Logic and social cognition in the canteen dilemma

Thomas S. Nicolet

April 30, 2019

Abstract

Successful social interaction often requires people to reason about the mental states of others, also called having a theory of mind. This equivalent to the ability of being able to make a mental model of other people's beliefs, desires or intentions. While research show that the ability is acquired in childhood, the application of this ability is limited in several different ways. The canteen dilemma is a game which investigates the extend of this limitation. Results from the canteen dilemma indicate that the capacity for higher order theory of mind is finer grained than simply being able to reason in n orders of theory of mind but not $n + 1$. This thesis works as an introduction to the accompanying article by providing sufficient background theory to the related research field.

Contents

1 Introduction (revise)

Successful social interaction often requires us to reason about and understand the mental states of others. This type of social cognition is called our theory of mind and is said to be the pinnacle of social cognition [17]. It is the capacity to attribute to others otherwise non-observable mental states such as beliefs, intentions and knowledge. Being aware of how others might reason, including their reasoning about oneself and so on, leads to efficient and intelligent social interaction. It is the ability that allows us to not just predict the behavior of others, but also predict how our own behavior might affect others. As such, it leads to a more complete and accurate understanding of other social beings around us and which has arguably lead to the centrality of this capacity when it comes to social cognition. This reasoning can be done recursively, modeling the mental state of others, including their model of one's own mental state, and so on. This recursive modeling of the mental states of others is called higher-order theory of mind and is the general focus of the canteen dilemma and this thesis.

Higher-order social reasoning can be described in epistemic logic, the modal logic for knowledge. In general, we can say zero-order theory of mind is like non-modal logic as it concerns facts about the world, while $k+1$-order theory of mind concerns facts about k-order reasoning of other people [68]. As higher order social reasoning can be described in various epistemic logics, the interface between logic and cognitive science has recently been receiving interest. There lies a two-fold question in this field: To what extend does real social cognition differ from the prescriptions from pure epistemic logic and how does logic and cognitive science react to possible divergence? In the first section below I'll discuss the relation between logic and empirical facts about human reasoning. I then present some epistemic logic before I go into insights from experimental cognitive science regarding higher order social cognition. This works as an introduction to the canteen dilemma experiment and the research field its situated in. As the essential details and results of the experiment are fleshed out in the accompanying article, this introduction will remain more general and describe related areas of research which motivated the experiment.

1.1 Logic and the facts - from Frege to Benthem

Human reasoning has a long but somewhat intricate relation to logic. Argumentative logic has been studied since antiquity and is still frequently taught to first year philosophy students in order to better understand and make valid inferences. Logic has traditionally been thought of as independent of empirical facts and the relation between logic and argumentation has only consisted in logic being a normative force of how we ought to make inferences. The relationship between human reasoning and logic is not clear-cut however,and it has been changing in recent years as modern logic has been undergoing a cognitive turn notably due to logicians like Johan van Benthem and Rineke Verbrugge [7, 68].

This cognitive turn is a step away from the anti-psychologism espoused by early logicians, notably Frege who 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" [29, p. 250]. Frege refers to the normativity of logic but his point rests on the more fundamental claim that logic is the study of objective mind-independent logical laws, as he writes: "the laws of truth are not psychological laws: they are boundary stones set in an eternal foundation" [28, p. 13]. In other words, logical laws are normative because of their objectivity. The norms for our reasoning and beliefs are structured by what is truth-conducing and the laws of logic are norms for reasoning exactly because they are objective truths, in fact necessary and eternal truths. Husserl followed Frege in this anti-psychologism and even argued that any argument that makes the logic independent from psychology rest on the distinction between how we ought to think and how do do think as implicitly psychologistic, since such an argument still depends on human reasoning [1]. For Frege and Husserl then, the normativity of logic is a side-effect of its objectivity.

This leads to a fundamental debate in philosophical logic concerning what makes logical truths true, i.e. what are logical facts facts about and are are they necessarily or contingently true? There is a rich historical discussion about this from Wittgenstein, Ayer, Carnap, Quine, Kripke and Putnam among others, but I'm going to sidestep this and focus on a more simple observation. Even if one argues for a non-Platonistic Realism somewhat like Maddy [51] or perhaps Quine [59], that is, the truth of logical laws are contingent on the right physical structuring, we can still hold onto the normativity of logic in relation to human reasoning. How does empirical facts relate to logic then?

When it comes to argumentative logic, there are certain inferences that humans make that deviates from logic in ways that can only be described as mistakes. Modus Ponens for example, from P and $P \rightarrow Q$, we can conclude Q, is true without any complications and failing to make such inferences can have drastic consequences. But there are complications when it comes to both richer epistemic logics and the ambiguity of natural language. There are several reasons why collaboration between logic and experimental cognitive science can be fruitful. Firstly, even if people do not actually reason like the prescriptions of logic dictate, people are are often highly successful in their endeavors none-the-less. That is, when people's reasoning deviate from logical prescriptions, we cannot necessarily deduce that their reasoning is illogical or randomly structured. There might be a more complex unknown logic underlying their decision, which can

be reconstructed and used to more accurately represent how people actually reason. Different people might reason differently and logic should be aware of this diversity if it wants to maintain the usefulness of its normativity.

Secondly, since AI has been described as the discipline of "making machines behave in ways that would be intelligent if a human were so behaving" by John McCarthy in 1956 [49], we need to be aware of how people actually reason if we want artificial intelligent technology to behave humanly. This is especially important when it comes to higher order social cognition. If we want to be able to interact with an AI in a way that is humanly intelligent, it will have to be able to reason about and understand the reasoning of humans. This is essential for understanding and predicting the actions of humans which is essential for cooperation. In other words, behavior which deviates from idealized logical prescriptions is not necessarily based on flawed reasoning and can't be *carte blance* dismissed as illogical, and AI systems should be aware of that if they are to understand and predict human behavior. In order to see how such an idealized representation of reasoning about knowledge might look I will start with an account of epistemic logic, originally used to formalize and analyze epistemological notions and concepts. Epistemic logic is useful for showing the depth and intricacy of real world examples, however it often involves too idealized assumptions about human reasoning. After going through these idealizations I will move onto real higher order social reasoning.

2 Epistemic logic - normative aspect

Epistemic logic is the modal logic of knowledge and as such can be very useful for keeping track of what agents know at specific static stages of an informational process. I will here introduce the most popular logical system $S5$ which is the epistemic logic most often used to symbolize knowledge, although with idealization problems we will turn to later.

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:

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

From this definition we can build formulas from atoms and more complex formulas using the negation "$\neg$", conjunction "$\wedge$" and knowledge operator "K_i". Before we explore the art of modeling in epistemic logic I'll define Kripke models and truth conditions for our epistemic language.

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 $for\, all\, worlds\, t\, such\, that\, R_a st,\ M, t \vDash \varphi$

We will focus on the prominent epistemic logic $S5$, which is the set of Kripke models where R is an equivalence relation. 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 is entails veridicality and is mostly uncontroversial: if you know something, it is true. The other two implies that if you're aware of what you know and do not know and are much less realistic, which we will get back to when looking at realistic social reasoning. $S5$ allows us to interpret R as an *indistinguishable* relation. It means that when agents consider worlds possible, they cannot distinguish between them. This means that we can view all worlds which are accessible for an agent as *epistemic alternatives*. 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 ..., we introduce the notion of *common knowledge*:

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

There is an intriguing but quite unintuitive difference between any finite number of iterations of the E-operator and the infinite iteration of common knowledge. This difference can be shown in the examples from epistemic logic called 'consecutive numbers' shown below. These examples are also very similarly structured to the canteen dilemma as we will see later.

2.0.1 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 their numbers will be 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]. This gives us the following diagram:

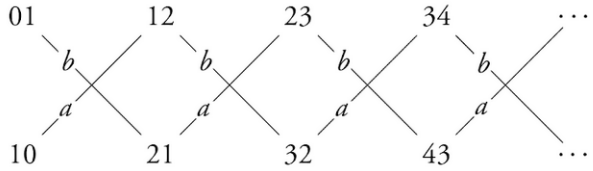

Figure 1. Consecutive numbers model (Ditmarsch & Kooi [20]).

Suppose the actual numbers given are 2 to Anne and 3 to Bill, denoted as a state $(2,3)$. In that case we can infer a few different facts. Most basically, Anne know 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 (if Bill has 1) or 2 or 4 possible (if Bill has 3), stated as: $K_a K_b(a_0 \vee a_2 \vee a_4)$. For the more interesting part, imagine that Anne and Bill has to guess whether they both have positive numbers, i.e. none of them have a 0, denoted as $M,(x,y) \models (positive)$ iff $(x \neq 0 \wedge y \neq 0)$.

If we ask both Anne and Bill in our supposed case if they know that the other's number is positive, they would both answer yes: $K_a(positive) \wedge K_b(positive)$, which can be expressed concatenated 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 respectively if they both know that no one has a 0 , i.e. if $E_{\{a,b\}}(positive)$ is true. Start with Bill. He considers $(2,3)$ and $(4,3)$ possible. Since we're only focused on 0, we can focus on the lowest number-pair, i.e. $(2,3)$, since the other direction only takes us away from 0. Bill thinks it's possible Anne has a 2, which means he knows both numbers are positive, and since Anne would then consider 1 as the lowest number, he knows that Anne knows it 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: $E_{\{a,b\}}(positive) \wedge \neg E_{\{a,b\}} E_{\{a,b\}}(positive)$. Another way to read this off the diagram, is to say that $(positive)$ is mutual knowledge in a world w if no one considers worlds with 0 possible. The E operator can then be iterated for each step an agent can take from their set of possible worlds towards a world with a 0.

The unintuitive aspect of this is that we can have k iterations of the E operator, but not $k+1$. That is, no finite iteration of mutual knowledge is equal to common knowledge. So in the consecutive number example,

[1]The following example is based on the example given in [20] and the more technical version in [19].

no matter what number pair Anne and Bill are given, it is never common knowledge that none of them have a 0! The practical implications of this can be exemplified by 'gamifying' the example. Imagine Anne and Bill have to make a binary choice between either both having positive numbers or remaining undecided, and they can only win by answering the same, without answering both having positive numbers if they do not. In our supposed case $(2, 3)$, both agents know they both have positive numbers. But since Anne has a 2, she is not sure that Bill knows. For all Anne knows, Bill might have a 1 in which case he would not know that both have positive numbers. For that reason Anne might not answer that no one has a 0, in which case Bill should answer the same, even though Bill has a 3. This is the case for any possible number pair for Anne and Bill. This is highly unintuitive however when the number of epistemic operators increase. If two real persons were given $(719, 720)$ and asked independently if they would agree none of them has 0, they would both be quite certain about their numbers being positive and probably hard to dissuade otherwise.

Agents reasoning about whether both have positive numbers in the consecutive numbers game is an individual process. When both have to reason about what the other will answer, coordination is required and it turns into social reasoning. It requires that each agent do not just consider their own possible worlds, but the possible worlds of the the other agent.

The canteen dilemma is structured very similarly to the consecutive numbers example and it shows this type of social cognition in action. Whenever we're faced with experimental facts concerning the differences between proscriptions from epistemic logic and how people actually think, we might be inclined to conclude that such results only show *limits* or *deficiencies* in human reasoning, following the anti-psychologistic tradition from Frege and seminal studies like [69]. This is partly true, since the validity of established logical systems does not change if someone makes different inferences. But when it comes to having a Theory of Mind and reasoning about the mental states of others, the salient feature must be the accuracy of such a representation, not how well it depicts what their mental model or reasoning ought to look like. If we want to interact intelligently with ambient technologies like self-driving cars and rescue-robots, they will also have to be able to reason about what humans think and as such they should be aware of the human cognitive limits when it comes to higher order social reasoning [21, 22, 24]. Non-human agents making erroneous assumptions about how people actually reason and basing their decisions on such assumptions is bound to lead to sub-optimal cooperation. That means that even if we stick to the traditional normative program of logic, when it comes to higher order social reasoning, the way that human or non-human agents *ought* to reason depends in part on how they *actually* reason. To rephrase Frege, if people are not to miss the truth, they sometimes have to consider how people actually think, and not just how they ought to think. This leads to looking into the insights about real higher order social reasoning gathered from behavioral experiments in cognitive science.

2.0.2 Byzantine generals problem [consider including]

Two generals unable to generate common knowledge through insecure communication (or delay).

3 Real higher order social reasoning

It is trivially accepted that people do not always reason perfectly and according to logical prescriptions. The Wason selection task [69, 70] was an early experiment showing difficulties with just propositional logic. It always involves four cards with members of a set A or B. In one treatment, subjects where shown sixteen cards with a letter on one side and a number on the other. Four of these, D, K, 3, 7 were used. Subjects 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 $p \to q$. 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 logical fallacy of affirming the consequent was part of the most popular answer (D and 3).

Such results seem to imply that humans are poor logical reasoners. But there are a few complexities. First off, Wason [70] that people were significantly better at answering correctly when the question was about cities and transportation devices, that is, *every time I go to Manchester I travel by car*. Griggs and Cox [36] 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 [64] argue that the abstract claim might be interpreted as merely checking satisfaction of instances instead of determining the truth of the rule. Wagner-Egger (Conditional reasoning and the Wason selection task: Biconditional interpretation instead of reasoning bias) argues that the 'error' made by most people may be due to interpreting the rule as a biconditional. The ambiguity of of the statement could also explain the results from Cheng et al. [15] which suggest 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. Traditional propositional logic also often sets a standard that is low enough that it can be met at least in specialized settings, like courtrooms or scientific journals, 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). When it comes to higher order social reasoning, modeled by our epistemic logic $S5$, it is another story. Before I go into empirical results on higher order social reasoning, let us pre-emptively look at some of the overly idealized aspects of $S5$.

3.0.1 Idealizations of $S5$

The properties of $S5$ imply that agents are agents know all logical truths. Since all logical tautologies are true in all possible worlds in an $S5$ model, every agent knows these as well. Transitivity ($K_a\varphi \to K_aK_a\varphi$)

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. Contrary to the cases of propositional logic above, there is no reason to believe such a strong idealization holds for human beings. These concerns leads some to say that if such properties are unacceptable for a given application, the possible worlds approach might not be the best option [19]. But following Frege, idealizations can still be used as guides towards truth, much like physicists might refer to 'spherical cows'.

Dynamic epistemic logics are often motivated by lying closer to real cognitive practices. Real knowledge is gathered in a dynamic social environment, where agents ought to update and change their knowledge and beliefs as they gain new information. Our epistemic logic can express this by adding action expressions as well as dynamic modalities for these. This means adding $[!\varphi]\varphi$, 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 facts 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. If agent a expresses the aciton formula $[!(p \wedge \neg K_b p)]$, b does not come to know both p and $\neg p$. When announcing non-epistemic propositions, agents *ought* to come to know it. But this rests 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 recall everything else they know. In fact, for an announcement to become common knowledge, perfect observation and recall must be common knowledge among agents as well. In reality publicly announcing non-epistemic facts to people (even assuming they're attentive) does not always entail them updating their knowledge with such facts. Statements can either be too complex or lengthy for agents to decipher, or the statements might be ambiguous, as some researchers argued 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. Liu [48] identifies this and other parameters for variation among agents which I will go through shortly before presenting more empirical facts regarding higher order social cognition.

3.0.2 Parameters for diverse cognitive capacities

Fenrong Liu identifies five novel parameters for diversity among epistemic agents [48, p. 25f].

(a) inferential/computational power: making all possible proof steps

(b) introspection: being able to view yourself in "meta-mode"

(c) observation: variety of agents' powers for observing current events,

(d) memory: agents may have different memory capacities, e.g., storing only the last k events observed, for some fixed k

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

Parameters (a) to (d) can be understood as logical norms describing limiting optimal 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. Sticking to one's beliefs come what may is hardly truth-conducing and neither is changing one's belief with every piece of conflicting information. 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 interact and accomplish difficult tasks. In other words, instead of noting what epistemic powers agents lack compared to perfection, we should focus on the epistemic powers agents do have. This includes effective heuristic measures which allows us to bypass certain computationally costly processes [35]. Like Liu mentions: "despite our differences and limitations, societies of agents like us manage to cooperate in highly successful ways!" [48, p. 26]. In other words, we should not just look at how far human reasoning is from upper bounds on a rationality spectrum but also positively in terms of how advanced and far it is from lower bounds of the spectrum. This means going into empirical data from experimental cognitive science and psychology concerning higher order social cognition.

3.1 The development and importance of Theory of Mind

The cognitive capacity of having a theory of mind is what allows humans to traverse and navigate our incredibly complex social world. It does this by enabling us to infer what other people might think, feel or desire. It means being able to attribute unobservable inner mental states to others based on their observable behavior. This in turn allows us to predict and understand their future behavior, as well as reason about what mental states they might attribute to us based on the behavior we exhibit. This is also why having a theory of mind is sometimes referred to as social cognition or intelligence, or perspective taking. It is argued that self-consciousness evolved exactly to enable primates to attribute mental states to others, as this mental attribution is essential for dealing with the trials of an increasingly complex social life [17, 42].

Studies show that children learn to distinguish their own beliefs from others between the ages of 3 and 5 [73], while children learn to make correct second-order attributions between age 6 to 8 [57]. Another study indicate that children acquire a capacity of theory of mind that outperforms our nearest primate relatives by age 2 [40]. I will return later to the discussion whether the use of theory of mind requires effortful attention or if it is applied more universal and automatically. Before I go into the human limitations within theory of mind, I will first briefly re-state why theory of mind is important and why it is important to understand the human limitations.

We often depend on successful higher order social reasoning in order to interact in meaningful social contexts. Realizing that other's might not know what you do can lead to better cooperation, negotiation and general understanding of others. As such, this insight into the inner mental workings of others might be an prevalent factor for having empathy for other people, although this normative aspect will be left out of this discussion.

As with everything else, people do not reason perfectly within higher order social cognition. Because of its complexity it is likely severely limited, in the sense that people can have problems at practically applying theory of mind considerations they might otherwise possess. Investigating these limitations are important for two reasons.

First, if our everyday theory of mind notions are flawed, we can only counteract and adjust for such shortcomings by being aware of them. In other words, it might allow us to learn both that we should not trust our common sense intuitions when it comes to higher order social cognition, and secondly, it allows us to learn better practices in cases of importance. Like Benthem writes, human cognition often manages to integrate formally designed practices (like games or puzzles) into our common sense behavior such that we can navigate and act within these structures without being attentively aware of the processing going into it. [**?**, p. 81f].

Secondly, even real applications of theory of mind which might deviate from certain logical proscriptions do not necessarily imply a limit or deficit in reasoning. After all, people cooperate successfully all the time without acting on perfect logical reasoning and rationality. The story of human social interaction is largely a success story and deviation from perfect logical practices does not change this. Furthermore, if we are to design artificial agents which has to understand and predict human actions, they ought not to assume that the humans reasons like themselves. In fact, as Benjamin Erb argues [24], when optimizing intelligent human-computer interaction it's 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 a human mental model and the usefulness of such a model hinges on how accurately the model represents the actual mental state of the human. 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 model also has to accurately represent the reasoning of the non-human and for humans to accomplish such a feat, the non-human might have to employ human-like reasoning.

"Knowledge of ToM and language use would be very useful in designing conversational agents, because if humans draw inferences differently, depending on the nature of the situation, artificial agents should also do so, and should be able to take into account that others may do so" [53].

"In intelligent, technical environments, humans may intrinsically apply ToM traits to their non-human

interaction partners. When designing humanoid agents, self-driving cars, or ambient technologies, it becomes important to factor in the human ToM. Once humans assign intentions, motives, or beliefs to their non-human counterparts, these thoughts inherently become part of their interaction space and must be considered carefully" [24]

3.2 The difficulty of applying Theory of mind

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 capped at first or second-order reasoning, indicating that people do not use what's called backwards induction. I will go through this first before presenting the limit relating to spontaneous versus reflective use of theory of mind, the so-called *curse of knowledge* and the task dependence problem.

While studies show children learn different levels of theory of mind at certain ages, it is never or rarely more than 2. 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 [27, 39, 67].

Having a higher order theory of mind might be difficult for some of the same reasons why backwards-induction reasoning is difficult. Take the consecutive numbers game for example and the question "are you certain that you will both answer that none of you have a 0?". A player will answer negatively if the player has a 0 themselves. If a player has a 1, they might also answer negatively, since the other person might have a 0. But even if a player has 2, they know none of them have a 0, but are still not sure that both will answer this, since the other player might have a 1, in which case they would not answer that both numbers were positive (if they followed the same reasoning as just before), in which case the first player shouldn't answer positively either. And this holds for all $n \in \mathbb{N}$by backwards induction. This is the same line of reasoning that dictates that one should always opt out of so-called centipede games at their first turn. In such games, two players takes turns either opting out of the game, taking a larger share than the opponent, or pass the choice to the other player, which leads to a increase in the total share. In such games people rarely choose the supposed rational strategy of opting out at the beginning of the game [61], and instead often reason by forward induction instead [32]. But these studies mostly show cognitive limits in thinking of limit cases, that is, thinking backwards from the end of a game or thinking about whether common knowledge can be achieved and not just a finite iterations of mutual knowledge. There are signs of other more severe limitations, indicating that even when having a first or second-order theory of mind, this is not always acted upon or applied properly.

3.2.1 Spontaneous versus reflective use of theory of mind

Experimental studies show that there is a relevant difference between having a theory of mind and actually using it. Keysar et al. [45] 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 relatively late developing theory of mind aspect of representing beliefs as separate from corresponding reality, that is, acknowledging that others might have a myriad of mutually exclusive beliefs. It suggests a difference between being able to utilize one's theory of mind in reflective tasks and utilizing it for guiding one's spontaneous actions.

Flobbe et al. [27] findings also show examples of children who could understand second-order reasoning in story tasks but were failing to properly perform second order reasoning in game tasks.

3.2.2 Curse of Knowledge (we are biased towards ascribing our own beliefs onto others, perhaps bc it is normally effective)

Birch & Bloom [11] show that when adults attribute beliefs to others, they are more likely to attribute their own beliefs they know to be true than to attribute false beliefs. 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. When subjects did not know where the violin had been moved, they where generally good at predicting that Vicky would look in the blue container. When they knew the violin was moved to the red container however, they assigned significantly higher probabilities to Vicky looking in the red container. That is, when attributing beliefs to others, it seems as if subjects to an extend simply modeled their mental state as their own. 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.

3.2.3 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 [68] lists a few possible explanations. Firstly, there might be a high processing cost associated with theory of mind, causing a failure in applying appropriate order of social cognition when the the processing cost becomes too high. Secondly, 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 [44].

Most importantly however might be the explanation that 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, one's theory of mind might be largely ignored until it becomes sufficiently apparent that it is worth the effort.

This is further supported by an argument in Keysar et al. [45] 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. In other words, most knowledge would be common knowledge. In that case, differentiating between one's own beliefs and the beliefs of others might not be very beneficial, especially if this activity would take up vital cognitive resources. Another related reason for failing to apply an existing theory of mind is that the the cost of failing to apply it might not make it worthwhile to apply it per default. 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 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 is a mental capacity at all. In other words, when people use their theory of mind to interpret the actions of others in terms of mental traits, there might not be much reasoning happening. When hearing an cyclist use their bell, people are likely to infer that the person doing it is impatient and want to make others aware of this. But this knowledge happens so spontaneously that it does not seem right to say that people hear the sound and then reason about what it might mean. Rather, the ringing bell *sounds* like someone is being impatient, that is, the behavior is understood directly as an example of impatience, rather than being processed as pure sound first. This explains why theory of mind can be used 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 general 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, conducted at the Center for Information and Bubble Studies at the University of Copenhagen.

4 Canteen Dilemma

The canteen dilemma is a game with structural affinity to the consecutive number example in section 2.0.1. It depicts a situation where two agents can have several iterations of mutual knowledge (everyone knows that ...) but no common knowledge (everyone knows that, everyone knows that ... etc). It is framed in a thematic story as studies show that this can significantly improve people's logical reasoning abilities [52, 70]. 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. Their preferences are ordered such that their prefer (1) going to the canteen together if both arrive before 9:00 am, (2) going to the office together at any time and (3) all other configurations, that is, discoordination or either player going to the canteen at 9:00 am or later. The game consists of a numbers of rounds where each player are given their own arrival time and have to decide between going to the canteen or the office. The payoff structure follow a logarithmic scoring rule, as 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 [62]. This required players estimating how certain they were that the other player made the same choice as them. Players were told the running results after each round, including the previous arrival times and choices of both them and the other player, including their current payoff standing.

The rules dictate that players should always choose office at 9:00 am and 9:10 as it is the only winning move. The rest of the game is not as trivial however. Assume player a and b are playing and a arrives at 8:50 am. Player a know that b arrived at 8:40 am or 9:00 am. If the b arrived at 9:00 am, a only winning move would be office. If b arrived at 8:40, they could choose to go to the canteen or the office, as long as they do the same. Player a knows they prefer going to the canteen but since one of the possible scenarios would require him to go to the office, let us assume player a chooses the office at 8:50 am. Now, assume a arrives at 8:40 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 am. This line of reasoning can be repeated for all possible arrival times, and as such, a might never choose to go to the canteen! If a player is to choose office at 8:40 or earlier, they need to not just consider possible arrival times for the other player (8:30 or 8:50) but also what choice the other player would make under such circumstances. Since these choices depend on what knowledge the other player has available, predicting an office choice necessitates having a theory of mind. Anyone can choose office randomly of course, but we would expect random choices to be normally distributed among arrival times, meaning that peaks of office choices are still signs of higher order social cognition.

Notice however the premise above where a assumes that b reasons like a. The premise does not only assume that a reasons in some prescribed way, but also that a is warranted in believing that b does as well. In other words, it does not just require that players are ideally rational, but that this is common knowledge as well. This means that regardless of what player a thinks b ought to do, a has to try and to figure out what b might actually do. It implies that without common knowledge about rationality, two rational players will will play suboptimally. So, choosing canteen does not seem to imply a lack of theory of mind in itself. This is where the certainty estimates can be valuable. Assume a logician plays the game and arrives at 8:40. By relying on higher order social reasoning, the logician might infer that the optimal choice would be going to the office. But if the logician is uncertain about what type of reasoner the other player is, the logician

[2]Calibrated is defined as: "a set of probabilistic predictions are *calibrated* if p percent of all predictions reported at probability p are true" [62].

might believe that the other player might play canteen and will do the same. The logicians uncertainty can be elicited in the certainty estimate however, which will also help offset some of the loss from possible discoordination. This makes it possible to distinguish between players who both play canteen, since those with higher order theory of mind considerations are likely to be less certain than others.

This also help deflect what Verbrugge [68] calls the danger of simplistic formal systems which posit fixed bounds on social cognition, that is, people can reason up to n order but not $n+1$. It is unlikely that natural social cognition is so neatly discontinuous. A group of first-order social reasoners might contain variation along all five parameters for cognitive resources described above, and as such some might be closer to second-order social reasoning than others. This leads us to look at some of the logical structure of the game and possible avenues for logical modeling. After this I will go through some of the main results and add supplementary discussion and results which were not included in the accompanying article.

4.1 Logical structure of the canteen dilemma

The canteen dilemma can be modeled much like the consecutive numbers game. The following diagram exhaustively represents the possible states that agents can find themselves in. A possible state (8:00, 8:10) denotes that agent a arrived at 8:00 and b arrived at 8:10.

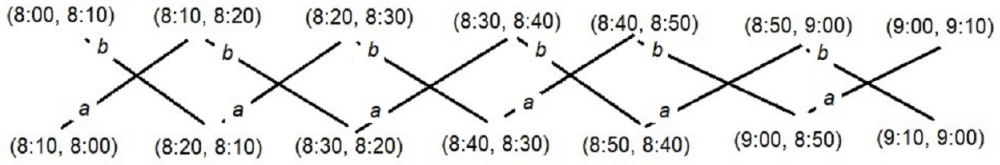

Figure 2: Epistemic model M_2 of the canteen dilemma

Figure 2 depicts an $S5$ epistemic modelM_2 following the same definitions as given earlier. Let us definite the proposition $(safe)$ as "it is not too late to go to the canteen", which is true in any state where 9:00 or 9:10 does not occur. M_2then shows the fact that in when a arrives at 8:30 and b at 8:40, both know that it is safe to go to to the canteen, but this epistemic fact is not known by both agents: M, (8:40, 8,30)$\models E_{\{a,b\}}(safe) \wedge \neg E_{\{a,b\}}E_{\{a,b\}}(safe)$. Agent b knows that everyone knows $(safe)$ but a does not. Agent b might think that knowing $(safe)$ is necessary for making a canteen choice. If that is the case, b would think that a would go to the office at 8:50, in which case b would go to the office at 8:40 too.This line of reasoning can again be iterated such that a and b would never go to the canteen. This is highly unintuitive of course, since in (8:00, 8:10), everyone knows that everyone knows that everyone knows that everyone knows that it is safe to go to the canteen, four iterations of the E modal operator. But since it does not hold for five iterations, there is no common knowledge that it is early enough to go to the canteen. Indeed, in the canteen dilemma, there is no such time that is early enough as to warrant the common knowledge that it is early enough to go to the canteen!

As has been the focus of previous sections, humans do not actually reason like this for several reasons, one of them being a limit in the capacity for higher-order theory of mind. Assume that a player's theory of mind makes them go to the office at 8:40 but the canteen at 8:30. One explanation for this would be that both choices were informed by their theory of mind, and that they simply went to the canteen at 8:30, since they believe that even if the other player arrives at 8:40, the other player would go to the office due to a lack of theory of mind. However, since we know that there is a cognitive limit when it comes to applying theory of mind, it is possible that there is a time t where a player thinks going to the canteen is unsafe due to their theory of mind, while the arrival time 10 minutes before that is deemed safe, since the theory of mind necessary to explain why it is risky simply is missing. That is, cognitive limits do not just imply that people play strategies involving canteen choices, they also imply that the inherent risk in any such strategy is out of epistemic reach for participants. We can now look at some of the results from the canteen dilemma. The accompanying article will contain the primary results and discussion, while the following sections will summarize and comment on supplementary results and considerations.

4.2 Results and discussion

4.2.1 Supplementary results

4.2.2 Improvements

4.2.3 Future research

5 Conclusion

References

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

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

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

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

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

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

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

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

[9] 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.)*

[10] 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

[11] 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

[12] 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.

[13] 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

[14] 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.

[15] 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

[16] 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

[17] 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.

[18] 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

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

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

[21] 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.*

[22] 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.*

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

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

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

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

[27] 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.)*

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

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

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

[31] 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.

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

[33] 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

[34] 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.

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

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

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

[38] 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.*

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

[40] 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.

[41] 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

[42] 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.

[43] 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.*

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

[45] 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.

[46] 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).

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

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

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

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

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

[52] 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.*

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

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

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

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

[57] 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.

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

[59] 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.

[60] 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.

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

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

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

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

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

[66] 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.

[67] 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.)

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

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

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

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

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

[73] 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.

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

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