CHAPTER 9

Discussion and conclusion

Experimental evidence shows that human participants make use of higher-order theory of mind in order to understand false belief stories and to play strategic games. However, it is controversial whether there is any other species fully capable of reasoning about the minds of others, let alone whether any non-human is capable of using theory of mind recursively. Over the previous chapters, we have constructed an agent model to investigate in which kinds of settings reasoning at higher orders of theory of mind can be shown to be beneficial for agents. This agent model allows us to determine the degree to which specific settings may have contributed to the emergence of higher-order theory of mind. Based on the hypotheses for the function of theory of mind outlined in Section 1.5, we have investigated a range of different types of scenarios, from competitive games such as rock-paper-scissors (Chapter 2) and cooperative games such as the Tacit Communication Game (Chapter 4) to the negotiation game Colored Trails (Chapter 5 and Chapter 6). In Section 9.1, we will give a brief summary of our main findings. In Section 9.2, we discuss what these findings tell us about the emergence and the function of higher-order of mind.

9.1 Summary

In Section 1.5, we outlined three specializations of the social brain hypothesis that attempt to explain the emergence of higher-order theory of mind. Each of these hypotheses points to a different setting in which higher-order theory of mind may be especially useful. Throughout this thesis, we have made use of an agent-based model to simulate social interactions in each of these settings to determine the plausibility of these hypotheses. The main findings of these simulations are listed in Table 9.1, which shows the additional advantage of reasoning at increasingly higher orders of theory of mind for the Machiavellian intelligence hypothesis, the Vygotksian intelligence hypothesis, and the mixed-motive interaction hypothesis. In this section, we provide a more detailed summary of these findings.

	Social brain hypotheses		
	Machiavellian	Vygotskian	Mixed-motive
Additional advantage of orders of theory of mind			
First-order	++	++	++
Second-order	++	+	++
Third-order	+	-	-
Fourth-order	-	-	+

Table 9.1: Summary of our main findings concerning the Machiavellian intelligence hypothesis, the Vygotskian intelligence hypothesis, and the mixed-motive interaction hypothesis. For each of these specializations of the social brain hypothesis, the table shows the additional advantage of reasoning using kth-order theory of mind compared to reasoning using $(k-1)$st-order theory of mind.

9.1.1 The Machiavellian intelligence hypothesis

According to the Machiavellian intelligence hypothesis (Byrne & Whiten, 1988; Whiten & Byrne, 1997), competition is the main driving force behind the emergence of higher-order theory of mind. Higher orders of theory of mind allow an individual a higher proficiency in deception and social manipulation. The ones with the highest social intelligence are therefore expected to outperform competitors of a lower order of theory of mind in competitive settings.

In Chapter 2, we investigated this hypothesis using our agent-based model of theory of mind. We simulated interactions between agents of different orders of theory of mind across four different games, which include repeated single-shot games such as rock-paper-scissors, but also repeated extensive form games like limited bidding, in which agents sequentially make multiple decisions within each game. Across these games, we find a similar pattern of diminishing returns for reasoning at higher orders of theory of mind, which is summarized in Table 9.1. Agents that made use of first-order theory of mind clearly outperform competitors that are restricted to associative learning techniques. Similarly, agents that can reason using second-order theory of mind, and that are therefore capable of reasoning about the theory of mind of their competitor, also clearly outperform first-order theory of mind agents. In comparison, the additional advantage of deeper recursion to third-order is limited, while fourth-order theory of mind appears to have no additional benefits.

Our agent-based simulations support the Machiavellian intelligence hypothesis by showing that there are strictly competitive settings in which agents can benefit from the use of higher-order theory of mind. Additionally, in Chapter 3, we find that characteristic behavior of human participants in an n-player extension of rock-paper-scissors known as the Mod game (Frey & Goldstone, 2013; Frey,

2013) is consistent with higher-order theory of mind reasoning. That is, in these strictly competitive settings in which agents can benefit from the use of higher-order theory of mind, human participants also appear to engage in higher-order theory of mind reasoning.

Chapter 2 shows that there are competitive settings in which reasoning at higher orders of theory of mind is advantageous. In particular, these settings include a specific class of games in which there is no pure strategy Nash equilibrium, which zero-order theory of mind agents would eventually learn to play. In these settings, zero-order theory of mind agents continuously change their behavior. Higher-order theory of mind agents can benefit from this behavior by anticipating the way lower-order theory of mind agents change their behavior in response to the observed actions of others. This means that the advantage of higher-order theory of mind in these settings depends on the inability of zero-order theory of mind agents to play according to a mixed strategy. Interestingly, experimental evidence suggests that human participants are indeed poor at generating random sequences (Wagenaar, 1972; Rapoport & Budescu, 1997).

9.1.2 The Vygotskian intelligence hypothesis

The Vygotskian intelligence hypothesis (Vygotsky, 1978) claims that rather than competition, cooperation is the main driving force behind the emergence of higher-order theory of mind in humans, and cooperative social interactions play a vital role in the development of theory of mind in children. According to the Vygotskian intelligence hypothesis, higher-order theory of mind allows groups of individuals a higher flexibility to organize cooperation through shared intentionality (Tomasello et al., 2005; Tomasello, 2009), and achieve higher levels of cooperation than those that can be obtained by individuals that are unable to reason about the mental content of their group partners.

To put the Vygotskian intelligence hypothesis to the test, we simulated interactions among agents of different orders of theory of mind in the Tacit Communication Game (De Ruiter et al., 2007; Newman-Norlund et al., 2009; Blokpoel et al., 2012). In the Tacit Communication Game, two players need to set up a novel communication protocol that allows one of the players, called the Sender, to communicate the common goal to his partner, the Receiver. Human participants are known to be highly proficient in completing this cooperative communication task (De Ruiter et al., 2007). In Chapter 4, our agent-based simulations show that agents can achieve a cooperative solution more efficiently through the use of first-order theory of mind, compared to a situation in which both agents are limited to zero-order theory of mind. In addition, second-order theory of mind can sometimes help agents to achieve a cooperative solution even faster than when both agents are limited to first-order theory of mind reasoning.

Our findings suggest that participants make use of theory of mind while playing the Tacit Communication Game, but we also found evidence that theory of mind alone is not enough to explain human performance. When a pair of human participants play the Tacit Communication Game, the Sender determines the meaning of the messages he decides to send. If a trial fails, the Sender tries to clarify his messages by emphasizing the part of the message that is meant to communicate the common goal. That is, according to the Sender, a trial fails because the Receiver incorrectly interpreted the message. In contrast, our simulations show that a cooperative solution is found more efficiently if the Receiver decides on the way messages should be interpreted. Whenever the Sender sends a message, the Receiver decides how that message will be interpreted from that moment on, irrespective of the outcome of the trial. In this case, a trial fails because the Sender failed to send the correct message. This difference between human performance and theory of mind agents suggests that participants rely on more than theory of mind alone.

Our results also show that in some cases, theory of mind reasoning can be detrimental to the efficiency of the cooperative effort in the Tacit Communication Game. In particular, when both players reason at the same order of theory of mind, each of them holds incorrect beliefs about the mental content of the other, which hinders the cooperative effort. Moreover, while higher-order theory of mind helps agents to achieve a cooperative solution more quickly, agents no longer benefit from the use of theory of mind once such a cooperative solution has been found.

9.1.3 The mixed-motive interaction hypothesis

The third hypothesis for the emergence of higher-order theory of mind in humans that we investigated is the mixed-motive interaction hypothesis (Verbrugge, 2009). According to this hypothesis, the main contributors to the emergence of higher-order theory of mind are neither purely competitive nor purely cooperative settings, but rather settings in which both cooperation and competition play a role. In these so-called mixed-motive settings, individuals need to balance competitive aspects with cooperative goals. In such settings, higher-order theory of mind may be especially useful.

We test the mixed-motive interaction hypothesis in a negotiation setting known as Colored Trails (Lin et al., 2008; Gal et al., 2010; De Jong et al., 2011; Van Wissen et al., 2012). This setting is a prototypical multi-issue bargaining setting, in which various aspects of the process of negotiation can be investigated. In Chapter 5, we consider one-shot negotiations, in which the cooperative and competitive aspects of the negotiation process were represented by two different agents. In this setting, we find that both first-order and second-order theory of mind allow agents to achieve better negotiation outcomes, both in terms of their own perfor-

mance as well as the performance of the group as a whole. That is, theory of mind allows agents to not only increase their own piece of the pie, but their efforts also increase the size of the pie as a whole. When both cooperative and competitive aspects play a role, reasoning at third-order theory of mind does not yield benefits over second-order theory of mind, but fourth-order theory of mind allows agents to obtain additional advantages over lower-order theory of mind agents. These results are summarized in Table 9.1.

We find the same pattern of advantages for increasingly higher orders of theory of mind across different specifications of the zero-order theory of mind agents. Additionally, our results show that the adaptive learning ability of zero-order theory of mind agents relevantly influences the way agents benefit from reasoning at higher orders of theory of mind. When zero-order theory of mind agents adapt their behavior in response to the actions of others, fourth-order theory of mind agents experience a competitive advantage over lower-order theory of mind agents. In contrast, when zero-order theory of mind agents follow a static strategy that does not change in response to the actions of others, this competitive advantage for fourth-order theory of mind is shifted up to fifth-order theory of mind.

In Chapter 6, we investigate a more dynamic Colored Trails setting, in which two agents alternate in making offers until an agreement is reached. We find that when agents have the opportunity to learn from their trading partner over multiple rounds of offers, agents without theory of mind are in principle capable of obtaining the same mutually beneficial outcome as more sophisticated agents. However, such zero-order theory of mind agents fail to balance competitive opportunities for a higher personal gain against the need for cooperation. In this scenario, natural selection favors the agents that make relatively few compromises to their position. This eventually leads to a situation where agents no longer negotiate, but insist on the outcome that maximizes their personal payoff. Although first-order theory of mind allows agents to prevent a complete breakdown in negotiation, such agents still experience evolutionary pressure to attempt to increase their own piece of pie at the expense of their trading partner. However, when both agents engage in this behavior, they end up reducing the size of the pie as a whole. Interestingly, second-order theory of mind allows agents to balance cooperation and competition.

In Chapter 7, we let computational theory of mind agents interact with human participants directly in the dynamic negotiation setting of Chapter 6. In interaction with the theory of mind agents, participants showed spontaneous use of theory of mind reasoning. Moreover, when paired with a second-order theory of mind agent, participants' offers were more consistent with second-order theory of mind reasoning than they were with reasoning at lower orders of theory of mind.

	Social brain hypotheses		
	Machiavellian	Vygotskian	Mixed-motive
Match with empirical findings			
Human theory of mind abilities	match	mismatch	match
Lack of non-human theory of mind	mismatch	mismatch	match

Table 9.2: Summary of the match of the theory of mind abilities of humans and non-human species to the predictions of the Machiavellian intelligence hypothesis, the Vygotskian intelligence hypothesis, and the mixed-motive interaction hypothesis.

9.2 Conclusion: The function of higher-order theory of mind

In this thesis, we have constructed agent-based models to understand the function of higher-order theory of mind. To this end, we have modeled theory of mind agents in a variety of settings to determine to what extent the use of higher-order theory of mind can be beneficial to agents. Settings in which individuals can benefit greatly from the use of higher-order theory of mind are more likely to have contributed to the emergence of this cognitively demanding ability in humans. Our main findings concerning the effectiveness of higher-order theory of mind are listed in Table 9.1. Table 9.2 summarizes how the predictions of each of the three specializations of the social brain hypothesis match up to empirical findings concerning the theory of mind abilities of humans and non-human species. In this section, we discuss what these findings can tell us about the emergence and the function of higher-order theory of mind.

Chapter 2 shows that there are competitive settings in which agents experience diminishing returns to the use of higher-order theory of mind. In such settings, agents can benefit greatly from reasoning using second-order theory of mind, while the additional advantage of deeper recursion to even higher orders of theory of mind is limited. These results coincide with empirical evidence that suggests that the behavior of human participants in these competitive settings is more consistent with the application of second-order theory of mind than with the application of lower orders of theory of mind (Devaine et al., 2014a, also see Chapter 3 of this thesis).

However, it seems unlikely that these strictly competitive settings are the main driving force behind the emergence of higher-order theory of mind. Compared to other species of animals, humans are considered to be a cooperative species (Tomasello, 2009; Burkart et al., 2014). This is particularly true when humans are compared to other great apes, such as chimpanzees. The Machiavellian intelligence hypothesis predicts that species that are characterized as being more competitive

can benefit more from the use of higher-order theory of mind than species that are characterized as being more cooperative. However, only the hyper-cooperative human species makes use of higher-order theory of mind. That is, as Table 9.2 suggests, there is a mismatch between predictions of the Machiavellian intelligence hypothesis and the theory of mind abilities of non-human species.

Competitive settings in which agents have been shown to benefit from higher-order theory of mind reasoning rely on the inability of zero-order theory of mind agents to play randomly. Interestingly, although experimental evidence shows that human participants are poor at generating random sequences (Wagenaar, 1972; Rapoport & Budescu, 1997), chimpanzees do appear to be able to learn to play according to a mixed strategy Nash equilibrium (Martin, Bhui, Bossaerts, Matsuzawa, & Camerer, 2014).

Chapter 4 shows that agents benefit from the use of higher-order theory of mind in the Tacit Communication Game. Through the use of higher-order theory of mind, agents were able to achieve a cooperative solution more quickly. However, cooperative settings may not be the main driving force behind the emergence of higher-order theory of mind. Many non-human species engage in cooperative activities without a need for theory of mind. For example, marmoset monkeys altruistically offer food to unrelated individuals (Burkart, Fehr, Efferson, & van Schaik, 2007), and cooperation even occurs among microorganisms (Crespi, 2001). Simulation studies confirm that many forms of cooperation can emerge and be maintained using simple mechanisms, without the need for a cognitively demanding ability like theory of mind (Boyd & Richerson, 1992; Boyd et al., 2003; Nowak, 2006; Sigmund, 2010; De Weerd & Verbrugge, 2011; Gärdenfors, 2012; Van der Post et al., 2013, 2015). In addition, our results show that while higher-order theory of mind can help agents in finding a cooperative solution, the need for theory of mind seems to disappear once such a solution has been found. As a result, Table 9.2 shows that the predictions of the Vygotskian intelligence hypothesis do not seem to match the abilities of human and non-human species.

Of the three types of settings we have considered in this work, mixed-motive settings are more likely to be the main contributor to the emergence of higher-order theory of mind in humans than either purely competitive or purely cooperative settings. Chapter 5 shows that higher-order theory of mind reasoning allows agents in one-shot negotiations to obtain a better outcome, both in terms of their own payoff as well as the payoff of their trading partner. Chapter 6 furthermore shows that when agents have an opportunity to learn from their trading partner, higher-order theory of mind allows agents to balance competitive and cooperative aspects of the game and reach a mutually beneficial solution. Agents without theory of mind try to increase their personal gain without regard for the goals of others. This presents zero-order theory of mind agents with an environment similar to the prisoner's dilemma, in which natural selection favors the agents that defect on cooperation. Even though zero-order theory of mind agents could coop-

erate to obtain a better result, the process of natural selection eventually leads to a breakdown of negotiation in the population. Higher-order theory of mind allows agents to recognize the need for cooperation and prevent this breakdown. This suggests that higher-order theory of mind is not just beneficial, but may even be essential in mixed-motive situations. As Table 9.2 suggests, this may explain why humans make use of higher-order theory of mind, while no other species appears to be fully capable of reasoning about the minds of others.

The main findings in this thesis show that agents can benefit from the use of higher-order theory of mind across competitive, cooperative, and mixed-motive settings. Interestingly, the way in which higher-order theory of mind benefits these agents is slightly different in each setting. In competitive settings, agents benefit from theory of mind when zero-order theory of mind agents do not exhibit equilibrium behavior. That is, zero-order theory of mind agents continue to adjust their behavior in response to the actions of others. Reasoning at increasingly higher orders of theory of mind allows agents to predict these changes in the behavior of lower-order theory of mind agents and act accordingly.

In cooperative settings, on the other hand, higher-order theory of mind can help agents to reach equilibrium behavior more quickly. By reasoning about the way a lower-order theory of mind agent learns from its observations, a higher-order theory of mind agent can predict the outcome of this learning process and reach a stable cooperative solution more quickly. Once this equilibrium has been reached, however, theory of mind is no longer needed.

Finally, in mixed-motive settings, higher-order theory of mind can actively prevent the selection of an undesired equilibrium behavior. In mixed-motive settings, lower-order theory of mind agents may experience an incentive to increase their own gain at the expense of others. Higher-order theory of mind agents can reason about these incentives and adjust their behavior to try and remove them, similar to the way a strategy like tit-for-tat can prevent mutual defection in the prisoner's dilemma.

Our results show that depending on the specific setting, higher-order theory of mind can benefit agents in different ways. The effectiveness of theory of mind is therefore not limited to a specific type of setting. While there may be a specific type of setting that has been the main contributor to the emergence of higher-order theory of mind in humans, theory of mind also allows an agent to learn quickly across different settings. In Chapter 5, we have shown that theory of mind allows agents a convenient way to generalize across different scenarios, which allows them to make meaningful predictions of the behavior of other agents in unfamiliar settings (cf. Robalino & Robson, 2012; Monte, Robalino, & Robson, 2012). Similarly, Chapter 4 shows that theory of mind allows agents to arrive at a cooperative solution more efficiently. The main driving force behind higher-order theory of mind could therefore be more general, as the social brain hypothesis suggests. Higher-order theory of mind could be needed to deal with social com-

plexity associated with a dynamic environment in which agents find themselves in many different settings that are sometimes cooperative, sometimes competitive, and sometimes involve mixed motives.