittgenstein, Ayer, Carnap, Quine, Kripke and Putnam among others, although its discussion is only indirectly related to how the purity of logic relates to the quirks of human reasoning. Assume we reject Frege's view and adopt a non-Platonistic Realism like Maddy [75] or even more radical empirical version from Quine [86], which states that the truth of logical laws are contingent on the right physical structuring. This does not affect the normativity of logic because even on such a view, human reasoning takes place within the physical structuring in question and is governed by the laws of logic as well. How is human reasoning relevant for logic then?

Take argumentative logic. There are certain inferences that humans make that can only be described as mistakes. Take Modus Ponens for example, from P and $P \rightarrow Q$, we can conclude Q, which can easily be relevant in situations where failing to infer Q can have adverse consequences. But there are more complex cases which can be modeled by more complex logics and where deviance from the logical model might not be a mistake in reasoning. There are few reasons for this which both encourage collaboration between logic and cognitive science.

First, if someone makes an inference which is invalid in classical logic, it does not follow that it is logically invalid in non-classical logics. There might be good reason for human reasoning does not always follow prescriptions from classical logic, such as when dealing with incomplete information or limited cognitive resources. See Drew & Doyle [32] and Brewka [18] for work on non-monotonic logic and default reasoning. Non-monotonic logic allows valid inferences on incomplete information for example such that new information might invalidate earlier inferences, which is not possible in classical logic. So, it does not follow that reasoning which deviates from some logical prescriptions is random or illogical. Such reasoning might simply not be properly formalized or represented. People might reason differently for various reasons and if logic wants to maintain its normativity, it should take these different reasoning patterns seriously.

Second, the problem of developing artificial intelligence was first described as that of "making a machine behave in ways that would be called intelligent if a human were so behaving" by John McCarthy in 1956 [73, p. 11]. If we want machines to be able to reason similarly to humans, it is necessary to be know the logical structure of how humans actually reason. This is especially important when it comes to higher-order social cognition. If we want to be able to interact intelligently with artificial agents, such agents will have to be able to reason about and understand the reasoning of

humans. In other words, if human reasoning deviates systematically from idealized logical prescriptions, AI systems should taking this into account when having to predict the behavior of humans.

I will now turn to epistemic logic as an example of an idealized representation of reasoning about knowledge. Epistemic logic was originally used to formalize and analyze epistemological notions and concepts and is also useful as a tool for showing the epistemic depth of certain real world situations. Some relevant idealizations will be discussed before moving onto real higher-order social reasoning.

2 Epistemic logic

Epistemic logic is a useful modeling tool for representing complex epistemic components of real world cases. An epistemic language is defined below with a subsequent focus on the system $S5$ as it is the most of often used to symbolize knowledge.

Definiton 1.1 (Epistemic language) Let P be a set of atomic propositions and A a set of agent-symbols. The language $\mathcal{L}_{\mathcal{K}}$ for multi-agent epistemic logic is then generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi$$

Disjunction, conditional and bi-conditional can be used standardly as abbreviations of formulas using the negation "$\neg$", conjunction "$\wedge$". Kripke models and truth conditions are also defined for our epistemic language, before going into the art of modeling in epistemic logic,

Definition 1.2 (Kripke models) A *Kripke model* for the epistemic language is a structure $M = (S, R, V)$, where S is a set of states, R is an accessibility relation $R_a \subseteq S \times S$ for every $a \in A$, where R_ast means t is accessible from s for agent a. V is a valuation function assigning truth values to propositions at states.

Definition 1.3 (Truth conditions) Epistemic formulas are interpreted on *pointed models* M, s consisting of a Kripke model $M$and a state $s \in S$. Truth conditions for formulas are then:

$M, s \vDash p$ iff V makes p true at s

$M, s \vDash \neg\varphi$ iff *not* $M, s \vDash \varphi$

$M, s \vDash \varphi \wedge \psi$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$

$M, s \vDash K_a\varphi$ iff *for all worlds t such that* R_ast, $M, t \vDash \varphi$

I will focus on the prominent epistemic logic $S5$, which is the set of Kripke models where R is an equivalence relation. I will also use *states* and *worlds* interchangeably. We get $S5$ by adding the following axiom schema to our logic (which can both be understood in terms of their epistemic consequences and the implication for the accessibility relation R):

$K_a\varphi \rightarrow \varphi$ (truth or reflexivity),

$K_a\varphi \rightarrow K_aK_a\varphi$ (positive introspection or transitivity)

$\neg K_a\varphi \rightarrow K_a\neg K_a\varphi$ (negative introspection or euclidity)

The first axiom entails veridicality and is mostly uncontroversial: if you know something, it is true. The other two implies that agents are aware of what they know and do not know, which are both often referred to as psychologically unrealistic idealizations. I later discuss other and more important idealizations in epistemic logic. $S5$ allows R to be interpreted as an *indistinguishability* relation. This entails that agents do not which of the worlds they consider possible is the actual one. It intuitively encapsulates the semantics above, which state that agents only know something when it is the case in all worlds they consider possible. We can also add a modal operator for *mutual knowledge,* symbolizing that every agent in group B knows φ, namely $E_B\varphi$, defined as the conjunction of all individuals in B knowing φ. That is, for every $B \subseteq A$:

$$E_B\varphi = \bigwedge_{b\in B} K_b\varphi$$

As the limiting case of the infinite conjunction mutual knowledge, where everyone knows φ, and everyone knows that everyone knows φ *ad infinitum*, we introduce the notion of *common knowledge*:

$$C_B\varphi = \bigwedge_{n=0}^{\infty} E_B^n\varphi$$

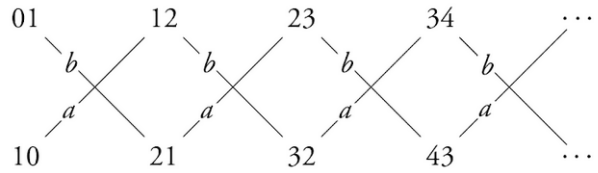
Common knowledge is interpreted semantically such that φ is common knowledge among a group A in a state s if and only if, for all worlds t in the reflexive transitive closure of the accessibility relations for all agents in A, φ holds in t. There is an intriguing but unintuitive difference between any finite number of iterations of the E-operator and the infinite iteration expressed in common knowledge. The difference can be made explicit in examples from epistemic logic like the 'consecutive numbers' example below. This example is structured very similarly to the canteen dilemma as we will see later.

2.1 Consecutive numbers example

Two agents a (Anne) and b (Bill) sit together. They are told they will each receive a natural number and that their numbers will be consecutive such that the numbers are n and $n+1$ where $n \in \mathbb{N}$. They are only told told their own number but it is common knowledge between Anne and Bill that their numbers are consecutive. This entails a structure where in any given situation, each agent know their own number and considers it possible that the other agent has a number one before or after (unless they have 0)[1].

[1]The following example is based on the consecutive numbers example in [30] and [29].

Figure 1: Consecutive numbers model (Ditmarsch & Kooi [30]).



Suppose the real situation is Anne being told 2 and Bill 3, denoted as a state $(2, 3)$. There are a few basic facts which are true in $(2, 3)$. Most basically, Anne knows her number is 2 and Bill knows his number is 3: $K_a a_2 \wedge K_b b_3$. Since they know their numbers are consecutive, they both know that there are only two possible numbers for the other (but these are not the same for Anne and Bill!): $K_a(b_1 \vee b_3)$ and $K_b(a_2 \vee a_4)$. Since Anne considers it possible that Bill has as 1 or 3, she considers it possible that Bill considers 0 or 2 possible for Anne (if Bill has 1) or 2 or 4 (if Bill has 3), stated as: $K_a(K_b(a_0 \vee a_2) \vee (K_b(a_2 \vee a_2))$. Now imagine that Anne and Bill has to guess whether they both have positive numbers, that is, none of them have a 0, denoted as $(positive)$.

If we ask Anne and Bill in state $(2, 3)$ if they know that the other's number is positive, they would both answer yes: $K_a(positive) \wedge K_b(positive)$ or expressed as $E_{\{a,b\}}(positive)$. This is equivalent to checking for each agent i whether every world which is i-accessible from $(2, 3)$ only has positive numbers. However, now imagine that we ask both Anne and Bill if they both know that $(positive)$ is true. Bill considers $(2, 3)$ and $(4, 3)$ possible. Since we are only focused on 0, we can focus on the lowest number-pair $(2, 3)$, since the other direction only takes us away from 0. Bill knows Anne's lowest possible number is 2 , which means he knows both numbers are positive. Since if Anne has a 2, she would consider 1 as the lowest number, Bill knows that Anne knows both numbers are positive as well: $K_b K_a(positive)$. But even though Anne in fact has a 2, meaning she knows that both numbers are positive since the lowest possible number for Bill would be 1, if Bill has a 1, he considers it possible that Anne as a 0. So $\neg K_a K_b(positive)$. In other words, while everyone knows that no one has a 0, not everyone knows this epistemic fact itself: $E_{\{a,b\}}(positive) \wedge \neg E_{\{a,b\}} E_{\{a,b\}}(positive)$.

The unintuitive aspect of this is the satisfaction of n iterations of the E operator but not $n + 1$. So in the consecutive number example, no matter what number pair Anne and Bill are given, it is never common knowledge that none of them have a 0! This also means that telling Anne and Bill that both of their numbers are positive is actually an informative statement in any possible situation of the consecutive numbers example.

The practical implications of this can be seen by 'gamifying' the example. Imagine Anne and Bill have to answer independently of each other whether they are certain that both have positive numbers or remain agnostic. Their objective is to (1) always give the same answer and (2) not

answering they have positive numbers if they do not. In our supposed case $(2, 3)$, both agents know they both have positive numbers. But as established above, Anne does not know that Bill knows. So Anne might remain agnostic and since Bill knows that Anne might have a 2, suppose he chooses to do the same. This line of reasoning generalizes to any natural number pair in the consecutive number game just devised. The reasoning becomes less intuitive as the numbers go up, since it depend on increasing orders of social reasoning. For example if two real people were given 719 and 720 in a structure as above and asked the same question, they would probably agree that they are certain about it. In fact, they might be hard to dissuade otherwise because remaining agnostic relies on higher-order social reasoning out of cognitive reach for normal people. Reasoning about the possible number of the other depends on zero-order reasoning, while reasoning about what numbers the other considers possible is first-order, and so on as iterations continue. As we will see in a few sections, adults are often said to be limited to second-order reasoning at most, implying that while Anne and Bill would remain agnostic in state $(2, 3)$, they might answer differently in $(3, 4)$, since Anne would remain agnostic due to second-order reasoning, while Bill would not since it would rely on third-order reasoning. The next section focuses on some of the details of the limitation on social cognition.

3 Higher order social reasoning in real life

3.1 Limitations to the concept of theory of mind

Since the term 'theory of mind' have been used in various different ways, some conceptual clarification is in order. As mentioned above, the term sometimes conflates varieties of cognitive capacities and the general cognitive capacity to represent the mental states of others have also been called various different terms. Research in economics often refer to higher-order rationality[67] or reasoning [94] since equilibrium models of game theory depend on higher order beliefs, while other fields often focus more on mental terms. Some researchers like Schaafsma et al. [90] have called for a reformulation of the term 'theory of mind' due to its heterogeneous usage. This includes deconstructing it into basic processes before reconstructing a scientifically tractable concept of theory of mind. While a proper reconstruction of the theory of mind concept is outside the scope of this thesis, it is necessary to understand some of the heterogeneity of its use and meaning in order to avoid unnecessary confusion.

First, there is a debate about whether social reasoning constitutes a *theory*, implying it is a cognitive process similar to that of constructing a scientific theory [50], or whether it is a more of an automatic intuitive process of *simulating* the mental state of someone else [44, 61, 62]. This distinction is also somewhat reflected in studies distinguishing between reflective and spontaneous social cognition. Verbrugge & Mol [98, 99] argue that there is a relevant gap between adults reflective understanding of theory of mind and its application in games. The distinction between reflective

and spontaneous social cognition is mirrored in the debate about explicit and implicit social reasoning. According to Heyes [57], implicit mentalizing involves thinking about mental states in a fast, automatic way whereas explicit mentalizing is done in a slow and controlled way.

Second, there is a debate about *nativist* and *constructionist* accounts of theory of mind. The nativist account assumes that the capacity to represent the mental states of others depends on dedicated cognitive processes which have been uniquely developed through human evolution for its purpose [57, 70]. The constructionist, or developmental account, assumes this mental capacity to be learned and culturally inherited [57].

It is important to critically evaluate the concept of theory of mind since the conceptual nomenclature we use can affect our understanding of the actual mental capacity. Heyes [57] for example argues that a lot of studies referring to implicit mentalizing can be explained by what she calls submentalizing, a cognitive mechanism which simulates the effects of mentalizing in social contexts. Bloom & German [17] also present two reasons why the standard false belief task is not suitable for testing theory of mind. First, passing the false belief task requires other cognitive capacities than just a theory of mind, like general task demands, which causes young children to fail regardless of any theory of mind. Second, having a theory of mind involves more than the ability to reason about false beliefs. That is, even though children younger than 2 years fail the false belief task, studies show that they can still appreciate some of the mental states of others, like attributing goals to others [27]. So besides the different terminology used, the capacity of attributing mental states to others is possible a complex made up of various cognitive processes, which would plausibly explain the heterogeneity of the methods used to test it.

The next section is a discussion on how people diverge from logical prescriptions, particularly those from epistemic logic.

3.2 Idealizations and imperfect reasoning

People never reason perfectly but the extent to which people's reasoning is flawed is subject to debate. The Wason selection task [102, 103] was an early experiment showing limitations of zero-order reasoning. In Wason's study, subjects were shown sixteen cards with a letter on one side and a number on the other. Four of these, D, K, 3 and 7 were used. Subjects were then asked which cards to turn in order to evaluate whether the following claim was true: "Every card which has a D on one side has a 3 on the other". The claim has the structure of the material conditional $p \rightarrow q$. The correct cards to turn are those with p and $\neg q$ but this answer (D and 7) was only the fourth most popular answer, while the most popular answer was D and 3, which includes the logical fallacy of affirming the consequent.

Such results seem to imply that humans are poor logical reasoners. But there are a few complexities. First off, Wason also found that people were significantly better at answering correctly when

the question was about cities and transportation devices, specifically framed as "every time I go to Manchester I travel by car". Griggs & Cox [52] also show that subjects perform near perfectly if the cards include ages and beverages and the claim "if a person is drinking beer, then that person is over 19 years old". There is a long discussion on the thematic effect on the Wason selection task. One of the reasons for the different performance is arguably that the different thematic presentations warrants different interpretations. Stenning and van Lambalgen [93] argue that the abstract claim might be interpreted as merely checking satisfaction of instances instead of determining the truth of the rule. Wagner-Egger [101] argues that the 'error' made by most people may be due to interpreting the rule as a biconditional. The ambiguity of the statement could also explain why Cheng et al. [23] find that people may even continue to do poorly after an introductory logic class.

While this shows that there might be a non-deficiency explanation even when people seem to fail to make correct inferences in propositional logic, the fact still stands that people of course do not always reason perfectly. However traditional propositional logic often sets a standard that is low enough that it can be met at least in specialized settings where there is a special requirement to avoid logical failures. So propositional logic might not be overly idealized when it comes to arguments (ignoring the fact that propositional logic does not capture the dynamic and rhetorical aspects of real arguments). It is another story when it comes to higher order social reasoning. Before I go into empirical results on higher order social reasoning, let us look at some of the overly idealized aspects of $S5$.

3.2.1 Idealizations in epistemic logic

The properties of $S5$ imply that agents are agents know all logical truths. Transitivity ($K_a\varphi \rightarrow K_aK_a\varphi$) and euclidity ($\neg K_a\varphi \rightarrow K_a\neg K_a\varphi$) implies that agents have unlimited introspection of their own epistemic states, that is, they have a perfect account of what they know and do not know. But there are other idealizations which are less discussed. To see this, note that real knowledge is gathered in a dynamic social environment, where agents update and change their knowledge and beliefs as they gain new information. There are many aspects of the social dynamics of knowledge which are outside the scope of this thesis, but some of this dynamic can be expressed by our epistemic logic by adding action expressions as well as dynamic modalities for these. This means adding the formula $[!\varphi]\varphi$, with the truth condition:

$$M, s \models [!\varphi]\psi \text{ iff } M, s \models \varphi \text{ implies } M|\varphi, s \models \psi$$

It is often mentioned that public announcement of an atomic fact makes it common knowledge, while the same is not always the case for epistemic facts. Statements like "p is true but you don't know it" are so-called "Moore-type" sentences which leads to unsuccessful updates. But when announcing non-epistemic propositions, agents *ought* to come to know it. This rests however on

the idealized presupposition that agents have perfect observation, since the announcement and its contents must be clearly understood by all, and perfect recall, since agents must be able to remember all of this as time passes. In fact, for an announcement to become common knowledge, perfect observation and recall must be common knowledge among agents as well. Publicly announcing non-epistemic facts to real people do not always bring about common knowledge about such facts. Statements can either be too complex or lengthy for agents to decipher, or the statements might be ambiguous, as some researchers argue in the Wason selection task. In other words, if the mistakes in the Wason selection has to be put in terms of cognitive limits, they might not indicate a limit in computational power but rather a limit in observational power. As we will see later, limitations in terms of social cognition does not seem to occur due to limited zero-order reasoning, it is rather some barrier to the modal depth of formulas agents can come to know. The next section gives a brief overview of such parameters for cognitive limitations as found in Liu [72]. Following this is a discussion on empirical findings regarding higher order social cognition.

3.2.2 Parameters for diverse cognitive capacities

Fenrong Liu identifies five novel parameters for diversity among epistemic agents [72, p. 25f].

(a) Inferential or computational power: making valid inferences

(b) Introspection: being aware of one's own knowledge

(c) Observation: correctly observing current events

(d) Memory: capacity to remember observed events

(e) Revision policies: varying from conservative to radical revision.

Parameters (a) to (d) can be understood as logical norms describing the limits of truth-conducing capacities. That is, even if no one is perfect, each parameter is a way in which agents could be epistemically better off. This is difficult to see with (e) however, where the optimal case is not clear, since neither complete epistemic conservatism or revisionism seems rational. As Liu writes, it might be better to move away from understanding deviations from such norms purely negatively as 'limits' or 'bounds' on cognitive capacity, and instead see them positively as the different resources that agents have and use to successfully accomplish difficult tasks. After all, the story of humanity is largely a success story dependent on our cognitive prowess, even if so limited. As Benthem mentions, paraphrasing Joerg Siekmann, the most admirable in a mathematics seminar is not presenting a well-oiled proof, it is rather the ability to discover a mistake and recover on the spot. Rationality on this view is owed to learning from one's mistakes, and as Benthem points out, the dynamic behavior involved in this surely depends on more mechanisms than just inferences. This is just to present an

alternative picture of discussion on cognitive limitations which are not simply deficit-based. I will now present and discuss some empirical results concerning higher order social cognition.

3.3 The development and importance of theory of mind

Studies indicate that children learn to distinguish their own beliefs from others between the ages of 3 and 5 [105], while children learn to make correct second-order attributions between age 6 to 8 [82]. Chandler et al. and Leslie [22, 69] argue that sociocognitive processes are present even earlier and Csibra et al. [27] even argue that basic theory of mind aspects as goal perception are present already in 9 to 12-month-old infants. Adults are generally limited to no more than second-order social reasoning, even though there is evidence that adults can rely on third and fourth-order reasoning in games when playing against programmed opponents which they know to rely on specific higher-order social reasoning [97]. Before discussing the limitations of higher-order social cognition even in adults, I will briefly motivate why it is important to understand such sociocognitive limitations.

Before moving onto specific limitations, it is important to note why such limitations are relevant. Having an insight into the mental workings of others play a vital role in both social interaction and basic human functions such as having empathy towards others [16, 34]. There two other general reasons why it is important to map apparent limitations of this vital mental capacity.

First, we can only hope to counteract and adjust for possible cognitive shortcomings if we become aware of them.

Second, even when higher-order reasoning deviates from certain logical prescriptions it is not necessarily a logical mistake in the sense that the reasoning might simply fits some other logical framework. People interact and cooperate successfully with others all the time, so if social reasoning is limited, there must be other heuristics that take over. Finding out how reasoning is limited relative to logical frameworks makes it possible to figure out what such alternative heuristics might be.

Furthermore, ambient technology such as self-driving cars have to understand and predict human actions, and as such have to reason about how humans might reason. Such artificial agents would be wrong to assume that humans reasons like themselves. In fact, as Benjamin Erb [36] argues, when optimizing intelligent human-computer interaction it is important that both humans and their non-human counterparts can reason about the 'mental' states of each other. That is, for non-human agents to successfully predict and reason about the behavior of humans, they will have to simulate human beliefs and the usefulness of such a simulation hinges on how accurately the it represents the actual mental state of the human in question. Since the only salient factor here is accuracy of representation, questions of normativity can be ignored. This also holds the other way around, as Erb writes: "In intelligent, technical environments, humans may intrinsically apply ToM traits to their non-human interaction partners". Humans may in other words also attribute a 'mental' model to non-human interaction partners concerning the reasoning and intentions behind their actions.

This leads to some of various limitations of human social cognition.

3.4 The difficulty of reading minds

There are several different ways in which human cognition seems to be limited in terms of higher order social reasoning. It is limited in the sense that it is seemingly capped at first or second-order reasoning. Studies show that most adults in game-like settings can utilize first-order reasoning and do a good attempt at second-order reasoning, while having a higher order theory of mind is rare [39, 55, 98]. There are signs of more severe limitations however, indicating that social reasoning is either entirely absent or otherwise incomplete in ways which are not encapsulated by simply being restricted to a certain order of social reasoning.

3.4.1 Spontaneous versus reflective use of theory of mind

Experimental studies show that there is a relevant difference between having a capacity to correctly attribute mental states to others and actually using this in practice. Keysar et al. [66] argue that there is a stark dissociation between having the ability to reflectively distinguish one's beliefs from others and the routine deployment of this ability in interpreting the actions of others. Their findings imply that even adults who are capable of forming reasonable beliefs about the beliefs of others do not consult this crucial knowledge when interpreting the actions of others. Their claim is specific to the capacity of representing beliefs of others separately from corresponding reality, that is, acknowledging that others might have false beliefs. It suggests a difference between relying on one's theory of mind in reflective tasks and utilizing as a decision-guiding mechanism.

Flobbe et al. [39] also show examples of children who could understand second-order reasoning in story tasks while failing to properly perform second order reasoning in game tasks.

3.4.2 Curse of Knowledge

Birch & Bloom [15] show that adults who have to attribute beliefs to others have a tendency to attribute their own belief instead of correctly attributing a false belief. They write "adult's own knowledge of an event's outcome can compromise their ability to reason about another person's beliefs about that event". College aged adults were told a story where a girl Vicky puts her violin in a blue container. While Vicky is outside, her sister Denise puts the violin in a different container, depending on the treatment, either (Ignorance) simply in another container, (plausible knowledge) moves violin to the red container or (implausible knowledge) moves the violin to the purple container. The plausibly and implausible condition relates to the red container being violin shaped, while the purple container being an odd shape for a violin. When subjects did not know where the violin had been moved, they where generally good at predicting that Vicky would look for her violin in the blue container. When the violin was moved to either the purple or the red container, participants assigned

significantly higher probabilities to Vicky looking in the purple or the red container respectively (even though Vicky would still believe it was in the blue container). Participants also assigned higher probabilities to the treatment with the red (plausible) container than with the blue (implausible) one. That is, participants relied on their own beliefs when attributing beliefs to Vicky, and this erroneous first-order social reasoning was strengthened by feeling more justified in their belief. The authors call this the the curse of knowledge and is related to the problem in Keysar et al. as well. Both show that theory of mind use might be task dependent, which we will look at now.

3.4.3 Task dependence

The limitations described above suggest that successful application of theory of mind can be task-dependent, successfully applied in some contexts but not in others. Understanding why this ability might be task dependent can lead to better understanding the nature of higher order social cognition. Verbrugge [99] lists a few possible explanations. First, there might be a high processing cost associated with theory of mind, causing a failure in applying appropriate order of social reasoning when the the processing cost becomes too high. Second, the capacity for performing higher order social reasoning does not necessarily transfer between domains. Being apple to apply this cognitive capacity across domains might require a development process like Representational Redescription to take place as suggested by Karmiloff-Smith [65]. Third, in order to utilize a higher order theory of mind in a situation, it is not just necessary to possess this capacity, but also to recognize that it is advantageous to incorporate this knowledge into actions, decisions or interpretation of the actions of others. This is in line with the explanation that applying higher order social reasoning has a high processing cost. In other words, due to the high processing cost of applying it, it might simply not be worth the effort to rely on it per default, so it might largely be ignored until its importance becomes sufficiently apparent.

This is further supported by an argument in Keysar et al. [66] as to why adults do not always deploy their existing theory of mind. They mention that in the real word, perspectives often tend to coincide. This means that a lot of knowledge important for social interaction is common knowledge, which reinforces the point above that differentiating between one's beliefs and the beliefs of others might unnecessarily take up cognitive resources. As Keysar et al. mentions, the dynamic nature of face-to-face interactions give people a feedback mechanism which allows them to be egocentric by effectively distributing the burden of applying appropriate order social reasoning across interlocutors.

Lastly, theory of mind might not be a uniform of social cognition at all. Without taking a complete behaviorist stance, one might even question to what extend theory of mind relates to mental states at all. It is possible that when researchers deduce implicit social reasoning, it might just be learned behavior. For example when a person hears a cyclist ring their bell they likely infer an intention behind the action, that the person doing it is impatient. But this is could perhaps be explained by simple learned behavior. It is also likely inferred so spontaneously that it does

not seem right to say that people hear the sound and then reason about what it might mean. In phenomenological terms, the ringing of the bell might *sound* like someone is trying to get a reaction from you, which prompts the reaction without much explicit consideration of underlying mental states. If social reasoning is computationally 'expensive', this could explain how it can appear so extensively in everyday interaction, but also why it is difficult in more complex situations, since people might rarely be aware of the explicit use of their higher order theory of mind.

This concludes the introduction to different logical and empirical aspects of higher order social cognition which underlies the canteen dilemma. Logical prescriptions often help us navigating how humans ought to reason, but as we have seen, in order to make prescriptive norms for higher order social reasoning, it is necessary to look at the actual reasoning done by real people. This leads us to the canteen dilemma, an experiment on coordination with imperfect information which involve exactly this type of reasoning.

4 Canteen Dilemma

The canteen dilemma was devised by Thomas Bolander and designed and implemented in conjunction with Robin Engelhardt and myself. The game has structural affinity to the consecutive number example in section 2.1. It likewise depicts a situation where two agents have mutual knowledge about something iterated several times (everyone knows that, everyone knows that ...) while never having common knowledge (everyone knows that, everyone knows that ... infinitely iterated). The structure is framed in a thematic story since studies show that this can significantly improve people's zero-order logical reasoning abilities [76, 103] and this reasoning is not the focus of the canteen dilemma. 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 task for the players is to coordinate their decisions, while knowing that both have to go to the office if they arrive at 9:00 or later. Their preferences are ordered such that their prefer (1) going to the canteen together if both arrive before 9:00, (2) going to the office together at any time and (3) all other configurations, that is, discoordinating or either player going to the canteen at 9:00 or later. The game consists of a number of rounds where each player are assigned an arrival time and have to decide between going to the canteen or the office. The payoff structure follows a logarithmic scoring rule, meaning participants gave an estimate p of how certain they were that they made the right choice, implicitly assigning $100 - p$ to being wrong. Research on eliciting belief have shown that forecasts from observers through proper scoring rules are significantly more accurate and calibrated[2] than those elicited using improper scoring rules [91]. Players were shown

[2]Calibrated is defined as: "a set of probabilistic predictions are *calibrated* if p percent of all predictions reported at probability p are true" [91].

all previous results after each round, including the previous arrival times and choices of both players and their payoffs.

The rules dictate that players prefer going to the canteen if they both arrive before 9:00, but that they should always go to the office if they or their colleague arrives at 9:00 or 9:10. Take the following example to see some of the complexity of the game. Assume player *a* arrives at 8:50. Player *a* know that *b* arrived at 8:40 or 9:00. If *b* arrived at 9:00, *a* should go to the office. If *b* arrived at 8:40, they would prefer going to the canteen together. Player *a* knows they prefer going to the canteen but since one of the possible scenarios would require her to go to the office assume player *a* chooses the office at 8:50. Now assume *a* arrives at 8:40 instead and reasons the following way. If *b* arrives at 8:50 am and reasons like *a,* player *b* they would go to the office at 8:50 as well, in which case *a* might go to the office at 8:40 just like they did at 8:50. This line of reasoning follows for all possible arrival times by backwards induction, and so *a* might never choose to go to the canteen!

Notice that a player who arrives at 8:40 knows that the other arrived at 8:50 at the latest and that it is therefore preferable for both to go to the canteen. However, players do not just have to consider the possible arrival times, they also have to consider what others might do in such circumstances. In order to predict the choice the other player might make in their situation, the first player has to consider what the other player considers possible, which is different from their own point of view. That is, the player who arrives at 8:40 has to appreciate the fact that while she knows that both would prefer going to the canteen, the other player does not know this. So making an office choice at 8:40 or earlier requires a player to not just consider facts about the world, that is, if both arrived before 9:00 or not, but also consider what the other player would do in this situation, which depends on the beliefs of the other player. This means that predicting an office choice at 8:40 or earlier relies on higher order social reasoning. Of course, players can always choose office for whatever reason, for example if they just make random choices. But we would expect this to be both evenly distributed among both canteen and office choices as well as across arrival times. This means that the relative difference can still be indicative of reasoning of participants.

Notice however the premise above that for *a* to go to the office at 8:40, she has to consider how *b* reasons if she arrived at 8:50. It means that *a* assumes *b* to reason like herself. So this example does not just assume that *a* performs first-order social reasoning but also that she believes *b* to correctly apply zero-order reasoning to reach a specific conclusion. So it does not just rely on players reasoning a specific way but also that it is common knowledge among players that they reason a specific way. This means that regardless of what player *a* thinks *b* ought to do, *a* has to try and to figure out what *b* will actually do. It implies that without common knowledge about rationality, two rational players cannot be expected to play rationally. They might both choose canteen because they do not believe the other to apply the sufficient order of social reasoning, or that the other believes that they do and so on. While office choices arguably does signify social reasoning, choosing canteen does

therefore not signify a lack of it. To summarize, I postulate towo conditionals: **(1)** If a participant chooses office at 8:40 or earlier, they rely on higher-order social reasoning and (2) if a participant only uses zero-order reasoning, they will choose canteen at 8:40 and earlier, office at 9:00 and later and 8:50 being a toss up. I argue that choosing office at 8:50 does not necessarily rely on first-order social reasoning, since it is a fact about the world stated in the rules that the other might arrive at 9:00 and that you should choose office in that case. The converse conditionals are not taken to be true.

The certainty estimates involved can be valuable to get further information here. Assume a rational player plays the game and arrives at 8:40. They might infer that the optimal choice would be going to the office by relying on higher order social reasoning, assuming the other thinks the same way. But if they are uncertain about what type of reasoner the other player is, they might believe that the other player might play canteen and they can therefore hedge their decision by being less certain. So the certainty estimates helps offset some of the loss of possible discoordination if it can be predicted. This makes it possible to distinguish between players who all choose canteen at a given arrival time, since those who assign a lower certainty estimate to it than others implicitly assign a higher chance to office and therefore make use of higher-order of social reasoning per (1) above. This also help deflect what Verbrugge [99] calls the danger of simplistic formal systems which posit fixed bounds on social cognition, that is, assuming that people can reason up to n order but not $n+1$. This will also be the focus when viewing the results later, since it shows that natural social cognition might be more continuous than previously described. Before looking into the results, I will now go through some of the logical structure of the game in terms of epistemic logic. This is followed by supplementary discussion and results of the canteen dilemma not included in the accompanying article.

4.1 Logical structure of the canteen dilemma

Since the canteen dilemma has a logical structure similar to the consecutive numbers example in section 2.1, it can be represented in a similar diagram. Note that a situation like (8:00, 8:10) denotes that agent a arrived at 8:00 and b arrived at 8:10.

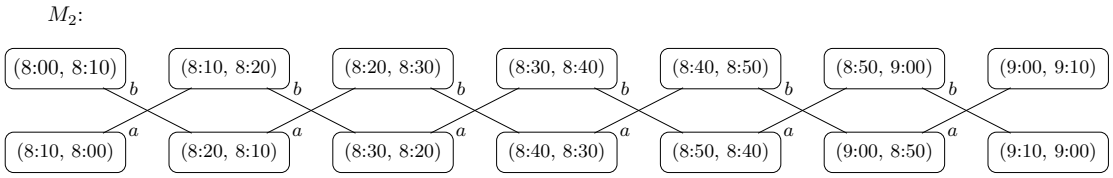


Figure 2: Unpointed Kripke model M_2 representing the possible arrival times in the canteen dilemma. Each state is named after the arrival times being true for (a,b) in that state. The relations between states indicates which states are indistinguishable for the players. Reflexive relations are omitted.

Figure 2 depicts the $S5$ model M_2. Let us first abbreviate the proposition that "both players arrived before 9:00" as (*early*). Recall that the highest possible payoff is both players choosing canteen, but only if (*early*) is true. The proposition is true in any state where 9:00 or 9:10 does not occur, so any but the four right-most states in M_2. Now suppose the actual situation is (8:30, 8:40) where a arrives at 8:30 and b at 8:40. Since the proposition (*early*) is true in every state which either a or b considers possible (connected to, in terms of the graph), they both know that they both arrived before 9:00 and that going to the canteen would be their preferred choice. This epistemic fact is not known by both agents however, since b consider it possible that she is in (8:50, 8:40), where a does not know (*early*): $M_2, (8{:}40,\ 8{,}30) \models E_{\{a,b\}}(early) \wedge \neg E_{\{a,b\}}E_{\{a,b\}}(early)$.

We can now show the structure of the canteen dilemma by relying on logical tools rather than joggling multiple iterations of knowledge. A player might think that it is not just necessary for *early* to be true in order to make a canteen choice, they must know that it is true. This excludes the four right-most states in M_2 as mentioned above. But since players in the canteen dilemma have to coordinate their actions, they have to consider how the other player might act, which requires reasoning about their beliefs. When a arrives at 8:40 then, she knows both players arrived before 9:00, but she does not know that b knows. If it is necessary for b to know (*early*) in order to go the canteen, then a considers it possible that b does not go to the canteen. If we couple this with the assumption that a does not want to go to the canteen if the other player might not make the same choice, then a will not go to the canteen at 8:40 either due to reasoning about what the other knows about their situation. This line of reasoning can be iterated for any earlier arrival time, implying that a and b will never go to the canteen. This is unintuitive however. Take for example state (8:00, 8:10) where everyone knows that everyone knows that everyone knows that everyone knows that it is early enough to go to the canteen.[3] That is four iterations of mutual knowledge: $E_{\{a,b\}}E_{\{a,b\}}E_{\{a,b\}}E_{\{a,b\}}(early)$. But it is not true for five iterations, that is, $\neg E_{\{a,b\}}E_{\{a,b\}}E_{\{a,b\}}E_{\{a,b\}}E_{\{a,b\}}(early)$ is also true in that state. So it is not common knowledge that it is early enough to go to the canteen and there

[3]Notice how unfit ordinary language is for portraying higher-order reasoning. Since our understanding of the is structured by language, it is possible that the limitations of ordinary language also play a role in limited higher-order reasoning.

is in fact no such time where (*early*) is common knowledge. This can be re-stated by saying that while choosing canteen at 8:50 involves a risk of discoordination, it is impossible to avoid any risk no matter how early you put the canteen choice. This nuance is hard to accept due to cognitive limits on our capacity for higher-order social reasoning. The infinitary modal depth is therefore another cognitively unrealistic idealizations of $S5$. The next section will present a case study of how epistemic logic might deal with this specific limitation.

4.1.1 Higher-order social reasoning in epistemic logic

Ditmarsch & Labuschagne [28] have proposed a logical framework capable of modeling various types of beliefs agents may have about the beliefs of others. This is done by modeling degrees of belief by partially ordered preference relations. Preference relations are motivated by the fact that without them, all epistemic alternatives in an $S5$ model are deemed equally possible. So instead of treating all accessibility relations between states equally, it becomes a preference ordering which express that agents may consider mutually exclusive states possible while considering one more likely than the other. This is the sense in which agents might *prefer* one state over the other.

According to Ditmarsch & Labuschagne, "a psychologically realistic model of an agent capable of interacting intelligently in social situations would be one in which each agent a would possess a ToM, comprising, relative to every agent b, a representation (which may or may not be accurate) of b's state of mind (accessibility relation, preference relation)" [28, p. 31]. This includes a zero-order theory of mind which does not represent a state of mind at all. The authors define a doxastic epistemic model, preference relation and degrees of belief. Their emphasis lies on the different classes of agents which are characterized by their preference relation. They first refer to a natural class of agents modeled after children with autism disorder. This class would consist of agents a who attributes to every agent b a preference relation which is identical to their own. The second class is an ideal agent a who would attribute a mental state to every agent b which is identical to the actual mental state of b. Ditmarsch & Labuschagne refer to this as precisely the situation modeled by the possible worlds semantics in epistemic logic. Third, a deranged agent a simply attributes random or innacurate mental states of others without systematic patterns. Fourth, there is the limited agent who can correctly attribute some but not all mental states to others. The limited agent is likely the psychologically most realistic one (possibly also for developmental disorders).

While Ditmarsch & Labuschagne takes cognitive insights seriously, there are some issues. As described in section 3.1, research on social cognition might not be as developed as logicians would like. According to those like Schaafsma et al., it is difficult even for experts to navigate what is meant by theory of mind. While Ditmarsch & Labuschagne refer to those with autism disorder as having a defective theory of mind due to failing a Sally-Anne experiment in a seminal paper on autism by Baron-cohen et al. [3], but 20% of the autistic children in the study in question passed

the first-order task while failing a different second-order task. Furthermore, some of those with autism disorder can pass theory of mind tasks consistently, making some postulate other cognitive impairments as explanatory forces for observed behavior [43]. Passing false-belief tasks might not be necessary nor sufficient for having a theory of mind either [17]. The problem of relating limitations of theory of mind to cognitive disorders is the contrast to normally developed adults, whose theory of mind is also limited, but which such a contrast ignores.

This does not bear on the logical validity of the work of the authors though. Logical theories have an important role to play in the study of higher-order social reasoning, but in order to make psychologically realistic models, it is important that the psychological reality is well understood and this is seemingly still under development.

For example as Verbrugge [99] writes, the simulation-theory might describe social cognition better than the theory-theory does. That is, we might simulate the minds of others rather than reason about them, but epistemic logic can still help cognitive science as Ditmarsch & Labuschagne writes, since logic can provide "computationally cheap" explanations for behavior.

Pol et al. [84] have also formalized theory of mind reasoning, specifically through updating beliefs about beliefs using dynamic epistemic logic (DEL). They found that theory of mind reasoning formalized as such is indeed intractable, meaning there is no efficient algorithm capable of solving it. The authors also argue that their findings suggest "that the intractability of theory of mind is not due to the computational demands imposed by 'higher-order reasoning,' as often assumed in cognitive science. Instead, our results suggest that intractability of theory of mind may be better sought in the computational demands posed by a more general form of reasoning about steps of change in time" [84, p. 291]. Such results are important for improving our understanding of what makes social reasoning so notoriously difficult as well as for implementing and simulating social cognition in artificial agents. As the authors mention it would be interesting to see other ways to possibly bound the order of reasoning in DEL models, since their logic did not bound reasoning but only prove it to intractable. Contrary to agents in DEL, human reasoning is bounded by physical constraints and not just computational complexity. That is, biological organisms are not necessarily capable of solving any tractable problem either. I now move on to the lack of introspection in relation to limited social reasoning.

4.1.2 Lack of introspection

While some of cognitive limitations comes with appropriate introspection, it is likely that limited social cognition is accompanied with a lack of introspection. While introspection in terms of zero-order reasoning is arguably easier, it becomes more complex for higher-order reasoning. Take an example from the canteen dilemma, where a player utilizes their first-order social reasoning to make an office choice at 8:40 but is otherwise sure about going to the canteen at 8:30. We might explain

the canteen choice at 8:30 by higher order social reasoning, because the person might believe that it is right to go the office at 8:40 but while thinking that the other player wont make that inference. A more plausible explanation could be that going to the office at 8:40 might have required n-orders of social reasoning, but going to the office at 8:30 required $n + 1$-orders, which were cognitively unavailable to the player.

So there might be a time where a player becomes certain that canteen is the right choice due to limited social reasoning and not due to lack of faith in the reasoning of others. However, this is due to a limitation of the same cognitive processes which are required to know that it is limited. Imagning oneself performing higher-order reasoning in 'metamode' seems to only make it harder. The next sections relate to results from the canteen dilemma. The accompanying article will contain the primary results and discussion, while the following sections will summarize supplementary results, discussions as well as considerations and improvements.

4.2 Results and discussion

The primary results of the canteen dilemma experiment are listed in the accompanying article.[4] The experiment was conducted both on Amazon's Mechanical Turk and at the Technical University of Denmark (DTU). Results from the canteen dilemma indicate that adults might not just have varying cognitive capacities in regards to the orders of social reasoning they can apply, but that there might also exist diversity even within those applying n-order social reasoning but not $n + 1$-order reasoning. This was indicated by the certainty estimates given by participants regarding how certain they were that the other player made the same choice as them. The argument for this conclusion is the following. Office choices can be interpreted as indicating some higher-order social reasoning. Those making canteen choices had varying degrees of certainty at different times, meaning they choose canteen at different arrival times, with less certainty towards later arrival times. Since the canteen/office choice was binary, being less certain about canteen means more certainty towards office being the right choice. So, if office choices indicate social reasoning, then lowering certainty about canteen choices does as well. But if certainty can be indicative of social reasoning, and certainty can be expressed in degrees, it seems possible that social reasoning can be exhibited in degrees similar to certainty estimates. This would imply in other tests on theory of mind that of those failing tasks requiring second-order reasoning, some might be closer to answering correctly than others due to a more developed theory of mind, even though this detail is ignored in binary results.

Another interesting result is the participant's lack of introspection into the limitations of their higher-order social reasoning. The results of the canteen dilemma generally show that players choose

[4] See https://github.com/thomasnicolet/canteen-dilemma-thesis for the Python code for all visualizations presented.

the office from 8:00 to 8:40, at 8:50 they choose somewhat randomly before they opt for the office at 9:00. Their certainty estimates indicate social reasoning in the sense that they are less certain about going to the canteen at later arrival times. But at earlier times, they become certain that the other player has also chosen canteen. This makes sense if it is exactly the type of social reasoning which is limited that is required to realize that the other player might chose differently from them. Instead of being less certain in such cases, it is plausible that people default to zero-order reasoning, assuming that the iterated mutual knowledge is simply common knowledge. It means that interactions within certain structures like the canteen dilemma can not just lead to unsuccesful interaction, it does so without participants being able to predict it or understand why it happens. We will see evidence for this when we look into free-text answers given by certain participants, which I will go through below as well as some supplementary results not included in the article.

4.3 Supplementary results

The first supplementary result is shown in Figure 3 below. It is motivated by the question whether participants always made the same choice for each arrival time or whether they changed their decisions throughout the game. The blue bars show the percentage of players who chose canteen at time t and then then chose office when they got time t again. The orange bar shows the converse for choosing office and later changing to canteen. Error bars show 0.95% confidence intervals.

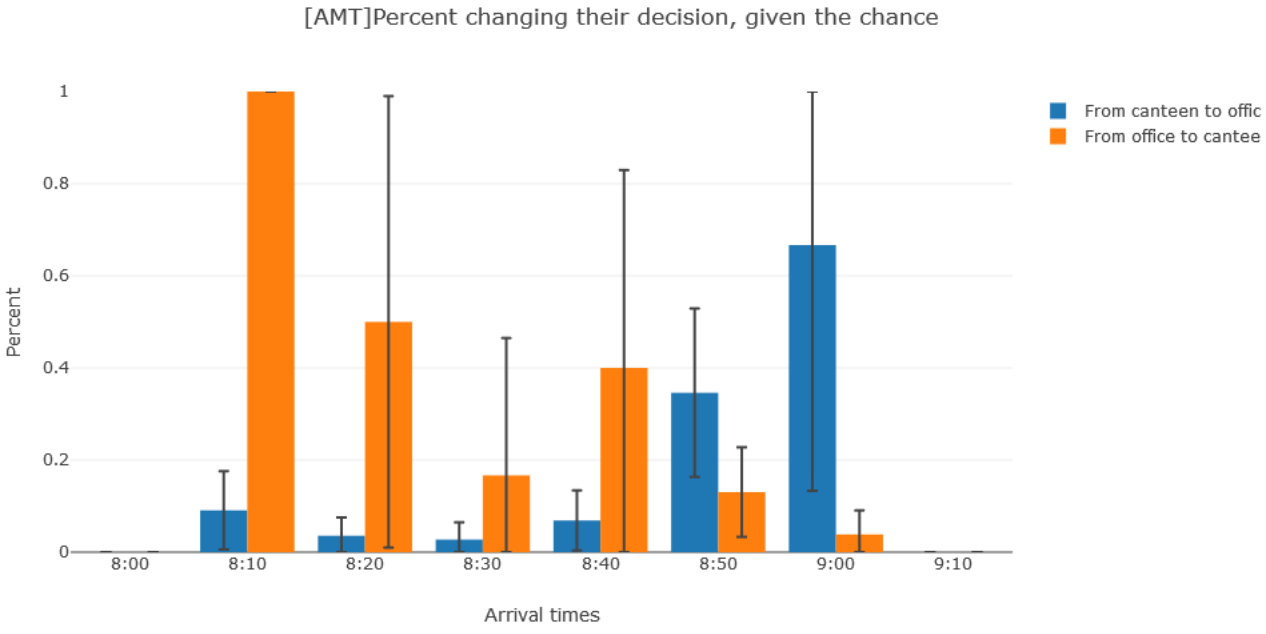


Figure 3. Grouped bar-chart indicating percent changing decisions throughout the game.

The confidence intervals show that participants did not frequently get the same arrival times multiple times, which results in wide confidence intervals. The plot generally shows that those choosing canteen at earlier arrival times did not change their decision later, while the few that chose office were more inclined to change to canteen. There is one interesting point at 8:50 however. The plot shows that those arriving at 8:50 were more susceptible to change from canteen to office than the other way around. This indicates that there is a type of learning happening at 8:50 but not at 8:40. The apparent lack of learning of higher-order social reasoning in games is also shown in Verbrugge & Mol [98].

4.3.1 Strategy categorization

Participants in both the AMT and DTU trials were asked the question “What strategy did you use while playing this game?". While these answers are harder to quantify than other answers, they provide insight into the reasoning and thoughts of the participants. This is helpful both because even if behavior in the canteen dilemma is correctly explained by a specific type of higher-order social reasoning, it only provides evidence of implicit reasoning. Participants may make decisions because of social reasoning which they are not themselves explicitly aware of. Explicit strategy answers provide insight into this aspect. These answers were quantified by devising a set of categories for various types of answers. The table below includes both the AMT and the two DTU experiments.

Strategy Categories	AMT	DTU1	DTU2
1. Guessing / miscellaneous answers	125 (18.6%)	7 (8.9%)	1 (2.4%)
2. Non-random strategies, not fitting in other categories	145 (21.8%)	9 (11.4%)	8 (19%)
3. Loosely time dependent strategies	163 (24.3%)	13 (16.5%)	5 (11.9%)
4. Canteen at 8:50, office at 9:00 and later	46 (6.9%)	6 (7.6%)	8 (19%)
5. Canteen at 8:40, office at 8:50 or later	56 (8.3%)	12 (15.2%)	7 (16.7%)
6. Canteen at 8:30, office at 8:40 and later	8 (1%)	5 (6.3%)	1 (2.4%)
7. Explicit first-order social reasoning	28 (4%)	9 (11.4%)	0 (0%)
8. Explicit second-order social reasoning	0 (0%)	2 (2.5%)	0 (0%)
9. Behavior based, reacting to the other	81 (12%)	3 (3.8%)	1 (2.4%)
10. Learning or changing strategy through game	18 (2.7%)	9 (11.3%)	6 (14.3%)
11. Preferring office only strategy	0 (0%)	3 (3.8%)	0 (0%)

Figure 4. Table of categories of strategy answers in both AMT and DTU trials.

One of the dominating categories was (3) which included answers which simply state that their strategy was based on the time they arrived, without much clarification. Many of these answers had the form of going to the canteen early and office later, see for example the AMT answer “How close to the 9:00 cutoff I arrived at work determined where I would go”. This answer is consistent with most if not all strategies, but the most plausible interpreation is that it refers to going to the canteen with certainty until some time before 9:00, where they either became less certain or started to make office choices. Other answers also allows for better interpretation of behavior in terms of higher-order social reasoning. See for example the AMT answer “I tried to think of the other guy arriving 10 minutes later than I did and choose accordingly”. This plausibly implies going to the office at 8:50 and canteen at 8:40. But does it involve infer representational mental states about the other? Possibly not, because the other person’s arrival time is a fact about the world which they can infer based on their own arrival time. That is, when arriving at 8:50, the other person might have arrived at 9:00 and this can be deduced regardless of the other persons mental state. See also the AMT answer “The farther from 9 the more certain I was of canteen. If it was 9 or later I was more certain of office”. Being less than maximally certain about going to the canteen at 8:40 and earlier requires higher-order social reasoning, but the answer in question does not refer to this reasoning at all. Of course they might just not mention it, but it is also possible that there is an implicit or automatic reasoning process which intuitively informs the participant’s choices. The reliance of intuition and common sense is also indicated by some answers, for example in the AMT answer, “Common sense, at least I thought :(“ indicating both that they relied on common sense while also acknowledging that something seemed to go wrong.

4.3.2 Supplementary free text questions

Participants in the DTU trials were asked a few supplementary questions besides those in the AMT trial. Participants in the DTU trial were asked the same question as in the AMT trial: “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 most popular answer was 8:40, while around 73% answered either 8:30, 8:40 or 8:50. See Appendix A for a barchart for these results. After answering this question, DTU participants were asked the question “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?”. The results were categorized in positive, negative and miscellaneous answers, which can be seen in the following barchart.

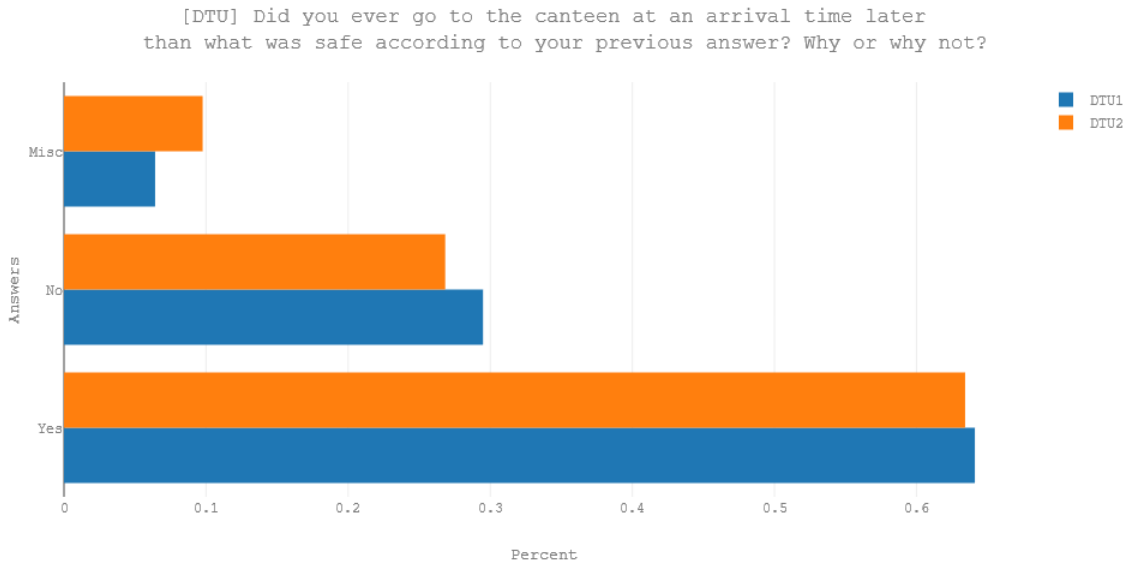


Figure 5. Grouped-barchart of categorized answers to question about whether participants went to the canteen later than what was deemed safe.

These answers generally indicate that participants in both DTU trials answered positively (63% and 64%) that they did go to the canteen at times later than when would have been if they could have made an agreement with their colleague. Some of these answers are qualfied with reference to lower certainty, see this answer from the first DTU trial: " Yes I did, but then I could not be very certain ". Other answers are similar, like "Yes, at 8:40. Because it seemed almost safe ". Other answers refer to intuition as well: " Yes, because it was possible, to go there safely and intuitively made sense". The last answer was given by a participant answering answering that their cutoff for safely going to the canteen would be 8:20. This means going to the office at 8:30 (and later) which is indicative of second-order reasoning. When choosing the canteen at later times anyway, it is possibly because this reflective reasoning does not seem as important in practical cases. The reference to intuition possibly means that even when their higher-order social reasoning entails that going to the canteen is not safe at later times, their intuition overrides this and they rely on some other default reasoning.

4.3.3 Changing strategy during the game

Participants in the DTU trials were also asked "Did you ever choose differently after seeing the same arrival time again at a later point in the game? Why or why not?". These answers were also put into three categorizes, positive, negative and miscellaneous. The results are below.

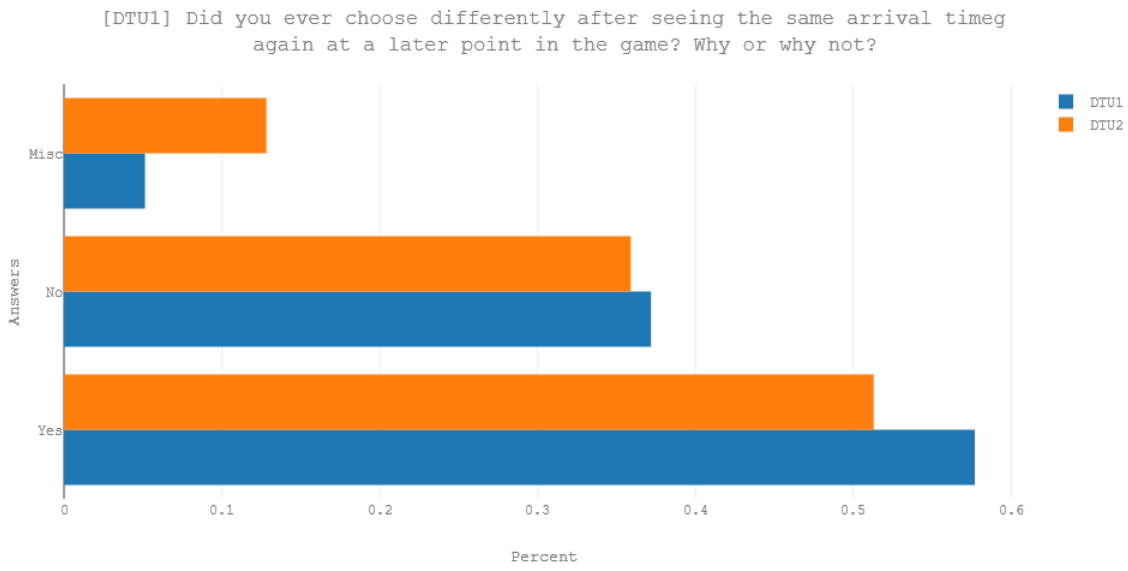


Figure 6. Grouped-barchart of categorized answers pertaning to whether participants changed their strategy.

Answers are grouped around positive affirmation again (51% and 57%) rather than negative (35% and 36%). The positive answers refer to either adjusting their strategy based on their partners earlier choices or from learning more about the game. One person explicitly states: "Yes because after a few rounds I realized it is never safe to go canteen.". The positive answers from the first DTU trial indicated a focus on learning about the structure of the game and the impossibility of safely establishing canteen coordination compared to the second DTU trial which relied more on reacting to their partner. The general dominance of positive answers indicate that participants do not rely on backwards induction alone, viewing each round as a one-shot game. Rather, many took earlier behavior into account, besides also reasoning more about the structure of the game.

4.4 Supplementary discussion

An essential part of experiments which elicit beliefs from participants depend on other implicit assumptions about their beliefs. Inferring and understanding why participants might have behaved like the did in the canteen dilemma therefore depends on implicit assumptions about their understanding of the structure of the game. There is as such a few aspects of the results which warrants methodological discussion.

Participants in both AMT and two DTU trials chose canteen at 8:50, 47%, 43% and 64% of the time, respectively. The rules of the game is intended to entail the three following propositions: (a) you are given an arrival time each round and know the other player arrives 10 minutes before or after yourself, (b) you have to chose between canteen and office in every round, (c) if you arrive at 9:00 or 9:10 you have to go to the office and (d) you have to do the same as the other player.

Why does a significant amount of participants chose canteen at 8:50 then? It is possible that they understood (a) to (d) above and simply took a chance, knowing the other might arrive at 9:00 and knowing that that the other would choose the office. This explains why it is centered around 50%, as it is interpreted to be 50/50 what the other player does. This is also shown in some strategy answers, see for example the following:

> "I tried to play it safe. It's obvious to go the canteen anytime before 8:40, but when 8:50 comes up it's a gamble. I messed up by assuming my partner would agree once, but after that I played it safe so I would lose less." (AMT strategy answer)
>
> "I tried to guess where they would go. In some cases this was a 50/50 chance." (First DTU trial)
>
> "If i arrived at 0900, go to office. At 0850, assume other person went to office and go to office. At 0840 assume other person assumed I went to office and go to office At 0830 50/50 bet At 0800 to 0820 assume other is also betting on canteen. " (First DTU trial)

The last strategy answer refers to going to the office at 8:50 and 8:40 as well. Notice that at 8:30, they apply second-order reasoning, as they imagine the other arriving at 8:40 and attributes first-order reasoning to them. This means that the random choice at 8:50 is moved to 8:30 for this participant, while higher-order reasoning is ignored for earlier arrival times. The first answer written above even states that it's obvious to go to the canteen at earlier times, likely due to lack of higher-order social reasoning.

It is possible that participants chose canteen at 8:50 due to different reasons too however. They might have interpreted the rules such that they did not infer the propositions above. They might not have inferred (c) and thought they could go to the canteen at 9:00, meaning it would be okay to go to the canteen at 8:50 if the other did the same at 9:00. They might have also have ignored (d) and thought that they could go to the canteen as long as *they* arrived before 9:00. The rules in the AMT and first DTU trial stated "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" which can be reasonably interpreted as 8:50 being an acceptable time to go to the canteen. While a reasonable interpretation, the rules continue stating: "If you or your colleague arrive at 9:00 am or after, you should go straight to your offices". This arguably overrides the previous statement, but since the other statement comes first, it is possible it has been prioritized. This was tested however in the second DTU trial, where the

sentence was changed to "If you *both* arrive before 9:00 am ..." where 62% of participants choose canteen at 8:50 (see Appendix B for plot of their choices). While different populations, it provides some evidence that the formulation of the rules did not lead to canteen choices at 8:50. Canteen choices at 8:50 might therefore be understood as gambles, which were somewhat calculated as we also see the lowest certainty estimates at this arrival time.

There is also the question why some (circa 10% in the AMT trial) have chosen canteen at 9:00 and 9:10 which are zero-order reasoning failures. Part of the explanation is random answers, since we would expect some participants to either choose randomly or intentionally choose opposite their preferences (trolling). This accounts for some of the answers which also explains some of the around 7-10% of participants choosing office from 8:00 to 8:30. In other words, this might be the level of background noise in the data, such that these office choices are not necessarily indicative of reasoning either. This seems to be supported by the first DTU trial. The first DTU trial included students enrolled in a course focused on logical models of artificial agents. The demographic was therefore both primed in terms of theory of mind and in a setting where they are used to be tested on their interpretation of logical of statements. The first DTU trial had 4% and 0% choosing canteen at 9:00 and 9:10 respectively, and 0 percent choosing office at 8:00 and 8:10. It indicates that these DTU students generally had a good a zero-order understanding of the structure of the game, which means that it was at least possible to make such an interpretation of the rules. Another possibility as to why the amount of canteen choices seem high is the effect of the framing narrative. Participants might rely on their own preferences over those stated in the game, implying that they do not really believe that there is any benefit of going to the office together before you absolutely have to. While they might intuitively understand the preference of going to the canteen with their colleague, it might be less intuitive to understand why going to the office is only preferable if the other does so as well. This explanation assumes participants to have understood the rules well, but without being sufficiently aware of the payoff structure.

4.4.1 Improvements

Continuing from the discussion above, there are a few ways the canteen dilemma experiment could have been methodologically improved.

First, pilot-testing is always important in such experiments and the pilot-testing of the canteen dilemma could possibly have been improved by meta-questions. These could be general questions about the game, such as whether they were able to understand and recall the rules, if they had had enough time in the experiment or similar questions which would make it easier to optimize the description of the game. This could also involve participants playing a set of practice rounds before entering the game.

Second, the behavior of participants could have been interpreted more robustly by including

linguistic controls. A linguistic control question would be question that could be answered correctly if and only if a person has made a specific interpretation of the rules. Such a question could ask participants what they should do in various crucial scenarios in the game.

Third, the general interpretation of the rules could also have been aided by re-stating them in a more intuitive story. The implicit possibility of going to the office at any time could also have been emphasized, or tested for in a linguistic control. The rules could for example emphasize a story where two colleagues generally want to work on an important project together, which requires both of them to work on and which they have to work on from 9:00 at the latest, but also possibly earlier. This could mean pumping the intuition of always being allowed to work on a project together, while not always being allowed to go the canteen due to an important deadline. Even though the payoff structure implicitly emphasized the importance of avoiding discoordination (more than half of the participants in the AMT trial lost their entire bonus due to discoordination), the facilitating narrative could be strengthened by emphasizing the importance of avoiding discoordination. The participants could also have been given the same arrival times in a fixed order with this being common knowledge. The game could also be structured differently such that participants were automatically assigned to the office at 9:00 and 9:10. A more drastic change would be that payoff for choosing office was independent on the other player while payoff for choosing canteen was not.

Fourth, the structure of the game and its questions could be tailored towards more narrowly defined hypotheses, as well as having a more narrow focus within social cognition. Some existing questions could simply be altered, for example 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?" could be changed such that the arrival time in question was randomly generated. Delimiting terminology and focus within social cognition is of course difficult due to the diffuse taxonomy of the field. The focus of the canteen dilemma could have been delimited by focusing either on explicit/reflective social cognition or implicit/spontaneous social cognition. This could also be testing the degree to which such reasoning could be replicated by other cognitive processes which does not involve representation of mental states, like 'submentalizing' described by Heyes [57], which she argues could be both a substrate and subsitute for thinking about the minds of others.

4.4.2 Future research

The cognitive capacity of attributing mental states to others have been the focus of numerous studies but still remains surprisingly enigmatic. It has often been described as 'theory of mind' but this term has been criticized by researchers like Schaafsma et al, who argue that its meaning is often "vague and inconsistent, its biological bases are a subject of debate, and the methods used to study it are highly heterogeneous [90, p. 65]. The cognitive capacity in question is undeniably central to

our social life however and it is therefore important to understand its complexity. Appreciating that others have minds of their own could consist of a conglomerate of implicit and explicit cognitive processes, some which might relate more to behavior than mental states. This calls for research of the possibly unexplored diversity of processes and aspects of social cognition, including subsequently improved terminology, which would aid numerous research fields.

Cognitive science and psychology could benefit in terms of improving our understanding of both cognitive development in children as well as the fully fledged capacities and limitations of adults. By getting a more accurate description of how adults infer representational mental states of others, logic and computer science could make more realistic computational models of human reasoning. If artificially intelligent agents are to understand and predict human reasoning, including prediction of how their own action might be perceived by humans, it is important that they have an idea of how humans might reason and not just how they might ought to reason. As Heyes [57] argues, some cases where mentalizing is inferred might better be understood as sub-mentalizing, that is, domain-general cognitive mechanisms which simulate mentalizing in social contexts. As has been mentioned previously, this supports the argument from Erb [36], that humans might very well start assigning intentions, beliefs and motives to non-human counterparts. As development continues on ambient technology such as self-driving cars, voice assistants or other technology involving human-computer interaction, it comes increasingly important for computational agents to be able to both make human-like inferences such that humans can understand and predict their behavior, and make inferences about human inferences in order to predict their behavior.

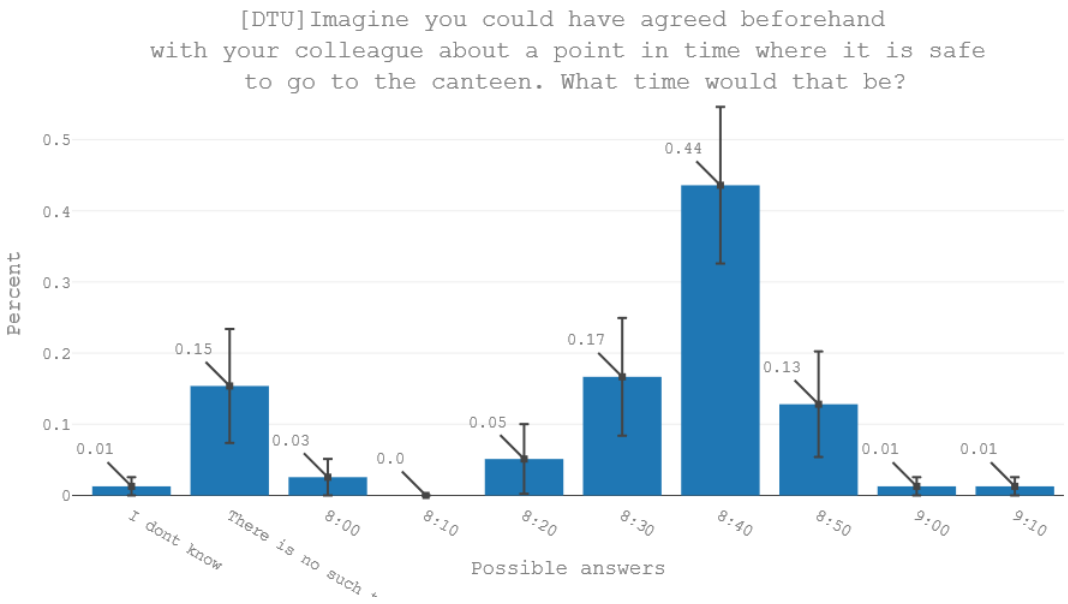
Due to the central role which intentions, beliefs and mental states have in philosophy, developing a nuanced understanding of social cognition is relevant for philosophy as well, from epistemology, philosophy of mind, ethics and logic. While social cognition by nature ignores some philosophical problems like solipsism, it can have immense importance in relation to modern problems. Agent-centered deontological moral theory states for example that actions are permissible depending on the intentions of the agent, but this means that any moral evaluation of the actions of others depends the capacity to understand the intentions of others. That is to say that insights from social cognition can put moral problems in an epistemological light. Problems in social epistemology like pluralistic ignorance are also highly related to such research.

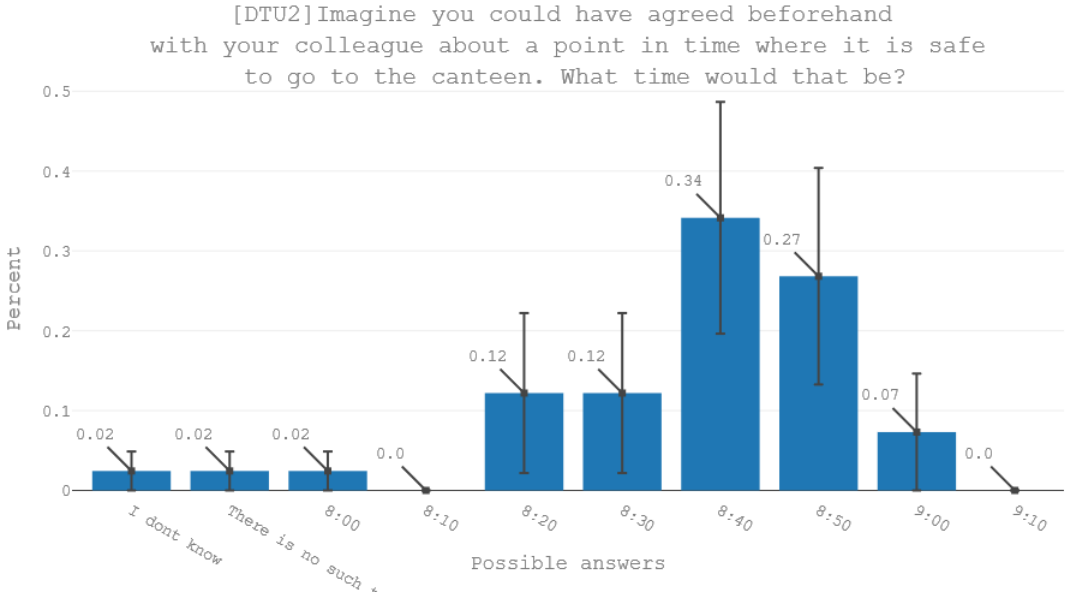
The canteen dilemma added to the possible diversity of this cognitive capacity, as it indicates that there might be variation even among just n-order social reasoning. A possible avenue for future research would be to map some of these details, for example testing if and under what circumstance some first-order reasoners might be closer to second-order reasoning than others. Empirical nuance also makes it easier to establish a scientifically tractable terminology for accurately describing this interesting cognitive phenomena. While the term 'theory of mind' is sometimes used vaguely, it is understandingly difficult for researchers to use better terms for a concept if its complexity is not yet sufficiently explored.

5 Conclusion

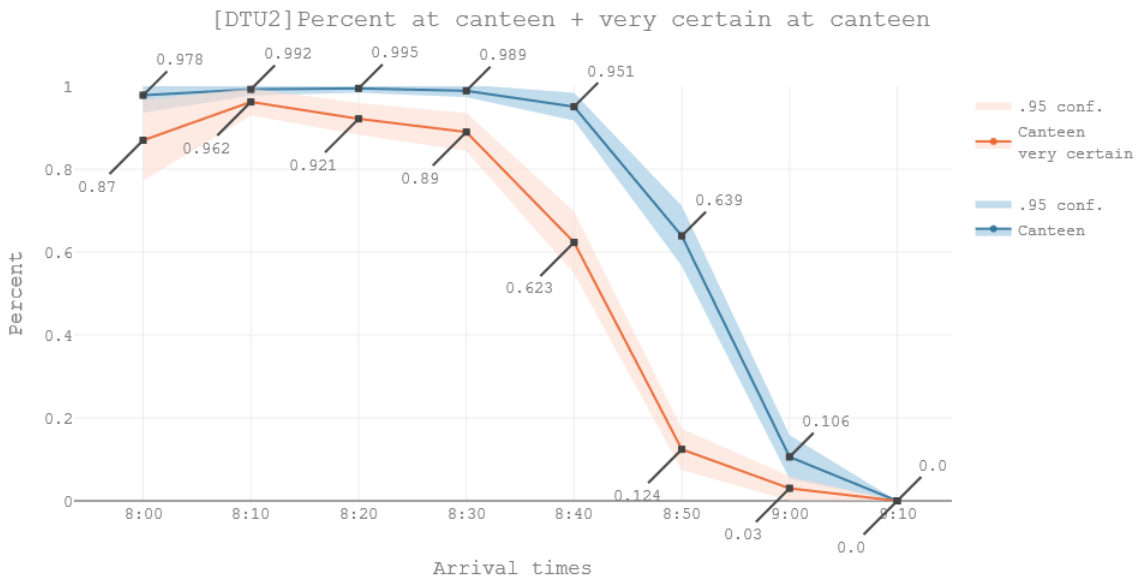
This thesis presented and introduced a large and still developing research field on higher-order social cognition. This is the study of the cognitive capacity to mentally represent possible mental states of others, such as beliefs, desires or intentions. The thesis included a critical discussion of the heterogeneity of terms and methods used to study this capacity. I also discussed the importance of researching actual cognitive capacities in relation to logic and computational models for reasoning in computer science, often thought to be normative in nature and independent of how humans actually reason. This works as an introduction to the accompanying article on the canteen dilemma experiment conducted at the Center for Information and Bubble Studies (CIBS). Various results are discussed, such as limited higher-order social cognition as well as a lack of introspection. Results indicated that even within those who only have the capacity for n-order social reasoning, some might be closer to $n+1$ than others. This implies some type of continuity of the various degrees of higher order social reasoning which are ignored by typical descriptions of higher-order reasoning in integers. If this result can be corroborated, it can improve both (i) our understanding of the developmental aspect of social cognition, since it does not posit some jump between orders of reasoning and (i) psychologically realistic models of human reasoning in logic and economy, since it shows evidence of types of agents otherwise ignored.

Appendix A


[DTU]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?
Percent
0.5
0.4
0.3
0.2
0.1
0
0.01
0.15
0.03
0.0
0.05
0.17
0.44
0.13
0.01
0.01
I dont know
There is no such ...
8:00
8:10
8:20
8:30
8:40
8:50
9:00
9:10
Possible answers


[DTU2]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?
Percent
0.5
0.4
0.3
0.2
0.1
0
0.02
0.02
0.02
0.0
0.12
0.12
0.34
0.27
0.07
0.0
I dont know
There is no such ...
8:00
8:10
8:20
8:30
8:40
8:50
9:00
9:10
Possible answers

Appendix B


[DTU2]Percent at canteen + very certain at canteen
Percent
Arrival times
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

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 (1991) *Language in action: Categories, lambdas and dynamic logic.* North-Holland, Amsterdam (paperback edition 1995, MIT Press, Cambridge)

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

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

[9] 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.

[10] 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.).

[11] 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)

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

[13] 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.)

[14] 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

[15] 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

[16] Birgit A. V., Alexander N.W. T., Paul R., Rhiannon C., John S., Shane M., John F.W. D., Rebecca E. (2006). *Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task.* NeuroImage, Volume 29, Issue 1, pages 90-98, ISSN 1053-8119.

[17] Bloom, P. and German, T.P. (2000). *Two reasons to abandon the false belief task as a test of theory of mind.* Cognition 77, B25–B31

[18] Brewka G. (2012). *Default Reasoning.* In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA

[19] 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.

[20] 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

[21] 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.

[22] 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

[23] 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

[24] 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

[25] 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.

[26] 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

[27] 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.

[28] 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.

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

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

[31] 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.

[32] Drew M., Doyle J. (1979). *Non-monotonic logic 1.* Artificial Intelligence, Volume 13, Issues 1–2, 1980, Pages 41-72, ISSN 0004-3702.

[33] 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.

[34] Dvash, J., Shamay-Tsoory, S. G. (2014). *Theory of Mind and Empathyas Multidimensional Constructs.* Top Lang DisordersVol. 34, No. 4, pp. 282–295

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

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

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

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

[39] 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.)

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

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

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

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

[44] Gallese, V. and Goldman, A. (1998). *Mirror neurons and the simulation theory of mind-reading.* Trends Cogn. Sci. 2, 493–501

[45] 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.

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

[47] 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

[48] 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.

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

[50] Gopnik, A. and Wellman, H.M. (1994). *The theory theory.* In Mapping the Mind (Hirschfeld, L. and Gelman, S., eds), pp. 257–293, Cambridge University Press

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

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

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

[54] 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.

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

[56] 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.

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

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

[59] 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

[60] 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.

[61] Hurley, S. (2005). *Social heuristics that make us smarter.* Philosophical Psychology, 18(5), 585–611.

[62] Hurley, S. (2008). *The shared circuits model: How control, mirroring and simulation can enable imitation, deliberation, and mindreading.* Behavioral and Brain Sciences, 31, 1–22.

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

[64] 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.

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

[66] 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.

[67] Kneeland, T. (2015). *Identifying Higher-Order Rationality*. Econometrica: Sep 2015, Volume 83, Issue 5. p. 2065-2079.

[68] 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).

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

[70] Leslie, A. M., Friedman, O., & German, T. P. (2004). *Core mechanisms in "theory of mind."*. Trends in Cognitive Sciences, 8, 528–533.

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

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

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

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

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

[76] 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.

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

[78] Nagel, T. (1974). *What Is It Like to Be a Bat?*. The Philosophical Review. 83 (4): 435–450. doi:10.2307/2183914. JSTOR 2183914.

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

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

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

[82] 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.

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

[84] Pol, I. v. d., Rooij, I. v., Szymanik, J. (2018). *Parameterized Complexity of Theory of Mind Reasoning in Dynamic Epistemic Logic.* J Log Lang Inf (2018) 27:255–294 https://doi.org/10.1007/s10849-018-9268-4.

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

[86] 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.

[87] 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

[88] 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.

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

[90] 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

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

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

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

[94] Strzalecki, T. (2014). *Depth of reasoning and higher order beliefs.* Journal of Economic Behavior & Organization, Volume 108, 2014, Pages 108-122.

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

[96] 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.

[97] 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.

[98] 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.).

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

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

[101] Wagner-Egger, P. (2007). C*onditional reasoning and the Wason selection task: Biconditional interpretation instead of reasoning bias.* _Thinking and Reasoning_ 13 (4):484 – 505.

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

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

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

[105] 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.

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

[107] http://www.glascherlab.org/social-decisionmaking/ (15-01-2019)

[108] https://plato.stanford.edu/entries/ethics-deontological/#DeoThe (14-05-2019)