I Do Know What You Think I Think: Second-Order Theory Of Mind In Strategic Games Is Not That Difficult

Ben Meijering[1] (b.meijering@rug.nl), Hedderik van Rijn[2], Niels A. Taatgen[1], & Rineke Verbrugge[1]
[1]Department of Artificial Intelligence, University of Groningen, and [2] Department of Psychology, University of Groningen

Abstract

This paper is about higher-order theory of mind such as "I think that you think that I think …". Previous studies have argued that using higher-order theory of mind in the context of strategic games is difficult and cognitively demanding. In contrast, we claim that performance depends on task properties such as instruction, training, and procedure of asking for social reasoning. In an experiment based on a two-player game, we manipulated these task properties and found that higher-order theory of mind improved by providing step-by-step instruction and training. It also improved during the experiment when participants were explicitly asked to predict the opponent's next move.

Keywords: Theory of Mind; Social Cognition; Higher-Order Social Reasoning; Strategic Games.

Theory of Mind

Whenever the outcomes of our actions depend on the decision of others, and vice versa, we need to reason about one another. For example, if a researcher wants her paper to be accepted, she not only needs to have interesting empirical results, she also needs to get her story across. She needs to reason about what an intended reader knows beforehand and about what he will infer from reading her story. She may even wonder whether a particular reviewer knows that she knows that he was the one who wrote that glowing review. The ability to reason about the knowledge, beliefs, desires and intentions of others, in this case the reader, is often referred to as Theory of Mind (Onishi & Baillargeon, 2005; Wimmer & Perner, 1983; Premack & Woodruff, 1978).

So far, empirical findings have shown theory of mind to be far from optimal, especially in more complex social interactions (Flobbe, Verbrugge, Hendriks, & Krämer, 2008; Keysar, Lin, & Barr, 2003; Hedden & Zhang, 2002; McKelvey & Palfrey, 1992; but see Goodie, Doshi, & Young, 2010). The conclusion often drawn from these findings is that theory of mind is difficult and cognitively demanding (e.g., Verbrugge & Mol, 2008). In contrast, we claim that performance depends on the task. For example, participants seemed to have little difficulties applying theory of mind in false-belief story tasks (Flobbe, et al., 2008).

We claim that suboptimal performance due to task difficulties can be overcome by providing appropriate instruction and training. Social reasoning involves interplay of multiple serial and concurrent cognitive processes, and learning to apply theory of mind in a particular task might benefit from instruction and training that structure this interplay of processes. Besides instruction and training, the procedure of asking for social reasoning can also contribute to providing a supporting structure, so-called scaffolding, for the interplay of processes that underlie social reasoning.

In the current study, we show that providing supporting structure that maps with the reasoning steps required by the task facilitates social reasoning and improves performance.

Orders of Reasoning

Complex social interactions such as rescue operations and negotiations are cognitively demanding because of the depth, or order, of reasoning they require (Verbrugge, 2009). To illustrate orders of reasoning, imagine a social interaction between Ann, Bob, and Carol, and that Bob's birthday is tomorrow. Furthermore, Ann knows: "Bob's birthday is tomorrow". This is an example of zeroth-order or non-social reasoning, because Ann is not yet reasoning about someone else's mental state. She merely recalls a fact.

If Bob thinks: "Ann knows my birthday is tomorrow", he is applying first-order reasoning, because he ascribes knowledge to Ann. First-order reasoning covers a great deal of social interactions. Another example of a first-order attribution is Bob' thought: "Ann intends to throw me a surprise party, because she always throws surprise parties". In this example Bob ascribes an intention to Ann.

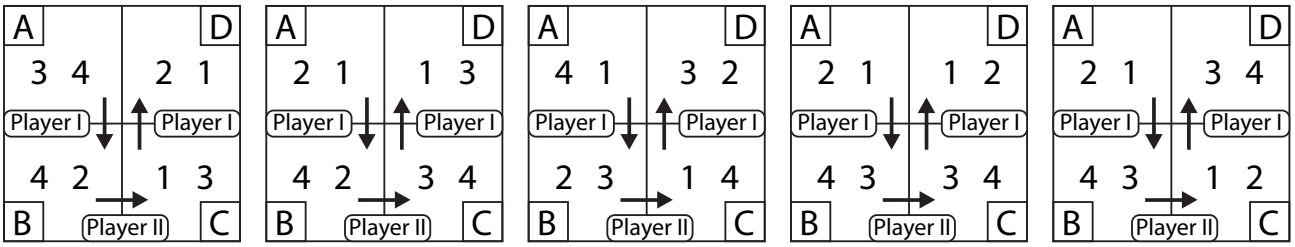
A social interaction between Ann, Bob, and Carol may demand reasoning of one order deeper: Suppose that Ann will not throw Bob a surprise party and expressed this to Carol. Now, Carol knows that "Bob falsely believes that Ann intends to throw him a surprise party".

Carol applied second-order reasoning, which is a complex skill that starts developing around the age of 6 to 9 years and apparently remains challenging throughout the later lifespan (Perner & Wimmer, 1985). Second-order reasoning is the main focus of the current study.

Hedden and Zhang's Experiments

So far, most studies showed suboptimal application of second-order reasoning in social interactions that involved the perspective of a participant and one other person or player (Flobbe, et al., 2009; Hedden & Zhang, 2002; McKelvey & Palfrey, 1992). For example, McKelvey and Palfrey (1992) presented participants with games in which they had to reason about an opponent's decisions. Participants' behavior in these games was not optimal. Hedden and Zhang (2002) found that participants correctly used second-order reasoning in approximately 65% of the trials, but not until the end of the experiments.

Hedden and Zhang presented participants with matrix games, which are two-player sequential-move games. Figure 1 depicts examples of matrix games. Each cell of a



(a) (b) (c) (d) (e)

Figure 1: Five example matrix games. Each cell in a game contains two payoffs. The first payoff is Player I's, the second Player II's. Each game starts in cell A. For both players, the goal is that the game ends in a cell that contains their highest possible payoff. In example game *a*, a rational Player I should decide to continue the game to cell B, as a rational player II should decide to stop in cell B, because Player I should decide to continue from cell C to cell D. Example games *b* to *e* were excluded because they did not require second-order reasoning, or a correct second-order prediction or decision was equal to a first-order prediction or decision.

matrix game contained a pair of rewards, so-called payoffs, that both ranged from 1 to 4 (Figure 1). The first payoff in a cell was Player I's, the second Player II's. For both players, the goal was that the game stopped in a cell that contained their highest possible payoff. Participants, always assigned to the role of Player I, had to apply theory of mind because their outcomes depended on the decisions of Player II, and vice versa. The computer played the role of Player II.

Each game started in cell A. Participants (Player I) were asked to decide whether to stop the game in that cell or to continue it to cell B. If a participant decided to continue the game, the computer (Player II) decided whether to stop the game in cell B or to continue to cell C. If the computer decided to continue, the participant was asked to decide whether to stop the game in cell C or in cell D. When a game was stopped in a particular cell, both players received their payoffs in that cell. The game required second-order reasoning because participants had to reason about the computer's decision in cell B, and thus reason about what the computer thinks that a participant's decision should be in cell C.

Participants had difficulties playing matrix games (Hedden and Zhang, 2002), which was reflected in suboptimal performance. A possible explanation is task difficulty, instead of difficulties applying theory of mind. In our view, task difficulty can be overcome by providing participants with appropriate instruction and training that explain how social reasoning comes into play in matrix games. Hedden and Zhang explained the rules of matrix games and provided training, however we think that their training could be improved by providing support that structures the reasoning steps required by matrix games.

We also think that Hedden and Zhang's training may have been misleading. During the training phase, they presented participants with so-called trivial games that allowed for both first- and second-order reasoning (e.g., Figure 1b). In these games, Player II did not have to reason about Player I's last decision, because Player II's payoffs in cells C and D were both either lower or higher than Player II's payoff in cell B. Consequently, it would suffice for a participant (Player I) to apply first-order reasoning that includes a zeroth-order opponent that does not consider what Player I's last decision should be. If a participant had adopted first-order reasoning, she or he had to *unlearn* that in similar-looking non-trivial second-order games during the experiment, resulting in a gradual increase in performance. We think that performance could have started at a higher level if training consisted of games that unambiguously required second-order reasoning.

Participants' performance indeed improved over time (Hedden & Zhang, 2002). The results of Hedden and Zhang showed considerable learning effects as the proportion of games in which participants applied second-order reasoning started at approximately 25% and monotonically increased to approximately 65%.

Whereas training lacked supporting structure, Hedden and Zhang's procedure of asking for social reasoning during the experiment did provide support: In each game, Hedden and Zhang asked participants two responses: (1) predict the opponent's decision in cell B, and after the prediction, (2) decide what to do in cell A. This procedure provided supporting structure by closely mapping the reasoning steps required by matrix games with the responses asked for: Predicting the opponent's move precedes making a decision.

Supporting Structure

We hypothesized that performance in matrix games could be improved by providing supporting structure in instruction and training, besides supporting structure in the procedure of asking for social reasoning.

Incrementally explaining and training orders of reasoning can provide supporting structure. By first presenting zeroth- and first-order games, participants not only learn that social reasoning can involve multiple orders, they also learn the rules of the matrix game: a participant first makes a decision; if given a choice (in first-order games), the opponent decides second; if given a choice in second-order games, the third decision is the participant's. Figure 2 depicts example zeroth- and first-order games.

Hedden and Zhang did provide supporting structure in their procedure of asking for social reasoning by asking participants to predict the opponent's move before making a

decision. We explicitly tested whether such a procedure had a positive effect on performance.

Method

Participants

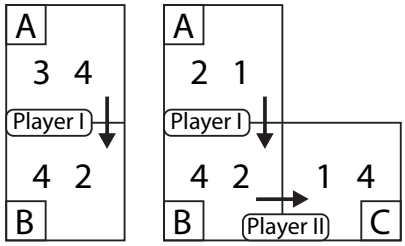
Ninety-five first-year psychology students participated in this study in exchange for course credit. Each participant gave informed consent prior to admission into the study.

Materials

Game Forty-eight participants played matrix games as described above, and forty-seven participants played game-theoretically equivalent games called Marble Drop. The latter games adhered to exactly the same rules as the former, but differed in appearance[1]. We have described the effects of game representation elsewhere (Meijering, Van Maanen, Van Rijn, & Verbrugge, 2010). As the focus of this study is supporting structure in instruction, training, and procedure of asking for social reasoning, we collapsed the data across the two levels of game representation (i.e., matrix game and Marble Drop). It is important to note that the main and interaction effects reported here did not change with or without the inclusion of the factor game representation.

Design We manipulated two factors: (1) structure in instruction and training, and (2) structure in the procedure of asking for social reasoning. We labeled these factors *training* and *asking*. Both factors had two levels: we either did or did not provide supporting structure for social reasoning, which we explain below.

Training that provided such structure consisted of separate instruction and training for zeroth-, first-, and second-order games (Figure 2). Furthermore, each zeroth-, first-, and second-order game was diagnostic of the corresponding order of reasoning: simpler less demanding strategies did not yield correct decisions. Training that lacked supporting structure consisted of 24 trivial games (e.g., Figure 1b), similar to Hedden and Zhang's training. Trivial games allowed for both first- and second-order reasoning, and therefore could not be diagnostic of second-order reasoning. We think that these games biased participants to apply the simpler less demanding first-order reasoning strategy.



(a) (b)

Figure 2: A zeroth-order (a) and first-order (b) matrix game. These example games unambiguously require zeroth- and first-order reasoning.

Asking for social reasoning was manipulated as follows. During the experiment (i.e., two test blocks), we either did or did not ask participants to predict the opponent's move in cell B before deciding what to do in cell A. By asking participants to predict the opponent's move, we provided supporting structure for social reasoning in matrix games: Predictions precede decisions. Supporting structure was absent when participants were not explicitly asked to predict the opponent's move.

Payoff Combinations The trivial games described earlier demonstrated that payoff combination determines which order of reasoning is required. A lot of the total number of combinations (i.e., 4! x 4! = 576) do not require second-order reasoning, or yield the same response for second-order reasoning and other strategies (e.g., first-order reasoning). We had to exclude these.

Combinations of which Player I's payoff in cell A was a 1 or a 4 were not included as stimuli, because zeroth-order reasoning would suffice. It is obvious that Player I should continue the game in the former case and stop in the latter. The game in Figure 1c is an example of a game in which Player I should decide to stop in cell A.

Of the remaining payoff combinations, we excluded the trivial ones in which Player II's payoffs in cells C and D were both either lower or higher than Player II's payoff in cell B. Figure 1b depicts an example of such a game: Player II does not need to reason about Player I's decision, as Player II's payoffs in cells C and D are both more preferable than Player II's payoff in cell B.

We also had to exclude payoff combinations that yield the same decision for a zeroth- and first-order Player II, as these would yield the same prediction (of Player II's move) for a first- and second-order Player I. Figure 1d depicts an example of such a game: a Zeroth-order Player II would continue the game because the Player II's payoff in cell C is higher than in cell B. A first-order Player II would also continue the game to cell C because Player I should stop the game in that cell. The prediction of Player II's move would be the same for a first- and second-order Player I.

Besides payoff combinations such as in the game in Figure 1d, we excluded payoff combinations that yield the same decision for a first- and a second-order Player I (e.g., Figure 1e). Hedden and Zhang did not exclude these payoff

[1] Matrix games and Marble Drop games are game-theoretically equivalent because they have the same extensive form (Osborne & Rubinstein, 1994), namely that of the Centipede game (Rosenthal, 1981). See http://www.ai.rug.nl/~leendert/Equivalence.pdf for an informal proof.

combinations as long as the prediction of Player II's move would be opposite for a first- and second-order Player II. However, due to our manipulation of the procedure of asking for social reasoning, half of the participants were not explicitly asked to predict what the opponent's next possible decision would be.

In line with Hedden and Zhang, we distinguished between so-called 2- and 3-starting games with payoff combinations of which Player I's payoff in cell A was either a 2 or a 3, respectively. For the final set of stimuli, we double-balanced for both the number of *stop* and *continue* decisions of Player I and the number of *stop* and *continue* decisions of Player II. As this was not possible for the 2-starting games, we excluded those. That left us with 16 unique payoff combinations, all 3-starting games, to present during the experiment.

Procedure

The experiment consisted of three blocks: one training block and two test blocks. In the training block, we familiarized participants with the game. They were assigned to instruction and training having or lacking supporting structure. This was counterbalanced between participants. Each game was played until the participant or computer decided to stop, or until the last decision was made. After each game in the training block, participants were presented feedback that indicated whether the decision was "correct" or "incorrect".

In Test Block 1, participants were presented with second-order games, and had to decide what to do in cell A. After entering a decision, the game stopped immediately and feedback was presented. Feedback indicated whether the decision was "correct" or "incorrect". Participants assigned to the condition *asking with supporting structure* were first explicitly asked to enter their prediction of the opponent's move in cell B, before making a decision in cell A. Feedback was presented after entering a prediction. This block consisted of 32 trials; each payoff combination was presented twice. The items were presented randomly.

Test Block 2 was similar to Test Block 1 except that participants assigned to the condition *asking with supporting structure* were not explicitly asked to predict the opponent's move anymore. This block also consisted of 32 trials.

Results

Accuracy of Decisions

To account for random effects of individuals and payoff combinations, we performed linear mixed-effects (LMEs) analyses (Baayen, Davidson, & Bates, 2008). Our analysis of participants' decisions consisted of a logistic LME model that included the fixed factors *training*, *asking*, and *block* and random effects of *participants* and *payoff combinations*. The mean accuracy of the decisions is depicted in Figure 3.

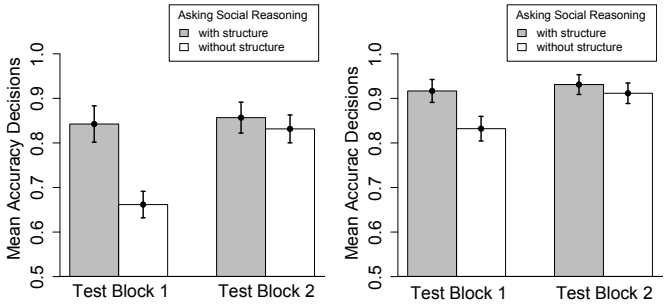


Figure 3: Mean proportions of decision scores in Test Block 1 and 2, depicted separately for participants that were explicitly asked to predict the opponent's move (grey bars) and those that were not (white bars), and separately for participants that were assigned to training without supporting structure (a) and participants assigned to training with supporting structure (b). The error bars depict standard errors.

Scaffolding (i.e., supporting structure) in training had a positive effect. Participants that were assigned to training without scaffolding significantly performed worse than participants that were assigned to training with scaffolding: $\beta = -1.23$, $z = -3.34$, $p < .001$. Mean accuracy was 80% (SE = 1.9%) in the former group, 89.8% (SE = 1.3%) in the latter.

Scaffolding in the procedure of asking for social reasoning also had a positive effect. The probability of making a correct decision was significantly higher if participants that had already predicted the opponent's move: $\beta = 1.28$, $z = 3.32$, $p < .001$. Mean accuracy was 88.7% (SE = 1.6%) for these participants. Mean accuracy was 81.1% (SE = 1.7%) for participants that were not explicitly asked to predict the opponent's move.

The probability of making a correct decision significantly increased over block: $\beta = 1.03$, $z = 6.55$, $p < .0001$. This effect was mainly due to learning of participants that were not explicitly asked to predict the opponent's move, which was reflected in a significant interaction between the factors *asking* and *block*: $\beta = -.90$, $z = -5.14$, $p < .0001$. Figure 3 shows that the difference in performance between participants that were asked to predict the opponent's move and those that were not became smaller in Test block 2.

The effects of scaffolding in the procedure of asking for social reasoning did not (significantly) differ between participants assigned to training with scaffolding and those assigned to training without scaffolding: $\beta = .25$, $z = .49$, $p = .63$.

Accuracy of Predictions

Hedden and Zhang's (2002) analyses mainly focused on participants' predictions. We analyzed accuracy of decisions, because each participant had made decisions whereas only half of the participants were explicitly asked to predict the opponent's move (i.e., in the condition *asking with scaffolding*). To make informal comparisons with

Hedden and Zhang, we analyzed the predictions of participants in the condition *asking with scaffolding* (in Test Block 1) in more detail.

Figure 4 shows an increase in performance that is qualitatively different from the gradual and linear increase that Hedden and Zhang observed (from 25% to approximately 70%).

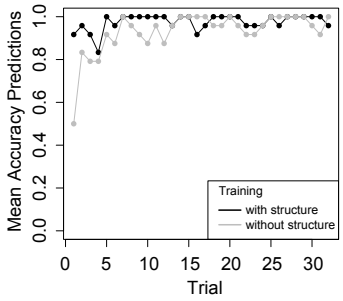


Figure 4: Mean accuracy of predictions in Test block 1, plotted separately for participants assigned to training without supporting structure (grey line) and training with supporting structure (black line).

We fitted a logistic LME that consisted of *training* as a fixed factor, the logarithm of *trial* as a covariate, and random effects of *participants* and *payoff combinations*.

Supporting structure in training had a positive effect on the accuracy of predictions. Participants assigned to training without supporting structure performed significantly worse than participants assigned to training with supporting structure: $\beta = -2.49$, $z = -3.36$, $p < .001$.

The effect of trial was significant: $\beta = .69$, $z = 2.90$, $p < .01$. The probability of correctly predicting the opponent's move increased with time. This effect was stronger for participants that were assigned to training without supporting structure instead of training with supporting structure: $\beta = .72$, $z = 2.43$, $p < .05$. This finding supports our hypothesis that training without supporting structure, which consisted of trivial games, biased participants to apply first-order reasoning. As first-order reasoning did not yield correct decisions during the test blocks, participants had to unlearn this strategy.

Discussion

We investigated effects of scaffolding (i.e., supporting structure) in instruction, training, and procedure of asking for social reasoning. We hypothesized that scaffolding would facilitate social reasoning by structuring the interplay of serial and concurrent cognitive processes that underlie social reasoning. The results corroborated our hypotheses.

First of all, the participants successfully applied second-order reasoning in a large proportion of the games, especially if supporting structure was provided in both training and procedure of asking for social reasoning. Mean accuracy was 92% (SE = 1.7%) in those conditions. In contrast, mean accuracy in Hedden and Zhang's (2002) matrix games started at approximately 25% and increased to approximately 65%, which is not far above chance level.

Supporting structure in training had a positive effect on social reasoning in matrix games. We assigned half of the participants to training similar to Hedden and Zhang's training, which consisted solely of trivial (second-order) games that allowed for both first- and second-order reasoning. We assigned the other half to training in which we provided supporting structure. Supporting structure consisted of zeroth-, first-, and second-order games that unambiguously required reasoning of corresponding orders. We hypothesized that participants assigned to training that lacked supporting structure preferred the simpler and less demanding first-order reasoning strategy over second-order reasoning and erroneously tried using that during the test blocks. Our results corroborated this hypothesis. Over the entire experiment, the probability of making a correct decision was higher for participants assigned to training with supporting structure instead of training that lacked supporting structure.

Besides a positive effect on performance of supporting structure in training, we found a positive effect of supporting structure in the procedure of asking for social reasoning. Supporting structure closely mapped the reasoning steps required in matrix games with the responses asked for. Participants that were asked to predict the opponent's move before making a decision had a higher probability of making a correct decision than participants that were not asked to make predictions. This finding corroborated our hypothesis that performance improves by providing supporting structure in the procedure of asking for social reasoning.

Supporting structure in training and in the procedure of asking for social reasoning both had positive effects on social reasoning. However, there was no interaction. Given the disadvantageous effect of missing supporting structure during training, one might expect that participants in this condition would benefit more from supporting structure in the procedure of asking for second-order reasoning than participants assigned to training with supporting structure. The results did not corroborate this idea, as the interaction between *training* and *asking* was not significant: Whichever training participants were assigned to, the probability of a correct decision was greater for participants that were provided with supporting structure in the procedure of asking for social reasoning.

The probability of correctly deciding whether to stop or continue a game increased over block. However, this was mainly due to participants that were not explicitly asked to predict the opponent's move, as the difference with participants that were asked to make a prediction became smaller (Figure 3). In other words, participants that were asked to predict the opponent's move initially benefitted from supporting structure that closely mapped the reasoning steps required by matrix games with the responses asked for, but eventually participants that were not asked to make predictions were catching up.

Hedden and Zhang analyzed changes over time in second-order reasoning and found profound learning effects. In line

with their analyses, we analyzed predictions as a function of trial, in addition to the effect of supporting structure in training and procedure of asking for social reasoning. The analyses showed that the probability of correctly predicting the opponent's move increased over time, which implies a positive relation between proficiency in applying higher-order theory of mind and experience with social reasoning in matrix games. This effect was stronger for participants assigned to training without supporting structure than for participants assigned to training with supporting structure. This finding corroborated our hypothesis that training without supporting structure, which consisted solely of trivial games, biased participants to apply first-order reasoning. They had to unlearn that strategy during the test blocks, resulting in a gradual increase in performance.

In sum, we found effects of supporting structure in instruction, training and procedure of asking for social reasoning. Participants that were assigned to training that provided supporting structure performed better during the experiment than participants assigned to training that lacked supporting structure. Also, participants that were asked to predict the opponent's move were better at making a decision than participants that were not asked to make predictions. These effects were more pronounced in the first of two test blocks.

General Conclusions

We found that applying higher-order theory of mind in strategic games is not too difficult to learn as long as it is introduced appropriately by providing step-by-step instruction and training. Moreover, higher-order theory of mind improved by explicitly asking participants to predict the opponent's possible future behavior. Participants learned to play matrix games, and learned to play them proficiently.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

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.

Goodie, A. S., Doshi, P., & Young, D. L. (2010). Levels of theory of mind reasoning in competitive games. *Journal of Behavioral Decision Making, n/a*, doi: 10.1002/bdm.717.

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

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25-41.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica, 60*(4), 803-836.

Meijering, B., Van Maanen, L., Van Rijn, H., & Verbrugge, R. (2010). The facilitative effect of context on second-order social reasoning. In R. Catrambone & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, Cognitive Science Society, Austin (TX), 2010, pp. 1423-1428

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge, Massachusetts: The MIT press.

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, 39*, 437 – 471.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515-526.

Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory, 25*(1), 92-100.

Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic, 38*, 649 – 680.

Verbrugge, R., & Mol, L. (2008). Learning to apply theory of mind. *Journal of Logic, Language and Information, 17*(4), 489-511.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128.