The effectiveness of higher-order theory of mind in negotiations

Harmen de Weerd[1], Rineke Verbrugge[1], Bart Verheij[1,2]

[1] Institute of Artificial Intelligence, University of Groningen, The Netherlands
[2] CodeX, Stanford University, United States

1 Introduction

When the outcome of a decision you make depends on the actions of others, it is important to be able to predict those actions. To facilitate this process, people reason about unobservable mental content of others, such as beliefs, desires, and intentions. People can also use this so-called *theory of mind* recursively, and reason about the way others make use of theory of mind. For example, to understand a sentence such as 'Alice *believes* that Bob *knows* that Carol is throwing him a surprise party', the reader has to use second-order theory of mind, by reasoning about the way Alice reasons about Bob's knowledge.

Behavioral experiments have demonstrated that people make use of higher-order (i.e. at least second-order) theory of mind [1, 2]. However, the extent to which non-human species are able to use theory of mind of any kind is under debate [3, 4]. The human ability to make use of higher-order theory of mind suggests that there may be settings in which this ability provides individuals with enough of an evolutionary advantage to support the emergence of reasoning about the minds of others, and even to use this ability recursively. One possible explanation is that higher-order theory of mind is needed to engage effectively in *mixed-motive interactions* [5] such as negotiation. Mixed-motive interactions involve partially overlapping goals, so that these interactions are neither fully cooperative nor fully competitive. In this paper, we make use of agent-based computational models to determine whether the use of higher orders of theory of mind allows agents to reach better outcomes in negotiation, both in terms of individual agent performance as well as in terms of social welfare.

2 Colored Trails

We study the effect of higher-order theory of mind in a particular negotiation game known as Colored Trails, a test-bed introduced by Barbara Grosz, Sarit Kraus and colleagues to investigate various aspects of negotiations [6, 7][1]. In our setup, the game is played by three players on a square board consisting of 25 colored tiles, as depicted in Figure 1. The three players, i, j, and r, are initially located at starting location S and want to end up as close as possible to their

[1] Also see https://coloredtrails.atlassian.net/wiki/display/coloredtrailshome/.

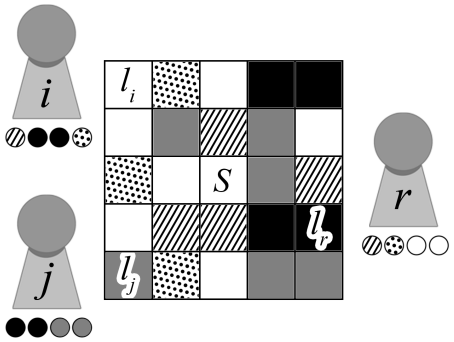

Fig. 1: Colored Trails is played by three agents. These agents start at starting location S and are assigned a random goal location. In this example, agents i, j, and r have goal locations l_i, l_j, and l_r respectively.

own goal location, l_i, l_j, and l_r respectively. Each player also receives a set of four colored chips (depicted as small circles in Figure 1), selected randomly from the same five possible colors as those on the board. These chips are used to move around on the board. Players may move to a tile adjacent to their current location by handing in a chip of the same color as the destination tile. For example, a player could move from starting tile S in Figure 1 to location l_r by handing in one striped chip and two black chips.

A player's score depends on how closely he approaches his goal location. A player receives 10 points for each step he takes towards his goal. Reaching the goal location yields an additional 50 points. Finally, any chip that has not been used to move around the board is worth an additional 5 points to its owner. Players are thus highly incentivized to reach their goal location, but they also compete over control of unused chips.

To get closer to their goals, players are allowed to trade chips. This trading takes the form of a one-shot bargaining game. Two agents i and j are assigned the role of *allocator*, while the third agent r is assigned the role of *responder*. The two allocators simultaneously choose an offer to make to the responder. An allocator suggests to trade any given subset of his own chips against any given subset of the responder's chips. The responder then accepts the offer that yields her the highest score. However, if both allocators have made an offer that would reduce her score, the responder rejects both offers and the initial distribution of chips becomes final.

3 Theory of mind

In our Colored Trails setup, the role of the responder is limited to selecting the offer that benefits her the most. We therefore focus on the theory of mind abilities of the allocators. A zero-order theory of mind (ToM_0) allocator is unable

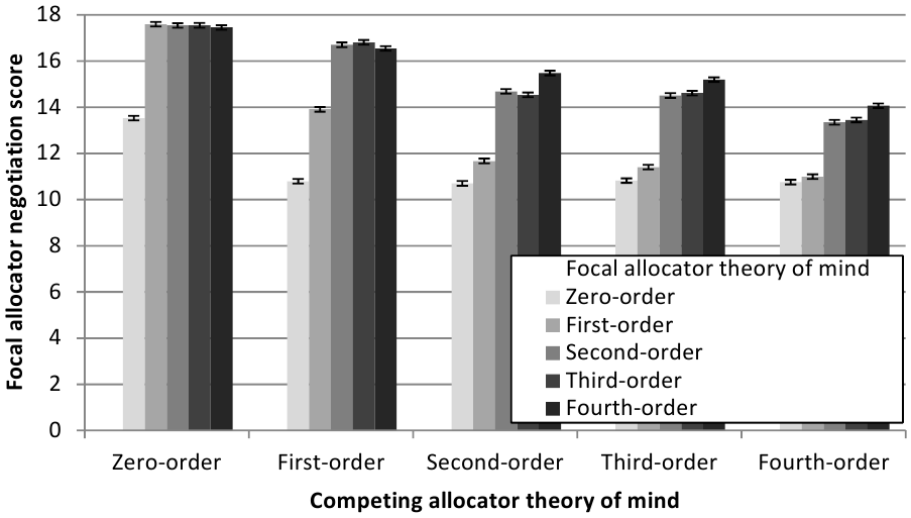

Fig. 2: Average negotiation score of a focal allocator in the Colored Trails game. Results are shown for different combinations of levels of sophistication of both allocators. Brackets indicate standard errors.

to reason about the goal of his trading partner. Instead, the ToM_0 allocator estimates the probability that his offer will be accepted based on how successful this offer has been in the past.

The first-order theory of mind (ToM_1) allocator can use theory of mind to put himself in the position of other agents and simulate their decision-making processes. By putting himself in the position of the responder, a ToM_1 allocator understands that the responder will reject any offer that would reduce her score. Similarly, by placing himself in the position of the competing allocator, a ToM_1 allocator can predict what offer his competitor is going to make. The ToM_1 allocator can use this information when making an offer himself.

For increasingly higher orders of theory of mind, a kth-order theory of mind (ToM_k) allocator considers the possibility of increasingly more sophisticated competitors. However, a ToM_k allocator retains the ability to reason at orders of theory of below the kth. For example, through repeated interaction with the same competitor, a ToM_6 allocator may come to believe that the competing allocator is a ToM_1 agent, so the ToM_6 allocator may choose to behave as if he himself were a ToM_2 agent.

4 Results

We performed simulations in which the theory of mind agents described in the previous section played repeated one-shot Colored Trails games. Each new game

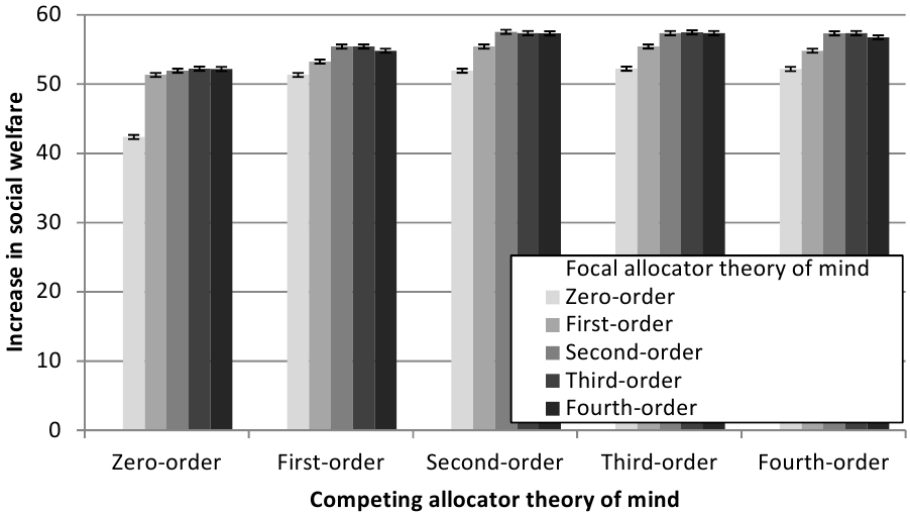

Fig. 3: Average increase in welfare, measured as the sum of the scores of all three agents. Results are shown for different combinations of levels of sophistication of both allocators. Brackets indicate standard errors.

was played on a a different board in terms of coloring and goal locations and with different sets of initial chips.

Figure 2 shows the average performance of a focal ToM_i allocator in the presence of a competing ToM_j allocator, which is calculated as the average difference between an agent's score after the end of a negotiation and his initial score at the start of negotiation. It turns out that even though ToM_0 allocators can learn to negotiate effectively, ToM_1 allocators outperform ToM_0 allocators, irrespective of the theory of mind abilities of the competing allocator. Similarly, ToM_2 allocators outperform ToM_1 allocators when the competing allocator uses theory of mind. We find no additional benefit for third-order theory of mind. However, surprisingly, ToM_4 allocators outperform lower-order agents when the competing allocator can use second-order theory of mind.

Figure 3 shows that the presence of ToM_1 allocators and ToM_2 allocators also increases social welfare, as measured by the sum of the negotiation scores of all three agents. Even higher orders of theory of mind were found not to influence social welfare any further. Interestingly, even though theory of mind agents act purely in their own interest, this increase in social welfare is not completely explained by increase in the score of the allocator; the score of the responder increases as well. It would be interesting to also investigate alternative notions of social welfare (see for example [8]).

5 Conclusion

Our results in the Colored Trails game show that there are mixed-motive settings in which the ability to make use of theory of mind allows individuals to reach better outcomes. We find that both first-order and second-order theory of mind allows agents to obtain a better score, but also to obtain a better score for their trading partner. Although we find no additional advantages for third-order theory of mind, we find that fourth-order theory of mind provides agents with a competitive edge. Interestingly, we did not find a competitive benefit for fourth-order theory of mind in strictly competitive settings [9]. This suggests that theory of mind may be more important for dealing with mixed-motive settings than it is in competitive settings.

Acknowledgments

This work was supported by the Netherlands Organisation for Scientific Research (NWO) Vici grant NWO 277-80-001, awarded to Rineke Verbrugge.

References

1. Perner, J., Wimmer, H.: "John thinks that Mary thinks that...". Attribution of second-order beliefs by 5 to 10 year old children. Journal of Experimental Child Psychology **39**(3) (1985) 437–71
2. Hedden, T., Zhang, J.: What do you think I think you think?: Strategic reasoning in matrix games. Cognition **85**(1) (2002) 1–36
3. Penn, D., Povinelli, D.: On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. Philosophical Transactions of the Royal Society B: Biological Sciences **362**(1480) (2007) 731
4. Tomasello, M.: Why we Cooperate. MIT Press, Cambridge, MA (2009)
5. Verbrugge, R.: Logic and social cognition: The facts matter, and so do computational models. Journal of Philosophical Logic **38** (2009) 649–680
6. Grosz, B., Kraus, S., Talman, S., Stossel, B., Havlin, M.: The influence of social dependencies on decision-making: Initial investigations with a new game. In: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems. Volume 2., IEEE Computer Society (2004) 782–789
7. Gal, Y., Grosz, B., Kraus, S., Pfeffer, A., Shieber, S.: Agent decision-making in open mixed networks. Artificial Intelligence **174**(18) (2010) 1460–1480
8. d'Aspremont, C., Gevers, L.: Social welfare functionals and interpersonal comparability. In Arrow, K.J., Sen, A., Suzumura, K., eds.: Handbook of social choice and welfare. North Holland (2002) 459–541
9. de Weerd, H., Verbrugge, R., Verheij, B.: How much does it help to know what she knows you know? An agent-based simulation study. Artificial Intelligence **199–200** (2013) 67–92