The Canteen Dilemma - An Experiment on Higher Order Social Reasoning

Thomas S. Nicolet

April 22, 2019

Abstract

Successful social interaction often requires people to reason about the mental states of others, i.e. having a Theory of Mind. While the ability to make a mental model of other people's beliefs, desires or intentions is required in childhood, there is substantial research showing that even adults fail to reliably incorporate the ability into certain decisions. This type of social cognition can be modeled in Epistemic Logic, but even the most successful applications of Theory of Mind fall short of the proscriptions from Epistemic Logic. The Canteen Dilemma is a game which investigates how real higher order social cognition occurs. Our findings show that while certain behavior in the game does not indicate having a ToM, the certainty estimates for success provided by the participants does indicate Theory of Mind. This provides evidence for a finer grained picture than the idea that people can make k iterations of Theory of Mind but not $k+1$. This introduction to the accompanying article provides some background theory for the research field plus some additonal discussion regarding the experiment which was not excluded in the article.

This type of social cognition is often described as having a Theory of Mind of others. While there is robust evidence that children acquire the ability to reason about the mental states of others at a young age, there is also evidence suggesting that this ability is not effectively drawn upon in many cases. Even when successfully applied, his type of social cognition is not iterated more than once or twice, i.e. making a mental model others which might include their mental model of oneself. Higher order social cognition can be described more clearly in epistemic logic, but even the best applications of ToM amongst adults fail to live up to the proscriptions of Epistemic Logic. The Canteen Dilemma is an experiment which offers insight into how higher order social cognition occurs.

I introduce the Canteen Dilemma experiment and presents how it combines interests from cognitive science and logic by looking at a specific type of social cognition, namely higher order social reasoning. Social interaction often rely implicitly or explicitly on

higher order social reasoning - what do you think, that I think and so on. Such reasoning has been modeled in epistemic logics, although real reasoning rarely follow the proscriptions of these logics. I here introduce the Canteen Dilemma as an experiment which offers insight into how real people reason and discuss the implications for logic as viewed as a theory of rational agency.

Contents

1 Introduction (revise)

Successful social interaction often requires us to reason about the mental states of others. This type of social cognition is often called our Theory of Mind. Its the capacity to model the mental states of other, that is, being able to attribute to others otherwise non-observable mental states such as beliefs, intentions and knowledge. Being aware of how others might reason leads to better understanding and prediction of not just the behavior of others, but also how our own behavior might affect them. This reeasoning 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 this thesis. 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 [61].

Higher order social reasoning can be described in various epistemic logics and 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. #This will include supplementary discussion and possible improvements for future experimental research.

1.1 Logic and the facts - from Frege to Benthem (Consider removing fact-logic relationship entirely and focus more on cognitive science)

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. But none the less, logic has traditionally been thought of as independent of empirical facts and only relating to human reasoning as a normative force of how we ought to make inferences. The relationship between facts 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.

This cognitive turn is a step away from the anti-psychologism od 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" [28, 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" [27, 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 for Frege 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 is 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 consequently are they necessarily or contingently true? There is a rich historical discussion about this from Wittgenstein to Ayer and Carnap to Quine, Kripke and Putnam, 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 [45] or perhaps Quine [52], 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.

However 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 deduce that their reasoning is illogical or randomly structured. There might be a more complex logic underlying their decision, which can be reconstructed and used to more accurately represent how people actually reason. Of course this is not important when people make simple logical errors, but even in cases where people might not follow logical prescriptions precisely, they are still successful in their endeavors. It might imply that there are different non-random reasoning employed by people, and logic should be aware of this diversity if it wants to maintain 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 [43], 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 and 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 in turn

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 show such an idealized representation of reasoning about knowledge 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 after seeing its idealized picture of human reasoning, I will move onto real higher order social reasoning, and lastly the canteen dilemma.

2 Epistemic Logic, the 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 Backus-Naur form:

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

From this definition we can build formulas from atoms and more complex formulas using the negation $'\neg'$, conjunction $'\wedge'$ and knowledge operator $'K'_a$. I will use standard abbreviations like $(\varphi \vee \psi) = \neg(\neg\varphi \wedge \neg\psi)$ and $(\varphi \rightarrow \psi) = (\neg\varphi \vee \psi)$, and bi-implication $(\varphi \leftrightarrow \psi)$ is shorthand for the conjunction of implication both ways. 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 $R_a st$ 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's true. The other two implies that if you're aware of what you know and don't 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, that is, even if one of them is the actual one, they do not know which one they're in. 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 their possible worlds. 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 the examples from epistemic logic called 'consecutive numbers' which I will show below. These examples are also very similarly structured to the canteen dilemma as we will see later.

Plan: Use logic notation with E and C to describe consecutive numbers and byzantine generals. Then show how S5 is idealized and go into real social cognition. Show how S5 is idealized and doesn't describe the real world. Describe why ToM is important and why an accurate representation of it is important. Describe how humans are limited in ToM, even when we can reflect upon it, we don't necessarily act on it. Describe curse of knowledge, how knowing P leads us falsely conclude others know P as well. Describe problems with discrete bounds on reasoning (k but not k+1 ToM). Go in to canteen dilemma and its results. How do we model this in logic?

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:

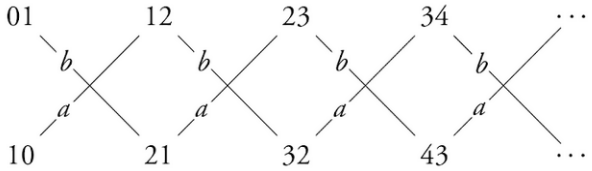

Figure 1. Consecutive numbers model (Ditmarsch & Kooi 2015)[19].

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

[1] The following example is based on the example given in [19] and the more technical version in [18].

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: $K_bK_a(positive)$. 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 $\neg K_aK_b(positive)$. 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, 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.

[Consider describing action formulas] [Consider arguing why having an accurate representation of the other's actual mental state and reasoning is more relevant than having a model of what the other ought to reason like]

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 [62]. 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 [20, 21, 23]. 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 [62, 63] 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 [63] 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 [33] 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 [57] 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 ($\neg K_a\varphi \to K_a\neg K_a\varphi$) 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 [18]. But following Frege, idealizations can still be used as guides towards truth, much like physicists might refer to 'spherical cows'.

Since real knowledge is gathered in a dynamic social environment, we might want to allow our epistemic logic to express agents updating and changing their knowledge and beliefs continually as they gain new information. This can be done by adding action expressions to our epistemic logic 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 the case for announcement of some epistemic facts . So-called "Moore-type" sentences are statements such as "p is true but you don't know it". If a publicly announces $[!(p \wedge \neg K_b p)]$, b cannot update her knowledge such that she both knows p and $\neg p$, and as such it is an unsuccessful update. But the statement that other public announcements leads to common knowledge is part of the idealized picture. Publicly announcing non-epistemic facts to real agents (even assuming they're attentive) does not always entail the agents 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, it might not be a limit in computational power but rather a limit in observational power. Liu [42] 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

Liu identifies five novel parameters for diversity among epistemic agents. These parameters can be understood as possible ways agents can deviate from logical norms, that is, for each parameter there is a limiting case which can be said to be rational. But as Liu writes, it would be better to eventually move away from understanding such variations in terms of 'limits' or 'bounds' and instead recognize how such diverse agents can accomplish difficult tasks. We have already seen the first two as idealizations in $S5$:

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

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

The next two came up as idealizations from public announcement logic.

(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.

The next describes the inherent dynamic nature of knowledge.

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

3.0.3 The importance of Theory of Mind

Successful social interaction often requires ToM. Furthermore, if we have to interact successfully with artificial agents, it's important that proper mental models are used. Studies show that children develop the ability to apply first order ToM in relation to a false belief task by the age of 4, while the ability to pass a similar task requiring second order ToM only develops a few years later. Studies with adults (Hedden and Zhang (2002)) suggest that they generally apply first order Theory of Mind successfully while their second-order Theory of Mind reasoning can also be seriously flawed.

Even when the use of ToM is severely limited, the study of its occurrence in real social reasoning is important for two reasons. First, if our standard intuitions concerning ToM are flawed, being made aware of possible shortcomings allows us to counteract and adjust for these shortcomings. This is important in the cases where the lack of ToM means foregoing pareto-efficient outcomes. 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. [7, p. 81f].

Secondly, even when real applications of ToM might deviate from certain logical proscriptions, it's not necessarily flawed. After all, people cooperate successfully all the time without acting on perfect logical reasoning and rationality, exactly because such behavior is rather robust. But if humans deviate from strictly rational behavior, artificial agent haves take this into account.

In fact, as Benjamin Erb argues [23], 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 on such a model hinges on how accurately the model represents the actual mental state of the human, which makes normative questions void for a moment. This also

holds the other way around turn, 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" [47].

"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"[23]

3.0.4 The limit of Theory of Mind use in adults - Spontaneous use vs reflective.

Even adults often don't apply sufficient ToM. However, there's evidence suggesting that this ability might not be effectively be drawn upon in spontaneous use cases.

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

3.0.6 Task dependence

ToM might not be a uniform type of social cognition. It's possibly task dependent, which could indicate that it's made up of different reasoning patterns. A possible explanation for this could be that theory of mind is more akin to a craft than a theory. A typical example is that we typically know how to ride a bike without having an explicit theory of how to do it. We can make a distinction between know-that and know-how. We can also make the inverse argument, that someone knows the theory of how to ride a bike, or play some sport, without being able to draw sufficiently on this knowledge in practice.

Consider (Wason 1971) to see that a thematic story might aid in logical reasoning much more than an abstract story. This is supported by (Meijering 2010), but this might have been counter productive in our case.

It's a possibility that employing ToM is a craft that's not sufficiently practiced among most people.

3.0.7 Problems of fixed bounds on social cognition.

3.0.8 Common knowledge about rationality

There's also a divergence about in games which could be explained by saying that rationality isn't common knowledge. If a player can't expect the other player to play rationally, she cannot be expected to play rationally either.

4 Canteen Dilemma

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] 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

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

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

[20] 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.*

[21] 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.*

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

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

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

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

[26] 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.)*

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

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

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

[30] 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.*

[31] 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

[32] 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.

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

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

[35] 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.*

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

[37] 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

[38] 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.*

[39] 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.

[40] 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).

[41] 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.

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

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

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

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

[46] 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.*

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

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

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

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

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

[52] 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.

[53] 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.

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

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

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

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

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

[59] 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.

[60] 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.)

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

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

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

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

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

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