Logic and Social Cognition

Article *in* Journal of Philosophical Logic · December 2009

DOI: 10.1007/s10992-009-9115-9

CITATIONS
37

READS
207

1 author:

Rineke Verbrugge
University of Groningen
203 PUBLICATIONS **1,410** CITATIONS

SEE PROFILE

J Philos Logic (2009) 38:649–680
DOI 10.1007/s10992-009-9115-9

Logic and Social Cognition

The Facts Matter, and so do Computational Models

Rineke Verbrugge

Received: 11 June 2009 / Accepted: 25 August 2009 / Published online: 10 October 2009

Abstract This article takes off from Johan van Benthem's ruminations on the interface between logic and cognitive science in his position paper "Logic and reasoning: Do the facts matter?". When trying to answer Van Benthem's question whether logic can be fruitfully combined with psychological experiments, this article focuses on a specific domain of reasoning, namely higher-order social cognition, including attributions such as "Bob knows that Alice knows that he wrote a novel under pseudonym". For intelligent interaction, it is important that the participants recursively model the mental states of other agents. Otherwise, an international negotiation may fail, even when it has potential for a win-win solution, and in a time-critical rescue mission, a software agent may depend on a teammate's action that never materializes. First a survey is presented of past and current research on higher-order social cognition, from the various viewpoints of logic, artificial intelligence, and psychology. Do people actually reason about each other's knowledge in the way proscribed by epistemic logic? And if not, how can logic and cognitive science productively work together to construct more realistic models of human reasoning about other minds? The paper ends with a delineation of possible avenues for future research, aiming to provide a better understanding of higher-order social reasoning. The methodology is based on a combination of experimental research, logic, computational cognitive models, and agent-based evolutionary models.

Keywords Epistemic logic · Cognitive science · Intelligent interaction · Cognitive modeling

R. Verbrugge (✉)
Institute of Artificial Intelligence, University of Groningen,
PO Box 407, 9700 AK Groningen, The Netherlands
e-mail: L.C.Verbrugge@rug.nl

1 Introduction

Most logicians are familiar with Johan van Benthem's ground-breaking contributions to modal logic, the logic of time and information dynamics, and to the fruitful exploration of the interfaces between logic and language and between logic and games. Less well-known is the fact that Johan van Benthem has also made worthwhile contributions to the interface between logic and cognitive science, as evidenced by [140, 141].

For many decades, logic and cognitive psychology were far apart and the only point of contact seemed to be a truck-load of papers describing psychological experiments purporting to show that people do not reason logically, most (in)famously [165]. These papers usually take a rather black-and-white view of logic, and they often do not describe the experimental task correctly in logical terms, as is poignantly described in Johan van Benthem's position paper "*Logic and reasoning: Do the facts matter?*":

> "Advertising 'mismatches' between inferential predictions of logical systems, usually without proper attention to the modelling phase, and what is observed in experiments with human subjects seems entirely the wrong focus to me—not to mention the fact that it is silly and boring. The much more interesting issue is to avail ourselves of what is involved in how people really reason."
>
> J.F.A.K. van Benthem [141]

This quote seems to suggest a negative stance towards psychological experimentation, but in fact the rest of Van Benthem's provocative piece does not at all deny the value of experiments—indeed, the facts do matter—but argues instead for more subtle distinctions and richer models of reasoning than have often been used in the past.

In recent years, many fruitful contacts along the lines suggested by Johan van Benthem have arisen between logic and cognitive science, witnessed by three recent special issues on this interface: a special issue of *Topoi* edited by Van Benthem et al. [143]; a special issue of *Studia Logica* edited by Leitgeb [84]; and a special issue of the *Journal of Logic, Language and Information* edited by Counihan [158].

The present article has been inspired by Johan van Benthem's ruminations in [141], especially by the question whether the facts matter for logic, and the question in which fruitful ways a combination of logic, psychology and computation could lead to plausible "formal models of real people". Our article focuses on just one type of reasoning, namely on social cognition, and more specifically on human higher-order reasoning about mental states of others. This type of reasoning is closely related to the core of the area that has come to fruition these last ten years at Johan van Benthem's initiative: the logic of intelligent interaction. This article is partly a survey on the state of the art on social reasoning from the combined viewpoints of cognitive science, logic and artificial intelligence, and partly a position paper on future research directions

for investigating questions about human social cognition by combining logic, behavioral experiments and computational models.

2 Intelligent Interaction and Higher-Order Social Cognition

As humans, we live in a remarkably complex social environment. One cognitive tool which helps us manage all this complexity is our *theory of mind*, our ability to reason about the mental states of others. By deducing what other people want, feel and think, we can predict how our actions will influence them, and how we should behave to be successful.

Thus, theory of mind is the cognitive capacity to understand and predict external behavior of others and oneself by attributing internal mental states, such as knowledge, beliefs, and intentions [109]. This is thought to be the pinnacle of social cognition [28].

Especially important in intelligent interaction is higher-order theory of mind, an agent's ability to model recursively mental states of other agents, including the other's model of the first agent's mental state, and so forth. More precisely, zero-order theory of mind concerns world facts, whereas $k+1$-order reasoning models k-order reasoning of the other agent or oneself. For example, "Bob knows that Alice knows that he wrote a novel under pseudonym" ($K_{Bob}K_{Alice}p$) is a second-order attribution.[1]

There has been an ongoing debate among philosophers and cognitive scientists whether our everyday understanding of mental states of others constitutes a *theory*, as claimed by the 'theory-theorists', or rather that, in order to understand and predict behavior of others, we directly *simulate* their mental states, as claimed by 'simulation theorists' [70, 71, 98]. Henceforth, I often use the term 'higher-order social cognition' in the sense of 'higher-order theory of mind'. The reason to do this is that in the controversy between 'theory-theory' and 'simulation theory', the term 'theory of mind' carries the unwanted connotation that 'theory-theory' is preferred.

2.1 Three Levels of Analysis

To delineate the perspective of this paper, it may be fruitful to keep in mind the three levels of inquiry for cognitive science that David Marr characterized [90]:

1. identification of the information-processing task as an input–output function: the computational level;
2. speciation of an algorithm which computes that function: the algorithmic level;

[1] Hence, 'higher-order' refers to a different phenomenon than higher-order logic (allowing quantification over sets and individuals). The orders roughly correspond to the modal depth of a formula.

3. physical / neural implementation of the algorithm specied: the implementation level.

Researchers aiming to answer the question what logical theories may contribute to the study of higher-order social cognition could be disappointed when it turns out that logic is not the best vehicle to describe higher-order theory of mind at the implementation level or the algorithmic level. For example, it may turn out that the simulation theory more closely describes people's actual reasoning than the theory-theory does. Still, logic surely makes a substantial contribution at Marr's first computational level by providing a precise specification language for cognitive processes; examples of this role of logic are sprinkled throughout in this paper. Quite possibly, logic may also have a fruitful role to play in theories of higher-order social cognition at the algorithmic level, in the construction of computational cognitive models.[2]

Let us first see how higher-order social cognition has started to be relevant for artificial intelligence and, in particular, multi-agent systems.

3 General Background: Artificial Intelligence and Multi-Agent Systems

In the proposal for the famous study at Dartmouth marking the birth of *Artificial Intelligence* (AI) in 1956, John McCarthy coined the term 'Artificial Intelligence' as the scientific discipline that is concerned with "making machines behave in ways that would be intelligent if a human were so behaving" [92]. He added that "the study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [92]. In the fifty years since 1956, the scope of research has broadened considerably to study natural intelligence both in people and in animals. Although a universally accepted definition of AI is still lacking, researchers have held true to the aim of seeking to understand and implement aspects of intelligence.

Currently, AI is often conceptualized in terms of building *agents* [118]. According to a widely accepted definition, "an *agent* is a computer system, situated in some environment, that is capable of flexible autonomous action in order to meet its design objectives" [74]. Their situated-ness distinguishes agents from classical AI expert systems that needed a user as intermediary of the information flow from and to their environment. Agents take initiatives without human interference and interact with other agents in order to further their own goals or those of others. The abstract concept of the agent has grown to provide fruitful tools and techniques for engineering complex computational systems [74].

[2] It is reassuring that in the slightly different context of closed-world reasoning, Stenning and Van Lambalgen managed to describe a reasoning task at all three levels: the information-processing task of credulously incorporating new information was shown to correspond on the algorithmic level to logic programming, which was then implemented in neural networks [129].

Multi-agent systems consist of dynamically cooperating computational systems, engineered to solve complex problems that require expertise and capabilities beyond the individual components [171]. Multi-agent systems have found application in complex situations that require multiple types of expertise, perspectives, and methods, such as air-traffic control [47], car manufacturing-line control [73], and negotiation [63, 86, 114]. Multi-agent systems display different types of complex interaction, from outright competition, through coordination and negotiation, to full cooperation.

In the last decades, research on multi-agent interaction has focused on idealized software agents with unlimited computational resources and perfect logical reasoning powers. Nowadays, interaction between humans and computer systems becomes vital for applications such as mixed rescue teams after earthquakes: a number of robots descend under the rubble to find victims, humans then rescue these victims, while software agents continually evaluate and re-plan the collaborative rescue action. In order to coordinate such complex teamwork, it is vital that team members understand one another. Therefore, it is high time to make more realistic models of intelligent interaction in mixed human-computer teams.

Fortunately, investigations into cooperative interactions in the behavioral sciences, logic and computer science show a marked convergence: after all, people cooperate, complex software systems cooperate [35, 85], and mixed teams consisting of software agents, robots and people cooperate, sometimes even better than people and computational systems separately [122, 132].

In order to develop intelligent systems in interaction, it turns out fruitful to specify agents in terms of their mental states, representing knowledge, beliefs, goals, intentions and plans, as well as recursively representing mental states of others [112]. As an illustration, let us turn to epistemic logic, the particularly elegant modal logic of knowledge.

4 From the Logical Point of View: Reasoning About Knowledge

As is probably familiar to the readers of this journal, epistemic logic was first introduced by Von Wright as a bare axiom system without semantics [164], with axioms such as $K_i\varphi \rightarrow \varphi$ (if agent i knows φ, then φ is true). The subject started to flourish after the invention of possible worlds semantics [68, 82]. As a reminder, one can view worlds that are possible or *accessible* for a certain agent i in world w as epistemic alternatives, worlds that are compatible with agent i's information in w. In general an agent i is said to *know* a formula φ in a world w_1 in model $\mathcal{M}$ (notation $(\mathcal{M}, w_1) \models K_i\varphi$), if and only if φ holds in all worlds w_2 that are accessible for i from w_1 (notation $(\mathcal{M}, w_2) \models \varphi$).

Group notions such as *common knowledge* are essential for formalizing intelligent interaction in multi-agent systems. Intuitively, φ is common knowledge in a group if everyone knows that φ, everyone knows that everyone knows that φ, and so on, ad infinitum. Semantically speaking, a proposition φ is *common knowledge* among group G in world w (notation $(\mathcal{M}, w) \models C_G\varphi$)

if and only if for all worlds w' in the transitive closure of the union of the accessibility relations for all agents in G, φ holds in w' $((\mathcal{M}, w') \models \varphi)$.[3]

4.1 Logic and Life: Children as Epistemic Logicians in the Science Museum

In a recent Kids Lecture about logic for children from the age of about eight in the Science Museum in Amsterdam, Johan van Benthem called three young volunteers to the front, let us call them Ann, Bob, and Carol. They received one card each from the set {red, white, blue}.[4]

They could all see their own card, but not those of the others. A possible worlds model of this situation is represented in Model I of Fig. 1a. Each world represents a card distribution in alphabetical order of the agents, with obvious color abbreviations. For example, *wbr* represents the state in which Ann has white, Bob has blue and Carol has red, corresponding to the real card deal in Van Benthem's Kids Lecture. We introduce propositional atoms such as w_{Ann} for "Ann holds white", which is true in *wbr* and *wrb* but not in the other four worlds. Because on the basis of her information, Ann cannot distinguish the factual situation *wbr* from the world *wrb* in which her colleagues have been dealt oppositely, an accessibility relation for Ann is drawn between *wbr* and *wrb* in the picture. Reflexive relations are assumed but have been left out of the picture. Let $\mathcal{M}_I$ be the name for the model as a whole, including the worlds, accessibility relations and the valuation of propositional variables.[5]

We find that in *wbr*, Ann doesn't know that Bob has the blue card $((\mathcal{M}_I, wbr) \models \neg K_{Ann} b_{Bob})$ because in at least one accessible world for Ann, namely *wrb*, Bob does not have blue. More complex propositions can also be seen to hold in the world *wbr*, such as "Bob knows that Ann does not know that Bob has blue" $((\mathcal{M}_I, wbr) \models K_{Bob} \neg K_{Ann} b_{Bob})$. This is because in all worlds accessible for Bob from *wbr*, namely both *wbr* itself and *rbw*, we can see again that Ann doesn't know that Bob has blue.

Under the assumption usual in epistemic logic that all participants have common knowledge of their being perfect logical reasoners, Model I in Fig. 1a applies; of course they do not know which is the real world, being able to see only their own card. Indeed when asked by Van Benthem, all three children said they did not know the cards of the others.

In the formal model, it is easy to see what ideally holds. In situation *wbr* it is common knowledge among Ann, Bob and Carol that Ann doesn't know that Bob has blue:

$$(\mathcal{M}_I, wbr) \models C_{\{Ann,Bob,Carol\}} \neg K_{Ann} b_{Bob}$$

This is because all six worlds can be reached from *wbr* in one or more steps by accessibility relations for agents in the group, and in each of those six worlds,

[3] For epistemic treatments of common knowledge, see [14, 50, 154].

[4] The experiment was inspired by a running example from [155].

[5] Precise mathematical representations of these notions can be found in introductions to epistemic logic [50, 145].

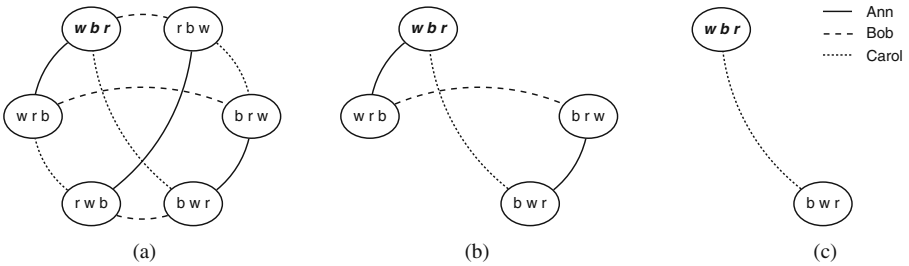

Fig. 1 Possible worlds models for the card-guessing experiment. **a** Model I. **b** Model II. **c** Model III

it is clear that $\neg K_{Ann} b_{Bob}$ holds there because from each world, Ann can access at least one world in which Bob doesn't have blue.

Through communication, agents gain knowledge, by which possible worlds models shrink. How did this work out in the Kids Lecture? Ann was allowed one question and asked Bob "Do you have the red card?". Johan van Benthem then asked, before the answer to Ann's question was given, if they now knew the cards of the others and Bob correctly said he did. Apparently he used a correct first-order knowledge attribution like "Because she asked me this question, clearly *Ann did not know* that I do not have red, so she does not have red herself. Therefore Ann must have white, and Carol has blue".[6] Formally, under the 'perfect reasoners' assumption, the possible worlds model after Ann's question is Model II of Fig. 1b, where *rbw* and *rwb*, exactly those worlds in which Ann has red, have been deleted from Model I.[7]

Now in the lecture, Bob answered Ann's question in the negative: "No, I do not have red". Johan van Benthem asked the three children again who knew the cards, and now Ann also raised her hand. Apparently she used a correct zero-order argument such as "if Bob does not have red, he has blue and Carol has red".[8] Also, both other kids understood that Carol did not know the cards, a correct first-order attribution.

Still under the 'perfect reasoners' assumption, the possible worlds model after Bob's public announcement results from Model II of Fig. 1b, from which *wrb* and *brw*, both worlds in which Bob has red, have been deleted. In the ensuing Model III, both Ann and Bob know the card distribution, whereas Carol still does not $((\mathcal{M}_{III}, wbr) \models K_{Ann}(w_{Ann} \wedge b_{Bob} \wedge r_{Carol}) \wedge K_{Bob}(w_{Ann} \wedge b_{Bob} \wedge r_{Carol}) \wedge \neg K_{Carol}(w_{Ann} \wedge b_{Bob} \wedge r_{Carol}))$. Indeed the only uncertainty left

[6]Note that some pragmatic reasoning about informativity is used here, as well as the assumption that it is common knowledge that the kids are cooperative and do not try to deceive each other. A very interesting recent take on such pragmatic reasoning, based on default logic, can be found in [17].

[7]For treatments of updating possible worlds models after public announcements, see [12, 156].

[8]In fact, Ann could already have come to the same conclusion at the previous step just after Bob's admission of ignorance, by way of the first-order attribution "If Bob had red, he would *not know* that I have white; so Bob has blue".

is for Carol, as witnessed by her accessibility relation between the two left-over worlds.

Colleagues from the department of psychology had warned Johan van Benthem that the children might not be able to solve the puzzle and understand each other's knowledge or lack of knowledge of the cards, but in my view the above scenario only requires correct first-order reasoning, combined with propositional logic, which children should definitely be able to do rather well by age 8.[9] It would be more impressive if the children in the audience, without seeing any of the cards, had reasoned after Bob's negative answer to Ann ("No, I do not have red") as follows: "Now Bob knows that Carol doesn't know his card" ($K_{Bob}(\neg K_{Carol} w_{Bob} \wedge \neg K_{Carol} b_{Bob})$), a second-order attribution, which 8 to 10 year olds often have great difficulty to apply in game situations [51].

Although in the Kids Lecture, the three participants were bright enough to provide the correct answers (for which first-order theory of mind was sufficient), nevertheless in general, dynamic epistemic logic may paint an overly idealized picture of human behavior.

5 Problems with Epistemic Logic as a Model for Human Social Cognition

In the field of epistemic logic, unlimited rationality is mistakenly taken for granted. Agents are assumed to be *logically omniscient*: they know all logical truths. Logical omniscience is a consequence of using possible worlds semantics [15, 126]. Logical truths hold in all possible worlds, so by the semantical definition of knowledge, all agents know them: If $\models \varphi$, then $\models K_i\varphi$. This is clearly not true for ordinary people, who do not know extremely complicated logical truths.

Moreover, epistemic logic assumes that agents have positive and negative introspection into their own knowledge: $K_i\varphi \rightarrow K_iK_i\varphi$ and $\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$ hold in the standard system **S5**.[10] In the second half of the twentieth century, however, cognitive scientists started to study phenomena like implicit cognition. Experimental subjects could correctly recognize well-formed strings of abstract languages by learning from examples, without managing to formulate the complex underlying rule [87, 169]. Hence, humans are often not aware of their own knowledge and beliefs.

Finally, the epistemic language allows reasoning on any modal depth and presupposes that agents can immediately decide whether a formula like

[9] Extracting from the story, Ann only needs zero-order reasoning, from "I have white" and "Bob does not have red" to "Bob has red and Carol has blue". Bob needs only one first-order attribution: From "I have blue" (zero-order) and "Ann did not know that I don't have red" (first-order), he concludes "Ann does not have red" (zero-order), and therefore "Ann has white and Carol has red" (again zero-order).

[10] Corresponding to the transitivity and euclideanness of accessibility relations, respectively; the veracity axiom $K_i\varphi \rightarrow \varphi$ corresponds to reflexivity [139].

$K_{Ann}\neg K_{Bob}K_{Ann}K_{Carol}\neg K_{Ann}\, w_{Ann}$ is true in wbr.[11] Common knowledge has an infinitary flavor, which makes it impossible to establish by communication in an untrustworthy communication medium, like the Internet [50, 62, 131].

The above three unrealistic properties of epistemic logic (logical omniscience, reflective powers and unlimited recursive knowledge attribution) occur in every logic that is based on possible worlds semantics. Therefore, the belief-desire-intention (BDI) logics, popular in multi-agent systems [112], are not sufficiently flexible to take the next step scaling up to mixed teams in which human participants cooperate with computational ones [132]. Nowadays, it becomes of paramount importance that software agents learn to reason about their human colleagues' cognitive limits with respect to higher-order social reasoning and their propensity to make mistakes [36, 41].

If one wants to make more realistic models of intelligent interaction in mixed human-computer teams, a fruitful avenue may be to computationally model human capacities to reason about mental states of other agents. So, after having explained ideal logical social abilities, let us have a look at results from psychology about social cognition in the context of natural intelligence.

6 Natural Intelligence in Interaction: Theory of Mind

Theory of mind, the capacity to reason about mental states of others, seems to be a characteristically human capacity. As for its *development*, Wimmer and Perner [170] showed that children between three and five years of age learn to distinguish their own beliefs from those of others, which may be false. In the famous 'false-belief task', children have to report their own belief about another child's mistaken belief (cf. [8, 33, 58, 100, 124, 166]). Interestingly, children model goals of others earlier than their beliefs [58, 120, 166]. Between around 6 and 8 years old, children learn to make correct second-order attributions [106].

Let us give an example of a second-order false-belief task, as reported in [51]. The participants heard a second-order false belief story, the Chocolate Bar story (see below), accompanied by drawings.[12] Afterwards, the participants answer several questions, modelled after [136]. The questions test different aspects of the participant's understanding of the story, among which their ability to correctly ascribe a second-order false belief such as "Mary believes that John believes that the chocolate is in the drawer".

In the Chocolate Bar Story, John and Mary are in the living room when their mother returns home with a chocolate bar that she bought. Mother gives the chocolate to John, who puts it into the drawer. After John has left the room, Mary hides the chocolate in the toy chest. But John accidentally sees Mary putting the chocolate into the toy chest. Crucially, Mary does not see John. When John returns to the living room, he wants to get his chocolate.

[11] It happens to be true.

[12] The Chocolate Bar Story is a second-order adaptation of a first-order story by [69].

Questions asked to the participants are: Where is the chocolate now? (reality control question), Does John know that Mary has hidden the chocolate in the toy chest? (first-order ignorance question), Does Mary know that John saw her hide the chocolate? (linguistic control question), Where does Mary think that John will look for the chocolate? (second-order false belief question), and Why does she think that? (justification question). If the children are not able to correctly attribute second-order false beliefs but otherwise are linguistically competent, they are predicted to answer the reality control question, the first-order ignorance question and the linguistic control question correctly, but give incorrect responses to the second-order false belief question and the justification question.

In [51], it turns out that a vast majority of tested children of 7 and 8 years old are able to correctly ascribe second-order attributions in the second-order false-belief task.

6.1 Adults and Higher-Order Reasoning in Games

Experimental research shows that most adult game-players exhibit first-order reasoning and many give a passable shot at second-order reasoning, whereas higher orders are rare [51, 64, 93, 163]. In classical game theory, it is often assumed that there is common knowledge that players are rational and capable of perfect reasoning. Game theory has often been criticized for the assumption of perfect reasoning. In fact, many experimental studies have shown that people do not always follow rational strategies, for example, they do not do so in so-called centipede games (introduced in [115]). Let us remind the reader of these games of perfect information, using the small example of Fig. 2.

In this centipede game, two players take turns choosing either to take a larger share of the current amount of marbles (down in the picture), thereby ending the game; or alternatively, to pass the choice to the other player (right in the picture), which leads to an increase of the total available amount of marbles. In the picture the pay-offs for each player are represented at the leaves, and the amount of marbles at the start of the game is one. If player 1 starts choosing down, he receives one marble while player 2 receives nothing; if, on the other hand, player 1 chooses right, the available amount increases. At

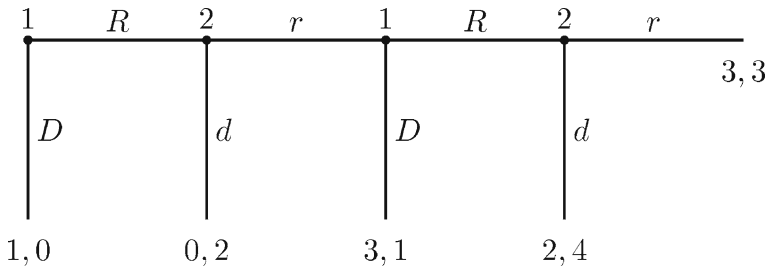

Fig. 2 Game tree for the centipede game

the next turn, if player 2 chooses down, she receives two marbles while player 1 receives nothing, and so on.

Now we can demonstrate the relevance of common knowledge of rationality by using backward induction (see also [145, 163]). Let r_1 denote that player 1 is rational, and r_2 that player 2 is rational. At the fourth and last choice point from the left, using r_2, we can infer that 2 will prefer four marbles to three ones and she will choose 'down'. Since rationality of both players is common knowledge, we know that $K_1 r_2$, hence, when in the third choice point, player 1 knows that player 2 will choose 'down' in the fourth choice point, and hence, since player 1 is rational and prefers three marbles over two ones, he will choose 'Down'. Since $K_2 K_1 r_2 \wedge K_2 r_1$, if player 2 were to reach the second choice point from the left, she would apply the same reasoning that we just did and conclude that player 1 will play 'Down' in the third choice point, so player 2, being rational, will play 'down' in the second choice point. Continuing this line of reasoning, and using the fact that $K_1 K_2 K_1 r_2 \wedge K_1 K_2 r_1$, we can conclude that player 1 will play 'Down' at the start. Therefore, if there were common knowledge of rationality, then backward induction could be used by the first player in a centipede game to conclude that he should immediately opt for the first dead-end and stop the game. The second player uses second-order social cognition to decide this, and we as observers use third-order social cognition to reach this conclusion.[13]

However, consider the following interesting alternative argument suggested by Ram Ramanujam. Just like the induction rule for epistemic logic ("from $\varphi \to E_G \varphi$, infer $\varphi \to C_G \varphi$"), one could formulate an induction principle for extensive form games presented as finite trees. For a two-player extensive form game g, the general induction principle would come down to the following, which is formalizable in a modal logic over trees:

> From "whenever both i and his opponent make a rational choice in each strict subgame of g, player i makes a rational choice in g" (for $i \in \{1, 2\}$), infer that "both players play rationally in g".

This principle is formalizable in any modal logic over finite trees. Both players follow the same principle, which is common knowledge to the players. Hence, both arrive at the same analysis as provided by backward induction. Thus, both players only apply first-order reasoning, and the outside observer only needs to apply a second-order inference predicting rational play!

It seems paradoxical that reasoning on the basis of common knowledge of rationality leads to a less than optimal outcome for both players. In [157],

[13] This reasoning applies to any n-player extensive form turn-based game presented as a finite tree. One level of knowledge is needed for a player reasoning at a node, the second level for the agent who is playing at the child node, and the third level for the outside observer's reasoning. The same applies to the subsequent argument based on the "induction principle" (as observed by Ram Ramanujam, personal communication).

Van Benthem and Van Eijck present a Platonic dialogue on *Game theory, logic, and rational choice* that offers an intriguing and sophisticated logical analysis of backward induction and its rival strategies.

Indeed, empirical research has shown that instead of immediately taking the 'down' option, players often show partial cooperation, moving right for several moves before eventually choosing to take the down option [93, 97]. Nagel and Tang suggest as possible reason for this deviation from the game-theoretic outcome that players sometimes have reason to believe that their opponent could be an altruist who always cooperates (by moving to the right).[14] In such a case, it is better to join the altruist in going to the right and then defect on the last round.

These empirical observations show that the premise of the induction principle above is questionable. After all, when a player is unsure that the other would make rational choices in a subgame of g, she cannot be expected to make a rational choice in g either. It seems that real-world players do not have common knowledge of a proposition such as "the other player is rational", but rather they get by with at most common beliefs about the procedures they employ.

Another possible explanation of human behavior in the centipede game involves error and cognitive limits: if the opponent has not correctly performed the full backward induction, it may be advantageous to cooperate in the first rounds. On this line of cognitive limits, Hedden and Zhang have argued that in reality, players hardly use backward induction at all and make use mostly of first-order, and only sometimes of second-order social cognition [64]. Hedden and Zhang used a game of perfect information very similar to the centipede game and concluded that adult subjects would start using at most first-order theory of mind. Gradually, a number of subjects would shift to second-order theory of mind when they started modeling their opponent as a first-order reasoner [64] (but see [51] for an alternative interpretation and experiments).

A difficulty with this kind of behavioral experiments is that it is very hard to be sure which kind of reasoning lies behind a player's choice in the game: does a player use backward induction or higher-order theory of mind? Or is he maybe working in parallel by concurrently analyzing the beginning and the possible endings of the game? A player may also adopt a strategy that is not couched in terms of others' mental states at all. For example, he can describe the game simply as "each step at which a player continues to the right, the pot is augmented; and each move, the player whose turn it is can terminate and collect the pot". He can then base his strategy on this specification: "Continue as long as you guess the other player will cooperate, and when you guess the other will not cooperate anymore, terminate". Such alternative models call into question what role logic can play at Marr's algorithmic level (see

[14] Note that in the induction principle above,"rational" could be replaced by "altruistic" to get a different pattern of reasoning (also suggested by Ram Ramanujam, personal communication).

Section 2.1).[15] Even asking a subject what reasoning he used may not be sufficient to be sure what's going on in his mind. Thus, experiments need to be designed very carefully.

6.2 Experiments: Understanding Theory of Mind Versus Applying It

A pilot study with children and adults turned up some surprising results for a game task [51]. Flobbe adapted Hedden and Zhang's centipede game experiments from [64]. Conversely to the earlier results, adults were mostly able to use correct second-order reasoning from the start and profitably adapted their strategy to their predictions about the other. Children between seven and eight years, however, acted much more variably. Many were still in the process of learning to apply second-order attributions, although they could already understand second-order reasoning in story tasks [51].

This points to a crucial gap between children's reflective understanding of theory of mind and its application in tasks such as games; a gap also shown by adults for first-order and second-order social cognition [78, 163]. Keysar et al. [78] report on experimental situations in which a speaker uses a term that could in principle refer to two objects known to the experimental subject, but only to one object for the speaker, as the latter is unaware of the existence of the second object, and this unawareness is clear to the experimental subject. The adult subjects nevertheless often perform as if the speaker referred to the object that is hidden from him, thus giving precedence to their own perspective rather than employing first-order social cognition.

The task-dependence of successful application of social cognition allows several explanations, all of which have implications for the nature of higher-order social cognition. A first, and very likely, possibility is that there is a processing cost associated with theory of mind, which causes a failure in applying the required order of social cognition when the processing demands of the task are high. Another explanation (not incompatible with the first) is that (higher-order) social cognition does not necessarily transfer from one domain of application to another. The ability to understand another's beliefs and intentions of a certain order may be present in principle, but to apply social cognition of the appropriate order, an individual must at least recognize that, in a given situation, it is to his advantage that this knowledge be incorporated in his decisions or actions.

In addition, higher-order social cognition may not be readily transferable from one domain to another until after a developmental process has taken place that makes this mental ability accessible to other domains, for instance Representational Redescription as proposed by Karmiloff-Smith [76]. Taking this reasoning one step further, it is even possible that what we call theory of mind is not one uniform mental ability to be drawn upon whenever the situation calls for it, but rather that different applications of social cognition

[15] This particular alternative was suggested by Keith Stenning (personal communication).

constitute different kinds of mental ability. These are all avenues of thinking about the nature of theory of mind that the scientific community may want to explore, however, their exploration is relevant only if first it is established to which extent there is task-dependence.

It is against this background that we placed the investigations presented in [51].We compared two groups of subjects, 8 to 10-year-old children and adults, on three measures. The first is a standard second-order false belief task, comparable to Tager-Flusberg and Sullivan [136]. The second is a strategic game, an adaptation of Hedden and Zhang [64], in which subjects play against a computer, trying to maximize their reward. The third measure is a linguistic task, which involves a linguistic phenomenon which is known to be acquired by children quite late, often after the age of ten: interpreting indefinite subjects of existential sentences such as "Er ging twee keer een meisje van de glijbaan af" versus "Een meisje ging twee keer van de glijbaan af". These are typical Dutch constructions that have as rough translations "Twice a girl went from the slide" and "A girl went twice from the slide".

Children's application of second-order social cognition was found to be highly dependent on the task to be carried out and the domain of application. Whereas almost all children succeeded on a verbal second-order false belief task, children's success rate in our second-order strategic game was only 57.2 %. With respect to the sentence comprehension task, only 40 % gave a bidirectionally optimal interpretation of the indefinite subject of an existential sentence. Thus, we have found that second-order social cognition is more difficult to apply than first-order social cognition, for children as well as adults, and that this is a pattern that does not only hold for verbal false-belief tasks, but also for a strategic game. Moreover, we have also found that successful application of second-order social cognition depends crucially on the domain in which it must be applied. This finding shows that, beyond the question of how human beings come to have a theory of mind, there looms another important question: How do we learn to use it?

Because humans do and other animals don't display higher-order social cognition, apparently somewhere during evolution hominids have acquired this capacity. It is important to investigate why and how higher-order social cognition evolved and which environments foster it. But we have only an approximate notion of the behavior and mental states of our ancestors. Evolutionary anthropologists have argued that change of environment led to a larger group size for hominids in pre-history. These groups necessitated new ways of bonding and establishing hierarchies. This social complexity was in turn correlated with higher relative brain-size and neocortex size, so that higher levels of social cognition emerged [2, 37, 38, 46, 52, 81].

Still, controversy about cognitive evolution remains. Some authors claim that higher-order social cognition arose because of the need for cooperative planning [57], others that it provided social glue by enabling gossip and language [38, 123]. Still others maintain that the main purpose of higher-order social cognition was to manipulate and deceive competitors, the so-called 'Machiavellian Intelligence' hypothesis [22, 130, 168].

7 Problems with Accounts of Natural Social Cognition

Human failures to apply higher-order social cognition are worrisome, because correct higher-order reasoning often spells the difference between failure and success in today's complex society. For example, Begin did far better than Sadat in the 1978 Camp David negotiations, partly because Sadat had written a letter to mediator Carter detailing his fallback positions on all major issues, but then refrained from drawing an important higher-order conclusion: "Begin may know that Carter knows my fall-back position" [99].

Everyone uses higher-order social cognition to negotiate, cooperate and compete. Still, the important question how higher-order social cognition works, how it is learned, how it has evolved, how it sometimes fails, remains largely an enigma. Hence it is still impossible to design effective interaction in mixed multi-agent teams including human participants. If software agents work together with human teammates, it is very important that they take into account the limits of social cognition of their human counterparts. Otherwise an international negotiation, for example, fails, even when it has potential for a win-win solution. In a time-critical rescue mission, a software agent may depend on a human teammate's action that never occurs.

Unfortunately, behavioral and neuro-psychological research on human *higher-order* social cognition is still scarce, in contrast to the wealth of research on first-order theory of mind [30, 31, 55, 56, 79, 100, 120, 124]. The 'higher-order' literature only investigates second-order social cognition [51, 89, 107, 163]. Group attitudes such as common belief have been investigated implicitly, but their complexity has been ignored. For example, Mant and Perner [89] asked children to judge the moral responsibility of the father in two versions of a story. His child was disappointed when he changed his previously communicated plan to go swimming. In one version, both had mutually agreed to go swimming, in the other version there was no agreement. Contrary to adults, children younger than nine years judge the father in the no-agreement version harshly [106]: "He said he would go, so he should have gone".

Strangely enough, Perner's analysis by second-order belief attributions [106] does not explain why children understand the difference between the two versions only around the age of ten, much later than they understand second-order beliefs. The concept of social commitment, as defined in [40], illuminates Mant and Perner's results. If an agent i *socially commits* to another agent j to do action α, then the first agent *intends* to do so. Moreover, the second one is *interested* in this intention. Finally, the agents have a *common belief* ("we believe that we believe that we believe...etc.") about these individual attitudes [40]:

$$\begin{aligned}\mathsf{COMM}(i,j,\alpha) \leftrightarrow\ & \mathsf{INT}(i,\alpha) \wedge \mathsf{GOAL}(j,\mathit{done}(i,\alpha)) \\ & \wedge\ \mathsf{C\text{-}BEL}_{\{i,j\}}(\mathsf{INT}(i,\alpha) \wedge \mathsf{GOAL}(j,\mathit{done}(i,\alpha)))\end{aligned}$$

Due to its recursive character, reasoning about common beliefs is more complex than attributing second-order attitudes. This explains why children

master agreements later than second-order tasks. Other behavioral research highlights striking limitations in adults [51, 64, 93, 163].

In conclusion, whereas standard epistemic logic idealizes higher-order social cognition of agents, empirical research lacks sophisticated representations needed to solve open problems. Still missing are accounts of how adults improve higher-order social cognition dynamically, how children develop theirs, in which contexts it arises and what is the nature of cognitive limitations. Theory and computational models of these issues are needed in order to implement higher-order social cognition in intelligent systems in interaction. The sequel of this article delineates some ideas about empirical, logical and computational methods aiming to fill this gap.

In particular, I will argue that computational cognitive models shed new light on how higher-order social cognition functions and how it is acquired. A logical perspective helps to formulate the right questions, design illuminating experiments, and precisely define suitable levels of aptitude [101, 102]. Finally, agent-based models settle disputes between theories about evolution of social cognition. Sections 8, 9 and 10 present the achievements and unmet challenges of these computational approaches.

Let us now more closely investigate the three computational approaches, namely *computational cognitive modeling*, resource-bounded *logical modeling*, and *agent-based modeling*.

8 Computational Cognitive Models Such as ACT-R: State of the Art

In cognitive science, a prime approach to investigating human cognition is by constructing computational cognitive models. These are used to understand experimental findings, construct and test theories, and develop new experimental research. The cognitive architecture ACT-R has been developed over the last thirty years. This 'implemented integrated theory of cognition' earned founder Anderson the first Heineken Prize for Cognitive Science [5–7].

Computational cognitive models are constrained by the architecture of ACT-R in the way they retrieve, store, and process information. All architectural constraints of ACT-R are derived from behavioral and neuropsychological experiments on human cognition.[16] ACT-R operates at a symbolic and a subsymbolic level. At the symbolic level, two kinds of memory operate. Declarative memory contains chunks of information representing facts ("knowing that"). Procedural memory contains IF-THEN rules, called production rules, representing actions ("knowing how").[17] Production rules

[16] See http://www.ai.rug.nl/niels/images/actr.jpg for an illustration of the ACT-R modules and their relation to brain regions.

[17] These facts and goal-subgoal production rules make ACT-R reminiscent of logic programming (see also [129]).

compete with each other at the subsymbolic level, where the production rule with the highest expected utility is executed. At this subsymbolic level, there is also competition for retrieval of chunks from the declarative level, dependent on the relevance, recency and frequency of their usage [6, 7].

Higher processing efficiency can be gained by learning through production compilation [135]. This occurs when two existing production rules are used consecutively. They are then integrated into one new production rule, resulting in more automatic processing, which is more efficient if used repeatedly. Production compilation has been successfully used to explain several well-known cognitive phenomena where non-superficial associations are required [66, 133–135, 160].

8.1 Cognitive Models of Reasoning About Mental States

Obviously ACT-R has not been developed with higher-order social cognition in mind. Still, there have been several cognitive models in ACT-R of social reasoning [77, 83, 91, 167]. Models of first-order theory of mind based on different cognitive architectures [16, 110] have not been independently validated by experimental studies. An intriguing ACT-R model with an interactive flavor is [66]. Hendriks et al. investigate why children, when speaking, choose correctly between "Bert washes himself" and "Bert washes him". When listening to somebody else, however, often they misinterpret *him* in "Bert washes him" as co-referring with Bert. The reason is that to interpret "Bert washes him" correctly, one needs to reason about the speaker: if he had intended co-reference, he would have chosen the reflexive form *himself* [18] [65]. The main hypothesis in [66] is that children do have the ability to optimize bidirectionally, but fail, because they lack processing efficiency to serially apply the two required unidirectional optimization processes. Higher-order social cognition has not yet been explicitly modeled in ACT-R.

In conclusion, the development of social reasoning raises a challenge for cognitive modeling. There are two main stages of children's development: roughly between three and five, they learn first-order social cognition; between six and nine, they learn second-order social cognition [106]. Why does this take so long? Do children need to overcome serial processing bottlenecks, as in the language interpretation task of [66]? This remains an intriguing open question.

[18] This does not rule out that language users, when viewed from the algorithmic level (see Section 2.1), could use correct alternative reasoning without any mention of mental states. For example, they could apply a rule such as "reflexives refer to the subject of a sentence, and reflexivization of pronouns co-referring within the clause is obligatory, from which it follows that a non-reflexive pronoun refers outside the clause" (as suggested by Keith Stenning, personal communication).

9 Logical Models: State of the Art

The aim to develop cognitively plausible logics for higher-order social cognition builds on two recent developments to take resource bounds seriously, in logic and artificial intelligence.

Resource-bounded reasoning In game theory and experimental economics, comparing bounded rationality to ideal rationality has an impressive history [11, 26, 117, 125], including a study of 'team reasoning', where agents decide on the basis of the question "what should *we* do to maximize joint utility" [10].

Until recently, however, logicians viewed the standards of rigor needed for mathematics as the norm for reasoning in general. They denounced most human reasoning as being incorrect or fallacious (see for example the classical [165]). The past few years, however, logicians have investigated human reasoning under resource bounds [41, 53, 80, 88, 129, 141].

Let us quote logicians Gabbay and Woods:

> "The theorist's second option is to accept what the empirical record reveals and give it a central place in his investigations. [...] Once we admit agents to logic, it is best to admit them as they actually are, warts and all" [53].

In opposition to the ideal theoretical agent, Gabbay and Woods characterize a practical agent as one who tries to achieve his cognitive goals "with relatively scant cognitive resources, such as information, time, and storage and computational capacities, and who sets the cognitive bar at heights that enable them to be negotiated with the resources at hand" [54]. Gabbay and Woods point to some possible resource strategies, such as approximating and cutting short lengthy processes, using efficient feedback mechanisms, and avoiding irrelevant considerations, thereby putting all reasoning at the service of the need to take quick action [53].

Gabbay and Woods give an interesting logical analysis of a case study in which a lawyer decides after a client's death whether to make a claim on behalf of the widow from the life insurance company or from the government [54]. The reasoning contains some informal theory of mind about the two institutions. They propose to formalize such practical cases not by using one logical language, but using several formalisms for different aspects and linking them together in a suitable meta-logical device.

This meta-logical approach fits well with a proposal by Fenrong Liu [88]. Liu considers some fruitful avenues along which agents may be distinguished, such as introspection ability, powers of observation, memory capacity, and revision policies. She considers the combination of different types of agents in one multi-agent system, leading to the combination of different logics [88].

Efficiency strategies have a strong history in artificial intelligence, too, because software agents also need to decide quickly on a course of action when facing uncertain information and scarce resources [24, 162]. BDI systems,

based on Bratman's practical reasoning paradigm, are tailored to resource-bounded action planning in dynamic environments [20, 39, 40, 112].

9.1 Logical Approaches to First-Order Social Cognition

Stenning and Van Lambalgen [127–129] combine logical models with a close analysis of experiments, taking into account neuro-psychological findings. They illuminatingly analyze children's first-order social reasoning in a first-order false-belief task, using the concept of closed-world reasoning [23]: "if you do not have evidence that the situation is abnormal for agent b (ab_b), then conclude that it isn't abnormal ($\neg ab_b$)", or formally, $\frac{\top : \neg ab_b}{\neg ab_b}$.

Van Ditmarsch and Labuschagne model first-order social cognition in terms of degrees of belief, modeled by a preference relation between possible worlds [153]. They characterize several general stances that agents might have with respect to another agent's preferences. For example, an autistic child has difficulty in distinguishing her own beliefs from those of others [13, 30] and may believe "another agent's preferences are exactly similar to my own". Such a general stance turns out to be frame-characterizable by a formula of a doxastic epistemic logic. Neither of these two studies explicitly treats higher-order social cognition.

In epistemic logic and multi-agent systems, resource bounds have been taken into account. However, in those fields the inspiration was not taken from cognitive science, but from bounds on computational resources and lack of information. In Section 4 three problems with epistemic logic were described. One of the ways to treat the problem of logical omniscience has been to introduce the notion of *awareness* of formulas [49, 95]; nowadays such approaches have become more fine-grained, also incorporating the notion of forgetting [152]. Recently, researchers have proposed alternative ways to account for limited versions of introspection and logical omniscience [1, 18, 103]. Other researchers have limited the complexity of multi-agent logics in terms of time and memory, using syntactic restrictions and limits on the application of deduction rules [4, 42, 43, 61, 162]. It is necessary to distinguish computational complexity from complexity of human cognitive processes. For example, many pattern recognition tasks are easy for children, but notoriously hard for machines [121].[19] Unfortunately, logicians have not yet investigated realistic complexity limits on higher-order social cognition.

In logic and cognitive modeling, usually only present-day cognitive capabilities have been investigated. For true understanding, however, it is also essential to investigate how such complex capabilities have dynamically evolved.

[19] At first sight these different complexity measures seem a matter best viewed at Marr's implementation level (see Section 2.1): Of course different types of computational complexity would fit the Von Neumann architecture and the brain. However, recently Van Rooij has shown convincingly that a complexity-theoretic analysis can help to improve computation-level theories of cognition as well. She posits the thesis that human cognitive capacities are constrained by computational tractability, where 'tractable' is interpreted as 'solvable in parametrized polynomial time' [161].

10 Agent-Based Models: State of the Art

The technique of agent-based social simulation has proved successful since the nineties, for modeling multi-agent phenomena as diverse as fighting in crowds [72], trust in negotiations [63], the evolution of agriculture [148], and the evolution of language [27, 32, 123]. In particular, models of the evolution of cooperation have gained tremendously in sophistication since Axelrod's [9], for example by modeling how organisms move through space [3, 75]. However, the emergence of social cognition has only been partially modeled (see [149–151]).

In general, inputs to a computational agent-based model are the attributes needed to match the model with a specific social setting, based on observations (from psychology, biology and anthropology). Outputs are the behaviors of the computational model through time in a dynamic environment [59].

It is important to distinguish the suggested methodology of agent-based modeling from that of evolutionary psychology as proposed by Cosmides and Tooby [34, 138]. Evolutionary psychologists generally combine historical information with experimental research on present-day human capabilities (see e.g. [29], criticized in [113, 127]). On the other hand, due to the complexity of the phenomena, evolutionary optimization models as used in theoretical biology [104] are not applicable to the evolution of higher-order social cognition.

Agent-based modeling provides a fruitful middle ground between the speculative evolutionary psychology and mathematical modeling. Based on repeated experiments, it enables to investigate how changes in relevant parameters affect the complex behavior of an agent society.

10.1 Agent-Based Models as a Laboratory for Theories of Evolution

Do animals have theory of mind? Controversy abounds, because empirical evidence cannot distinguish use of theory of mind from simple associative learning from experience [21, 25, 45, 67, 94, 108, 137]. Agent-based models come to the rescue by testing alternative theories.

Many of our primate cousins and corvids, like humans, must also deal with a continuously changing set of allies and enemies, dominants and subordinates. Does this mean that they, too, have a theory of mind? After decades of research, the answer seems to be, with the possible exception of the chimpanzee, probably not.

Nevertheless, to us as observers, it often seems as if apes and ravens are acting in ways that require them to think about the beliefs, desires, and intentions of others [45, 94, 137]. Only in carefully designed experiments do their limitations become apparent. In other words, although they may behave as if they have theory of mind, they seem to lack the underlying concepts. Chimpanzees, ravens, and scrub-jays all display behaviors that seem to imply that they can reason about who knows what. Yet, some maintain that, instead, animals are using some combination of instinct, experience and learning, to perform correctly [108]. Taking into account what humans find difficult about

information attribution should ease the assessment of this claim. Conversely, if animals can manage their complex social lives without information attribution, then perhaps humans solve most of their daily problems without it too.

Van der Vaart's TopTalent project in Groningen concentrates on the emergence of first-order theory of mind in birds. So far, the project has delivered a single computational model (inspired by ACT-R) that can replicate the outcomes of a sizable set of experiments on memory in two types of corvids, namely Clark's nutcrackers and scrub-jays. In this way, the first integrated computational account of different behavioral effects of memory in corvid food hiding and recovery is provided, and a new explanation for some hitherto unexplained experimental findings [149, 151]. Thus, the idea of constructing a single computational architecture of corvid cache and recovery cognition appears to be a fruitful one, which gives hope for useful computational cognitive models that can test whether seemingly very smart bird behavior such as 'it takes a thief to hide food where other potential thieves can't find it' really requires theory of mind or only depends on simpler mechanisms.

How would agent-based models help to explain the next step: the evolution of higher-order social cognition, supposedly only displayed by humans? Such models provide a laboratory to rigorously test several theories concerning the evolution of higher-order social cognition and its relation to teamwork. For example, one can investigate how people still perform effectively in mixed-motive contexts such as negotiations about task division in teams, in spite of difficulties with higher-order attributions.

After surveying the problems surrounding higher-order social cognition and the state of the art in computational cognitive models, logic, and agent-based modeling, let us now turn to the more speculative final sections of this paper: a sketch of possible avenues to investigate human higher-order social cognition.

11 Future Research

11.1 Computational Cognitive Models for Higher-Order Social Cognition

Some social cognition tasks are done correctly at an early age (recognizing intentions) [58], others take some years longer (the classic false-belief task) [33, 170], while very complex tasks, such as multi-attribute negotiation, are never reliably learnt by the general adult population [51, 64, 163]. For young children's failure with first-order attributions, experiments show that processing difficulties play the lead role rather than working memory limitations [60]. There may very well exist a processing bottleneck for higher-order social cognition as well.

This hypothesis can be tested by developing a computational cognitive ACT-R model for the step from first-order towards second-order attitude attributions occurring between six and nine years of age. At first sight, it seems that production compilation [133, 135] plays a role in learning to reason at higher levels. It is expected that a cognitively plausible model of higher-

order social cognition reasons about other minds, without exactly computing a complete representation of the other agent's standpoint, particularly at the higher orders (cf. [66]).

On the basis of the theory and the modeling results, including virtual experiments, new ideas for experiments can be developed, for example using card-guessing (Section 4) and the agreement task [89] (see Section 6). Thus, predictions about learning to attribute second-order and common beliefs can be tested. Experiments with normally developing children (6–12 years) may be performed, taking into account their reaction times. Regression models may deliver additional detailed information from reaction times about the strategies used in the children's reasoning processes [147]. Using tasks with artificial slow-downs, one can test the following *hypothesis*: children have the ability to take another's perspective about their own mental state already from the age of six, but fail to apply this correctly, because they lack efficiency to serially apply the necessary mental operations. Finally, based on the first cognitive model and the experiments, one may construct a combined cognitive model of the development of higher-order social cognition for children from 6 up to 12 years of age.

Whereas first-order social cognition is often applied seemingly without effort, adults find it hard to apply second-order social cognition, and even harder to apply third- and higher-order social cognition. It is interesting to investigate what causes this phenomenon. What are the bottlenecks in moving from level 1 to level 2 and from level 2 to level 3?

A plausible hypothesis is that, in order to apply social cognition, people need to store information about possible worlds and different viewpoints in their goal-related memory: the imaginal buffer in ACT-R, corresponding to a posterior region of the parietal cortex [5, 31]. Possibly, the processes found in adults learning to apply tasks that demand higher-order social cognition can best be explained by a combination of the imaginal effect and Salvucci's and Taatgen's threaded cognition model of multitasking [19, 119, 159]. For example, in a game of imperfect information the first task would be playing according to the rules of the game, and a second keeping track of the opponent's possible mental states [163].

Similarly as for children, one can incrementally build a computational cognitive ACT-R model on the basis of a cycle of behavioral experimentation, initial model building, predictions for further experiments, behavioral and fMRI studies, culminating in the construction of an integrated ACT-R model of higher-order social cognition.

11.2 How Logical can Higher-Order Social Cognition be?

One may design resource-bounded variants of standard modal logics for reasoning about other agents such as [50]. Results from the experiments and computational cognitive ACT-R models about processing bottlenecks in complex higher-order attributions may be used to design resource-bounded variants of standard modal logics for reasoning about other agents such as [50].

Both experimental results and computational models may help to tailor the logics to human cognitive capabilities. This deflects the danger of simplistic formal systems that posit a fixed bound on social cognition: everyone can reason up to order n but not on order $n + 1$" [11, 125]. Such a discontinuous idealization is risky. In physics, one expects one's idealizations to work out in a continuous manner: if one deviates from reality by a small amount, one can always make small corrections. However, in logic, when cutting off an approximation at a fixed order, the model changes discontinuously in terms of the truth values of propositions.[20]

One combined system may model several *types of reasoners* with different resource bounds, tuned by parameters such as inferential capabilities, reflective capabilities, and revision policies, taking off from [40, 88]. It would also be interesting to model the results of the second-order reasoning experiments in the constructed resource-bounded logic. Next, one may develop a resource-bounded logic for reasoning about *team attitudes* such as common knowledge (see Section 4), common belief, social commitment, and collective commitment [20, 39, 40, 116]. Then the results of the agreement-experiment [89], in which common beliefs play a role, can be modeled in the resource-bounded multi-agent logic.

Multi-agent systems operate in dynamic environments in which mental states of agents change because of observations and communication. The environment may in turn change as the result of agent actions. Therefore it is useful to integrate these *dynamic aspects*, thereby embedding interacting intelligent systems in their environment. A good starting-point for such more dynamic logics are dynamic epistemic logic [144, 155, 156] and temporal epistemic logics and coalition logics [105, 146], or the recently introduced synthesis of these two approaches [142]. For example, in dynamic epistemic logic the communication actions in the card-guessing task of Section 4 can be formalized in the language as two public announcements: "Ann does not have red", followed by "Bob does not have red". Their effects on the situation can be formally computed, leading from Model I to Model II and finally to Model III of Fig. 1 [155]. Van Eijck developed a useful model checker DEMO for dynamic epistemic logic, that automates this process [44]. Even though people may not be perfect at drawing the right conclusion immediately, with some communicative help they are amazingly good at revising their mistakes, and this should be modeled (cf. [141]).

11.3 Adaptivity of Higher-Order Social Cognition in Context

Why has higher-order social cognition evolved in the first place? One may investigate several hypotheses concerning the possible evolutionary advantages and the costs of higher-order social cognition using computational simulation.

[20]This concern was voiced to us by Johan van Benthem in 2006, referring to a Wittgenstein quote saying that in logic, every deviation causes a large error.

There are three main theories explaining why higher-order social cognition evolved [48]: (1) the need to cooperate with fellow humans [96], (2) the need to manipulate and deceive others—the Machiavellian hypothesis [22, 130, 168], and (3) as a by-product of the complex social life in large groups, that required a large brain [38]. I add another theory: (4) the need of mixed-motive interactions such as negotiations, with partially shared, partially competing interests.

Each theory can be tested by creating an environment that presents a selective problem. For example, for theory (2), one can choose a problem that cannot be solved individually, like hunting large animals. One can implement social and environmental selective pressures and see which variables promote represented higher-order social cognition. The next questions: how does selection happen and how do agents reason with higher-order representations? Special attention should be given to the logical representation of an agent's decision rules and to defining suitable measures of complexity, based on aspects such as formula-length and operator-depth.

One can investigate whether explicit representation of higher-order mental states of others gives *individuals* an advantage over those who only use behavioral association and first-order attributions. The costs of higher-order social cognition should be taken into account. One may investigate several contexts and tasks to test all four theories mentioned above.

Finally, one may study the role of higher-order social cognition in *teamwork*. For which tasks and contexts is the presence of higher-order social cognition in team-mates beneficial for the performance of the team as a whole? Several teamwork tasks and aspects of teamwork may be investigated, such as negotiation about task division. There, individual goals (such as avoiding a distasteful task) may conflict with goals of team-mates and the overall team goal. It will be interesting to find out whether correct higher-order reasoning about members' preferences helps or hinders team performance, and whether communication about such preferences improves the team's overall performance as well as individual agents' goals (as claimed in [111]).

For making these agent-based models, logic is vital to provide precision in the evolving representations of reasoning. Computational cognitive models provide clues for relevant types of decision rules.

12 Closing Remarks

After reviewing existing work on higher-order social cognition and sketching how logic may be joined with experimentation and computational modeling to shed light on some hard questions, let us end by quoting a hopeful vision from Johan van Benthem's "*Logic and reasoning: Do the facts matter?*" [141]:

> "Indeed, the above-mentioned logical theories of inference, update, and interaction all suggest interesting testable hypotheses about human behaviour, and one could easily imagine a world where a logician who

has created a new logical system does two things instead of one: like now, submit to a logic conference, usually far abroad, but also: telephone the psychologist next door to see if some nice new experiment can be done. And finally, going yet a bit further, I would think that logic can also contribute to a better understanding of how humans form and maintain representations of scenarios and their relevant information, the stage prior to any significant processing. What this would involve is a broadening of current 'model theory' to a 'theory of modeling'."

Acknowledgements One of the inspirations for this article was the international project *Games, Action and Social Software* at NIAS in 2006–2007, born from Johan van Benthem's initiative, which he transfered to Jan van Eijck and myself to lead. At NIAS, a host of international researchers built bridges between logic, multi-agent systems, and cognitive science through exciting new collaborations. I would like to thank all of them, next to Johan van Benthem and Jan van Eijck in particular the other longer-term participants: Barbara Dunin-Kęplicz, Krister Segerberg, Martin van Hees, Rohit Parikh, Peter Gärdenfors, Keith Dowding, Andrzej Szałas, Nicola Dimitri, Barteld Kooi, Marc Pauly, and Hans van Ditmarsch. I would also like to thank a number of other colleagues for discussions on social cognition and / or for being co-authors of pilot studies: Marian Counihan, Liesbeth Flobbe, Charlotte Hemelrijk, Petra Hendriks, Irene Krämer, Michiel van Lambalgen, Alice ter Meulen, Lisette Mol, Ram Ramanujam, Hedderik van Rijn, Keith Stenning, Jakub Szymanik, Niels Taatgen, Elske van der Vaart, and Bart Verheij. Both referees for this article were generous with their insightful suggestions. Finally, I would like to thank the Netherlands Organization for Scientific Research for Vici grant NWO-277-80-001, for the project *Cognitive systems in interaction: Logical and computational models of higher-order social cognition*.

References

1. Agotnes, T., & Alechina, N. (2007). The dynamics of syntactic knowledge. *Journal of Logic and Computation, 17*(1), 83–116.
2. Aiello, L. C. (1996). Hominine preadaptations for language and cognition. In K. Gibson, & P. Mellars (Eds.), *Modelling the early human mind*. (pp. 89–99). Cambridge: McDonald Institute for Archaeological Research.
3. Aktipis, C. A. (2004). Know when to walk away: Contingent movement and the evolution of cooperation. *Journal of Theoretical Biology, 231*, 249–260.
4. Alechina, N., Bertoli, P., Ghidini, C., Jago, M., Logan, B., & Serafini, L. (2006). Model checking space and time requirements for resource-bounded agents. In *Proceedings of the fourth international workshop on model checking and artificial intelligence* (pp. 16–30).
5. Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.
6. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060.
7. Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah: Lawrence Erlbaum.
8. Astington, J. W., Harris, P. L., & Olson, D. R. (Eds.) (1988). *Developing theories of mind*. Cambridge: Cambridge University Press.
9. Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
10. Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton: Princeton University Press.
11. Bacharach, M., & Stahl, D. O. (2000). Variable-frame level-n theory. *Games and Economic Behavior, 32*(2), 220–246.

12. Baltag, A., Moss, L. S., & Solecki, S. (2003). *The logic of public announcements, common knowledge, and private suspicions.* Technical report, Dept of Cognitive Science, Indiana University and Dept of Computing, Oxford University.
13. Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition, 21*, 37–46.
14. Barwise, J. (1989). On the model theory of common knowledge. In: *The situation in logic* (pp. 201–221). Stanford: CSLI.
15. Barwise, J. (1989). *The situation in logic*. Stanford: CSLI.
16. Bello, P., & Cassimatis, N. (2006). Developmental accounts of theory-of-mind acquisition: Achieving clarity via computational cognitive modeling. In R. Sun, & N. Miyake (Eds.), *Proceedings of twenty-eighth annual meeting of the cognitive science society* (pp. 1014–1019). Vancouver: Cognitive Science Society.
17. Benz, A., & van Rooij, R. (2007). Optimal 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.)
18. Bonnay, D., & Égré, P. (2008). Margins for error in context. In M. Garcia-Carpintero, & M. Kölbel (Eds.), *Relative truth* (pp. 103–127). Oxford: Oxford University Press.
19. Borst, J. P., Taatgen, N. A., & van Rijn, H. (2009). Problem representations in multitasking: An additional cognitive bottleneck. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 9th international conference on cognitive modeling*. Manchester: University of Manchester.
20. Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge: Harvard University Press.
21. Burkart, J., & Heschl, A. (2007). Perspective taking or behaviour reading? Understanding visual access in common marmosets (callithrix jacchus). *Animal Behaviour, 73*, 457–469.
22. Byrne, R. W., & Whiten, A. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes and humans*. Oxford: Clarendon.
23. Cadoli, M., & Lenzerini, M. (1994). The complexity of propositional closed world reasoning and circumscription. *Journal of Computer and System Sciences, 48*, 255–310.
24. Cadoli, M., & Schaerf, M. (1995). Approximate inference in default logic and circumscription. *Fundamenta Informaticae, 23*, 123–143.
25. Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*, 187–192.
26. Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton: Princeton University Press.
27. Cangelosi, A., & Parisi, D. (Eds.) (2002). *Simulating the evolution of language*. Berlin: Springer.
28. 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.
29. Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason Selection Task. *Cognition, 31*, 187–276.
30. Dapretto, M. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum. *Nature Neuroscience, 9*(1), 28–30.
31. David, N., Aumann, C., et al. (2008). Differential involvement of the posterior temporal cortex in mentalizing but not perspective taking. *Social Cognitive and Affective Neuroscience, 3*, 279–289.
32. de Boer, B. (2001). *The origins of vowel systems*. Oxford: Oxford University Press.
33. de Villiers, J. G., & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false beliefs. In P. Mitchell, & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 191–228). East Sussex: Psychology.
34. Dennett, D. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon and Schuster.
35. Dignum, F., Dunin-Kęplicz, B., & Verbrugge, R. (2001). Creating collective intention through dialogue. *Logic Journal of the IGPL, 9*, 145–158.
36. 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.

37. Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution, 22*, 469–493.
38. Dunbar, R. (1996). *Grooming, gossip and the evolution of language*. London: Faber and Faber.
39. Dunin-Kęplicz, B., & Verbrugge, R. (2002). Collective intentions. *Fundamenta Informaticae, 51*(3), 271–295.
40. Dunin-Kęplicz, B., & Verbrugge, R. (2004). A tuning machine for cooperative problem solving. *Fundamenta Informaticae, 63*, 283–307.
41. Dunin-Kęplicz, 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.
42. Dziubiński, M. (2007). Complexity of the logic for multiagent systems with restricted modal context. In B. Dunin-Kęplicz, & R. Verbrugge (Eds.), *Proceedings of the third workshop on formal approaches to multi-agent systems (FAMAS'007)* (pp. 1–18). Durham: Durham University.
43. Dziubiński, M., Verbrugge, R., & Dunin-Kęplicz, B. (2007). Complexity issues in multiagent logics. *Fundamenta Informaticae, 75*(1-4), 239–262.
44. van Eijck, J. (2007). DEMO—a demo of epistemic modelling. In J. van Benthem, D. Gabbay, & B. Löwe (Eds.), *Interactive logic—proceedings of the 7th Augustus de Morgan workshop, number 1 in texts in logic and games* (pp. 305–363). Amsterdam: University Press.
45. Emery, N. J. (2005). The evolution of social cognition. In A. Easton, & N. J. Emery (Eds.), *Cognitive neuroscience of social behaviour* (pp. 115–156). London: Psychology.
46. Erdal, D., & Whiten, A. (1996). Egalitarianism and Machiavellian intelligence in human evolution. In K. Gibson, & P. Mellars (Eds.), *Modelling the early human mind* (pp. 139–150). Cambridge: McDonald Institute for Archaeological Research.
47. Hill, J. C., et al. (2005). A cooperative multi-agent approach to free flight. In F. Dignum, et al. (Eds.), *AAMAS '05: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems* (pp. 1083–1090). New York: ACM.
48. Emery, N. J., et al. (2007). Cognitive adaptations of social bonding in birds. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362*, 489–505.
49. Fagin, R., & Halpern, J. (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence, 34*, 39–76.
50. Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*, 2nd ed., 2003. Cambridge: MIT.
51. 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.)
52. Foley, R. A. (1996). Measuring cognition in extinct hominids. In K. Gibson, & P. Mellars (Eds.), *Modelling the early human mind* (pp. 57–65). Cambridge: McDonald Institute for Archaeological Research.
53. Gabbay, D. M., & Woods, J. (2001). The new logic. *Logic Journal of the IGPL, 9*, 157–190.
54. Gabbay, D. M., & Woods, J. (2008). Resource-origins of nonmonotonicity. *Studia Logica, 88*, 85–112. (Special issue on logic and the new psychologism, edited by H. Leitgeb).
55. Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind reading. *Trends in Cognitive Sciences, 2*(12), 493–501.
56. Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8*(9), 396–403.
57. Gärdenfors, P. (2009). The communicative and cognitive demands of cooperation. In J. van Eijck, & R. Verbrugge (Eds.), *Games, actions and social software*. Oxford: Oxford University Press. (Earlier short version appeared in Hommage à Wlodek: Philosophical papers dedicated to Wlodek Rabinowicz.)
58. Gattis, M., Bekkering, H., & Wohlschläger, A. (2002). Goal-directed imitation. In A. Meltzoff, & W. Prinz (Eds.), *The imitative mind: development, evolution, and brain bases* (pp. 183–205). Cambridge: Cambridge University Press.
59. Gilbert, N., & Troitzsch, K. G. (Eds.) (2005). *Simulation for the social scientist*. Maidenhead: Open University Press.

60. Hala, S., Hug, S., & Henderson, A. (2003). Executive function and false-belief understanding in preschool children: Two tasks are harder than one. *Journal of Cognition and Development, 4*, 275–298.
61. Halpern, J. (1995). The effect of bounding the number of primitive propositions and the depth of nesting on the complexity of modal logic. *Artificial Intelligence, 75*, 361–372.
62. Halpern, J. Y., & Moses, Y. (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM, 37*, 549–587.
63. 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.
64. Hedden, T., & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. *Cognition, 85*, 1–36.
65. Hendriks, P., & Spenader, J. (2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition, 13*(4), 319–348.
66. Hendriks, P., van Rijn, H., & Valkenier, B. (2007). Learning to reason about speakers' alternatives in sentence comprehension: A computational account. *Lingua, 117*(11), 1879–1896.
67. Heyes, C. M. (1998). Theory of mind in non-human primates. *Behavioral and Brain Sciences, 21*, 101–148.
68. Hintikka, J. (1962). *Knowledge and belief*. Ithaca: Cornell University Press.
69. Hogrefe, G., & Wimmer, H. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development, 57*, 567–582.
70. Hurley, S. (2005). Social heuristics that make us smarter. *Philosophical Psychology, 18*(5), 585–611.
71. Hurley, S. (2008). The shared circuits model: How control, mirroring and simulation can enable imitation, deliberation, and mindreading. *Behavioral and Brain Sciences, 31*, 1–22.
72. Jager, W., Popping, R., & van de Sande, H. (2001). Clustering and fighting in two-party crowds: Simulating the approach-avoidance conflict. *Journal of Artificial Societies and Social Simulation, 4*(3). http://jasss.soc.surrey.ac.uk/4/3/7.html.
73. Jennings, N. R., & Bussmann, S. (2003). Agent-based control systems: Why are they suited to engineering complex systems? *IEEE Control Systems Magazine, 23*(3), 61–74.
74. Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-agent Systems, 1*, 7–38.
75. Kaplan, F., & Hafner, V. (2004). The challenge of joint attention. In L. Barthouze, et al. (Eds.), *Proceedings of the fourth conference on epigenetic robotics: modelling cognitive development in robotic systems* (pp. 67–74). Lund: Lund University.
76. Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge: MIT.
77. Kennedy, W. G., & Trafton, J. G. (2007). Using simulations to model shared mental models. In R. L. Lewis, T. A. Polk, & J. E. Laird (Eds.), *Proceedings of the eighth international conference on cognitive modeling* (pp. 253–245). London: Psychology / Taylor and Francis.
78. Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25–41.
79. Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: From self to social cognition. *Trends in Cognitive Sciences, 11*(5), 194–196.
80. Knauff, M. (2007). How our brains reason logically. *Topoi, 26*(1), 19–36. (Special issue on logic and psychology, edited by J.F.A.K. van Benthem, H. Hodges, and W. Hodges.)
81. Krause, J., & Ruxton, G. D. (2002). *Living in groups*. Oxford: Oxford University Press.
82. Kripke, S. (1959). A completeness theorem in modal logic. *Journal of Symbolic Logic, 24*, 1–14.
83. Lebiere, C., Wallach, D., & West, R. (2000). A memory-based account of the prisoner's dilemma and other 2x2 games. In N. A. Taatgen, & J. Aasman (Eds.), *Proceedings of third international conference on cognitive modeling* (pp. 185–193). Veenendaal: Universal Press.
84. Leitgeb, H. (2008). Introduction to the special issue. *Studia Logica, 88*, 1–2. (Special issue on logic and the new psychologism, edited by H. Leitgeb.)

85. Levesque, H. J., Cohen, P. R., & Nunes, J. H. T. (1990). On acting together. In *Proceedings eighth national conference on AI (AAAI90)* (pp. 94–99). Menlo Park: AAAI and MIT.
86. Lin, R., & Krauss, S., et al. (2008). Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *Artificial Intelligence Journal, 172*(6–7), 823–851.
87. Litman, L., & Reber, A. S. (2005). Implicit cognition and thought. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 431–453). Cambridge: Cambridge University Press.
88. Liu, F. (2006). Diversity of agents. In T. Agotnes, & N. Alechina (Eds.), *Proceedings of the workshop on resource-bounded agents* (pp. 88–98). European Summer School on Logic, Language and Information, Malaga.
89. Mant, C. M., & Perner, J. (1988). The child's understanding of commitment. *Developmental Psychology, 24*, 343–351.
90. Marr, D. (1982). *Vision*. New York: Freeman.
91. Matessa, M. (2001). Interactive models of collaborative communication. In J. D. Moore, & K. Stenning (Eds.), *Proceedings of the twenty-third annual meeting of the cognitive science society* (pp. 606–610). Mahwah: Erlbaum.
92. McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *Proposal for the Dartmouth summer research project on artificial intelligence.* Technical report, Dartmouth College.
93. McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica, 60*(4), 803–836.
94. Melis, A. P., Hare, B., & Tomasello, M. (2006). Chimpanzees recruit the best collaborators. *Science, 311*, 1297–1300.
95. Modica, S., & Rustichini, A. (1999). Unawareness and partitional information structures. *Games and Economic Behavior, 27*, 265–298.
96. Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 362*, 639–648.
97. Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economical Review, 85*, 1313–1326.
98. Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding of other minds*. Oxford: Oxford University Press.
99. Oakman, J. (2002). *The Camp David Accords: A case study on international negotiation.* Technical report, Princeton University, Woodrow Wilson School of Public and International Affairs.
100. Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258.
101. Pacuit, E., Parikh, R., & Cogan, E. (2006). The logic of knowledge based obligation. *Synthese: Knowledge, Rationality and Action, 149*, 57–87.
102. Parikh, R. (2003). Levels of knowledge, games, and group action. *Research in Economics, 57*, 267–281.
103. Parikh, R. (2007). *Logical omniscience in the many agent case.* Technical report, City University of New York, New York.
104. Parker, G.A., & Maynard Smith, J. (1990). Optimality theory in evolutionary biology. *Nature, 348*, 27–33.
105. Pauly, M. (2002). A modal logic for coalitional power in games. *Journal of Logic and Computation, 12*, 149–166.
106. 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.
107. Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that ...": Attribution of second-order beliefs by 5- to 10-year old children. *Journal of Experimental Child Psychology, 5*, 125–137.
108. Povinelli, D., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *Trends in Cognitive Sciences, 7*, 157–160.

109. Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 4*, 515–526.
110. Pynadath, D. V., & Marsella, S. C. (2005). PsychSim: Modeling theory of mind with decision-theoretic agents. In L. Kaelbling, & A. Saffiotti (Eds.), *Proceedings of the 19th international joint conference on artificial intelligence* (pp. 1181–1186). Edinburgh: Professional Bookcenter.
111. Raiffa, H. (1982). *The art and science of negotiation*. Cambridge: Harvard University Press.
112. Rao, A., & Georgeff, M. (1991). Modeling rational agents within a BDI-architecture. In R. Fikes, & E. Sandewall (Eds.), *Proceedings of the second conference on knowledge representation and reasoning* (pp. 473–484). San Francisco: Morgan Kaufman
113. Rose, H., & Rose, S. (Eds.) (2000). *Alas, poor Darwin: Arguments against evolutionary psychology*. New York: Random House.
114. Rosenschein, J. S., & Zlotkin, G. (1994). *Rules of encounter: Designing conventions for automated negotiation among computers*. Cambridge: MIT.
115. Rosenthal, R. (1981). Games of perfect information, predatory pricing, and the chain store. *Journal of Economic Theory, 25*, 92–100.
116. Roy, O. (2006). Commitment-based decision making for bounded agents. In T. Agotnes, & N. Alechina (Eds.), *Proceedings of the workshop on resource-bounded agents* (pp. 112–123). European Summer School on Logic, Language and Information, Malaga.
117. Rubinstein, A. (1998). *Modeling bounded rationality*. Cambridge: MIT.
118. Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach*, 2nd ed. Englewood Cliffs: Prentice-Hall.
119. Salvucci, D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115*, 101–130.
120. Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.
121. Schomaker, L., Hoenkamp, E., & Mayberry, M. (1998). Towards collaborative agents for automatic handwriting recognition. In *Proceedings of the third European workshop on handwriting analysis and recognition. Digest*, (Volume 1998/440 pp. 13/1–13/6). London: The Institution of Electrical Engineers.
122. Schurr, N., Marecki, J., Tambe, M., & Scerri, P. (2005). Towards flexible coordination of human-agent teams. *Multiagent and Grid Systems, 1*(1), 3–16.
123. Slingerland, I., Mulder, M., van der Vaart, E., & Verbrugge, R. (2009). A multi-agent systems approach to gossip and the evolution of language. In N. Taatgen, et al. (Eds.), *Proceedings of the 31st annual meeting of the cognitive science society (CogSci'09)* (pp. 1609–1614). Amsterdam.
124. Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.
125. Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior, 10*, 218–254.
126. Stalnaker, R. (1984). *Inquiry*. Cambridge: MIT.
127. Stenning, K., & van Lambalgen, M. (2001). Semantics as a foundation for psychology: A case study of Wason's selection task. *Journal of Logic, Language and Information, 10*, 273–317.
128. Stenning, K., & van Lambalgen, M. (2007). Logic in the study of psychatric disorders: Executive function and rule-following. *Topoi, 26*(1), 97–114. (Special issue on logic and psychology, edited by J.F.A.K. van Benthem, H. Hodges, and W. Hodges).
129. Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge: MIT.
130. Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition*. Oxford: Blackwell.
131. Stulp, F., & Verbrugge, R. (2002). A knowledge-based algorithm for the internet protocol TCP. *Bulletin of Economic Research, 54*(1), 69–94.
132. 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. Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition, 86*(2)(2), 123–155.

134. Taatgen, N. A., Huss, D., & Anderson, J. R. (2006). How cognitive models can inform the design of instructions. In D. Fum, F. del Missier, & A. Stocco (Eds.), *Proceedings of the seventh international conference on cognitive modeling* (pp. 304–309). University of Trieste.
135. Taatgen, N. A., & Lee, F. J. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors, 45*(1), 61–76.
136. Tager-Flussberg, H., & Sullivan, K. (1994). A second look at second-order belief attribution in autism. *Journal of Autism and Developmental Disorders, 24*, 577–586.
137. Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Sciences, 7*, 153–156.
138. Tooby, J., & Cosmides, L. (Eds.) (2000). *Evolutionary psychology: Foundational papers*. Cambridge: MIT.
139. van Benthem, J. F. A. K. (2005). Correspondence theory. In D. M. Gabbay, & F. Guenthner (Eds.), *Handbook of philosophical logic* (2nd ed., Vol. 3, pp. 325–408). Kluwer: Dordrecht. (An earlier version appeared in volume II of the first edition of the Handbook.).
140. van Benthem, J. F. A. K. (2007). Cognition as interaction. In *Proceedings symposium on cognitive foundations of interpretation* (pp. 27–38). Amsterdam: KNAW.
141. van Benthem, J. F. A. K. (2008). Logic and reasoning: Do the facts matter? *Studia Logica, 88*, 67–84. (Special issue on logic and the new psychologism, edited by H. Leitgeb)
142. 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.
143. van Benthem, J. F. A. K., Hodges, H., & Hodges, W. (2007). 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.).
144. van Benthem, J. F. A. K., van Eijck, J., & Kooi, B. (2006). Logics of communication and change. *Information and Computation, 204*(11), 1620–1662.
145. van der Hoek, W., & Verbrugge, R. (2002). Epistemic logic: A survey. In L. A. Petrosjan, & V. V. Mazalov (Eds.), *Game theory and applications* (Vol. 8, pp. 53–94). New York: Nova Science.
146. van der Hoek, W., & Wooldridge, M. (2003). Time, knowledge and cooperation: Alternating-time temporal epistemic logic and its applications. *Studia Logica, 75*(1), 125–157.
147. van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85*, 141–177.
148. van der Vaart, E., Hankel, A., de Boer, B., & Verheij, B. (2006). Agents adopting agriculture: Modeling the agricultural transition. In *From animals to animats 9: Ninth international conference on simulation of adaptive behavior. LNCS* (Vol. 4095, pp. 750–761). Berlin: Springer.
149. van der Vaart, E., Hemelrijk, C., & Verbrugge, R. (2009). A cognitive model for corvids: Learning where (not) to cache. In N. Taatgen et al. (Eds.), *Proceedings of the 31st annual meeting of the cognitive science society (CogSci'09)* (pp. 2420–2425). Amsterdam.
150. van der Vaart, E., & Verbrugge, R. (2008). Agent-based models for animal cognition: A proposal and a prototype. In *International conference on autonomous agents and multi-agent systems (AAMAS)* (pp. 1145–1152). New York: ACM.
151. van der Vaart, E., Verbrugge, R., & Hemelrijk, C. (2009). Memory in Clark's nutcrackers: A cognitive model for corvids. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 9th international conference on cognitive modeling*. Manchester: University of Manchester.
152. van Ditmarsch, H., & French, T. (2009). Awareness and forgetting of facts and agents. In *WLIAMAS, proceedings of the 2009 IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technologies (WI-IAT 2009)*. Milan: IEEE Computer Society.
153. van Ditmarsch, H., & Labuschagne, W. (2007). My beliefs about your beliefs: A case study in theory of mind and epistemic logic. *Synthese: Knowledge, Rationality and Action, 155*, 191–209.
154. van Ditmarsch, H., van Eijck, J., & Verbrugge, R. (2009). Common knowledge and common belief. In J. van Eijck, & R. Verbrugge (Eds.), *Discourses on social software. Texts in games and logic* (Vol. 5). Amsterdam: Amsterdam University Press.

155. van Ditmarsch, H. P. (2002). The description of game actions in Cluedo. In L. A. Petrosjan, & V. V. Mazalov (Eds.), *Game theory and applications* (Vol. 8, pp. 1–28). New York: Nova Science.
156. van Ditmarsch, H. P., van der Hoek, W., & Kooi, B. P. (2007). In *Dynamic epistemic logic*, *Synthese library series* (Vol. 337). Berlin: Springer.
157. van Eijck, J., & Verbrugge, R. (Eds.) (2009). *Discourses on social software*. *Texts in games and logic* (Vol. 5). Amsterdam: Amsterdam University Press.
158. 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.).
159. Van Maanen, L., & Van Rijn, H. (2007). An accumulator model of semantic interference. *Cognitive Systems Research, 8*, 174–181.
160. van Rijn, H., van Someren, M., & van der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science, 27*(2), 227–257.
161. van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science, 32*, 939–984.
162. Verberne, A., van Harmelen, F., & ten Teije, A. (2000). Anytime diagnostic reasoning using approximate boolean constraint propagation. In A. G. Cohn, F. Giunchiglia, & B. Selman (Eds.), *Proceedings of the seventh international conference on principles of knowledge representation and reasoning* (pp. 323–332).
163. 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.).
164. von Wright, G. H. (1951). *An essay in modal logic*. Amsterdam: North Holland.
165. Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology I*, (pp. 135–151). Harmondsworth: Penguin.
166. Wellman, H. (1991). From desires to beliefs: Acquisition of a theory of mind. In A. Whiten (Ed.), *Natural theories of mind* (pp. 19–38). Oxford: Basil Blackwell.
167. West, R. L., Lebiere, C., & Bothell, D. J. (2006). Cognitive architectures, game playing, and human evolution. In R. Sun (Ed.), *Cognition and multi-agent interaction: From cognitive modeling to social simulation* (pp. 103–123). New York: Cambridge University Press.
168. Whiten, A., & Byrne, R. W. (1997). *Machiavellian intelligence II*. Cambridge: Cambridge University Press.
169. Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
170. 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.
171. Wooldridge, M. J. (2002). *An introduction to multiagent systems*. Chichester: Wiley.