How much does it help to know what she knows you know? An agent-based simulation study (Abstract)

Harmen de Weerd Rineke Verbrugge Bart Verheij

Institute of Artificial Intelligence, University of Groningen

1 Introduction[1]

In everyday life, people regularly make use of *theory of mind*, by reasoning about what others know and believe. For example, we identify with characters in literature and movies, and accept that they may have beliefs and intentions different from our own. As part of a project that studies logical, psychological [4], and developmental [1] perspectives on theory of mind, and that determines to what extent complex behaviour can be explained using cognitively simple behaviour [5], in this paper we make use of agent-based simulations to explain the evolution of our ability to reason about the mental content of others.

According to the Machiavellian intelligence hypothesis, social cognition emerged because it allows individuals to use deception and manipulation to obtain a competitive advantage[2]. Following this hypothesis, we test whether higher-order theory of mind presents individuals with an advantage through agent-based simulations [3] across several competitive settings. We simulate interactions between pairs of agents that differ in their theory of mind abilities across three variations on repeated single-shot rock-paper-scissors games, as well as repeated interactions in a more complex, extensive form game named Limited Bidding, where agents need to consider how their actions affect the future of the game.

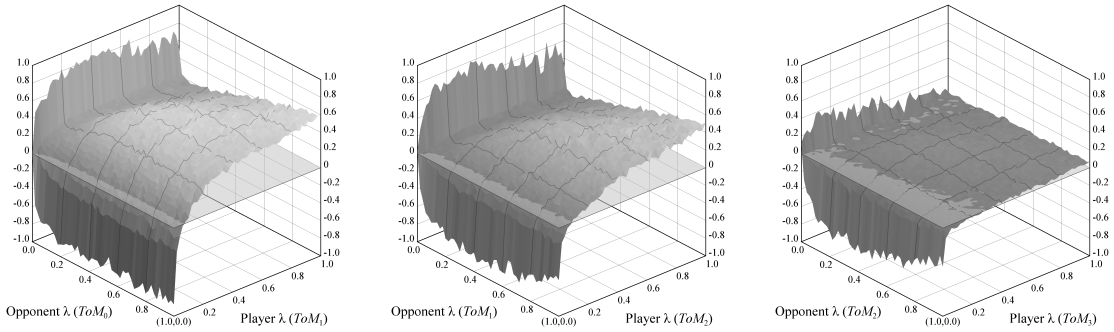
2 Theory of mind agents

In our agent model, agents without theory of mind predict their opponent's behaviour purely by analyzing her behaviour. A first-order theory of mind (ToM_1) agent can also consider the game from the perspective of his opponent, and determine what he would do himself in her position. The agent's own decision process thereby becomes a model of the decision process of the opponent that generates predictions for her behaviour. Depending on the observed accuracy of these first-order theory of mind predictions over repeated interactions with the same opponent, a ToM_1 agent can choose to either ignore or accept them.

Each additional order of theory of mind provides a theory of mind agent with an additional hypothesis of his opponent's future behaviour. For example, a second-order theory of mind (ToM_2) agent believes that his opponent may be using first-order theory of mind to predict his behaviour. By determining his own first-order theory of mind response from the position of his opponent, the ToM_2 agent obtains another prediction for her behaviour. The agent compares these hypotheses based on different orders of theory of mind with his opponent's actual behaviour to determine his own behaviour.

[1] This is an abstract based on [2].

[2] For a discussion of theories that explain the emergence of higher-order theory of mind, see [6].



(a) ToM_1 against ToM_0 (b) ToM_2 against ToM_1 (c) ToM_3 against ToM_2

Figure 1: Average performance of theory of mind agents playing Limited Bidding against opponents of a lower order of theory of mind. Each graph represents the result of 6.5 million individual simulated games.

3 Results

Across the four games that we investigate, we find a common pattern of diminishing returns on higher orders of theory of mind. Figure 1 shows the performance of theory of mind agents as a function of the speed λ at which an agent adjust his beliefs to his opponent's behaviour. The figure shows that ToM_1 agents and ToM_2 agents clearly outperform opponents that are more limited in their ability to model others. However, ToM_3 agents obtain a score close to zero, which shows that the competitive advantage for orders of theory of mind beyond the second is limited. Fourth-order theory of mind turns out to be beneficial only under specific circumstances. These results show that competition can indeed encourage the emergence of higher-order theory of mind. In future work, we aim to investigate whether theory of mind plays an important role in cooperative settings, for example in teamwork, as well as mixed-motive settings such as negotiations.

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] B. Arslan, A. Hohenberger, and R. Verbrugge. The development of second-order social cognition and its relation with complex language understanding and working memory. In N. Miyake, D. Peebles, and R.P. Cooper, editors, *Proc. 34th Annu. Conf. Cogn. Sci. Soc.*, pages 1290–1295. 2012.

[2] H. de Weerd, R. Verbrugge, and B. Verheij. How much does it help to know what she knows you know? An agent-based simulation study. *Artif. Intel.*, 199-200:67–92, 2013.

[3] J.M. Epstein. *Generative Social Science: Studies in Agent-based Computational Modeling*. Princeton University Press, Princeton (NJ), 2006.

[4] B. Meijering, H. van Rijn, N.A. Taatgen, and R. Verbrugge. I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In *Proc. 33nd Annu. Conf. Cogn. Sci. Soc.*, pages 2486–2491. 2011.

[5] E. van der Vaart, R. Verbrugge, and C.K. Hemelrijk. Corvid re-caching without theory of mind: A model. *PLoS ONE*, 7(3):e32904, 2012.

[6] R. Verbrugge. Logic and social cognition: The facts matter, and so do computational models. *J. Philos. Log.*, 38:649–680, 2009.