Estimating the use of higher-order theory of mind using computational agents[1]

Harmen de Weerd Denny Diepgrond Rineke Verbrugge

Institute of Artificial Intelligence, University of Groningen

1 Introduction

In social interactions, people often reason about the beliefs, goals, and intentions of others. People use this so-called *theory of mind* [5] to understand why others behave the way they do, as well as to predict the future behavior of others. People can even use their own theory of mind to reason about the way others make use of their theory of mind. For example, people make use of *second-order theory of mind* to understand a sentence such as "Alice *knows* that Bob *knows* that Carol is throwing him a party".

Both empirical and simulation studies suggest that the use of higher-order theory of mind may be particularly useful in simple competitive games [1, 3, 4]. In such games, however, it is difficult to distinguish between theory of mind reasoning and simpler, behavior-based strategies. In this paper, we combine a Bayesian model selection technique introduced by [7] with our theory of mind agents [1, 2] to estimate the extent to which participants make use of theory of mind in simple strategic games.

We consider behavioral data from two empirical studies in which participants play a simple repeated game known as matching pennies. In this game, two players simultaneously choose one of two possible actions. The first player wins if both players selected the same action, while the second player wins if the two players chose different actions. Sher et al. [6] let 69 children play a total of 12 rounds each against a confederate who always selected the action that would have won in the last round, while Devaine et al. [3] let 29 adult participants play this game against software agents that followed a theory of mind strategy for a total of 480 rounds per participant.

2 Agents as generative models of behavior

We made use of a technique known as group-level random-effects Bayesian model comparison (RFX-BMS), introduced by Stephan et al. [7]. This technique assumes that different participants may be best described by different models. Given a set of models of participant behavior, RFX-BMS determines the relative proportions of these models that best describes the aggregate behavior of participants.

Our model set consists of nine models of participant behavior. The *biased* model of behavior ignores all past behavior and selects each action with a fixed probability. The *other-regarding* model takes past behavior into account by playing the best response to the opponent's last observed action. The *self-regarding* model, on the other hand, repeats the action it performed in the previous round with a fixed probability. Finally, the *Nash* model of behavior selects an action to play at random.

In addition to the heuristics described above, we also included the theory of mind agents we developed to evaluate the benefits of higher-order theory of mind [1, 2]. A zero-order theory of mind (ToM_0) agent predicts the future behavior of the opponent based on her past behavior. A first-order (ToM_1) agent also considers the game from the perspective of his opponent, and determines what he would do himself in her position. Each additional order of theory of mind provides a theory of mind agent with an additional hypothesis of his opponent's future behavior. 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 behavior.

[1]The full paper has been published in *Proceedings of the Twelfth Conference on Logic and the Foundations of Game and Decision Theory*, 2016.

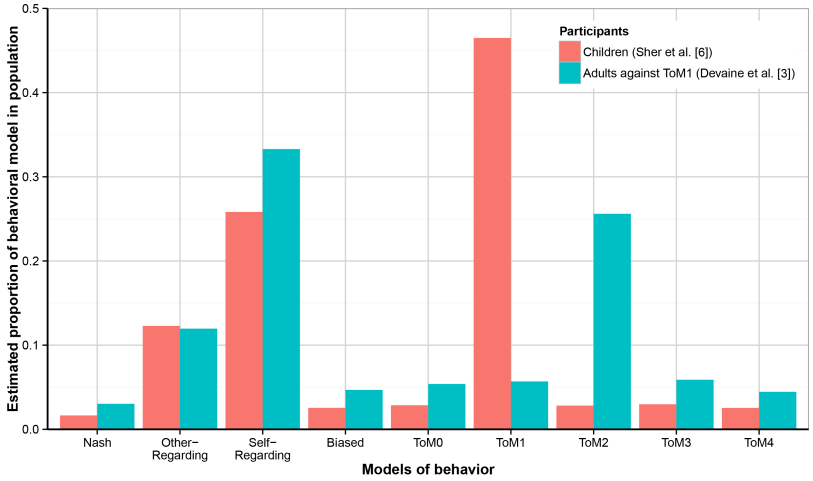

Figure 1: Estimated proportions of behavioral models in the matching pennies game.

3 Results

Figure 1 shows the estimated proportions of the nine models of participant behavior. These results show that Bayesian RFX-BMS estimation classifies over 45% of the children in the study by Sher et al. [6] as using first-order theory of mind. However, the behavior of many children is better described by the either the other-regarding or the self-regarding model of behavior.

Figure 1 also shows that many adult participants playing against a first-order theory of mind agent in the study by Devaine et al. [3] are classified as using second-order theory of mind. However, over 30% of the participants are better described by the self-regarding model of behavior. This suggests that a sizable proportion of participants may rely on simple heuristics when playing matching pennies.

Our results suggest that both children and adults engage in theory of mind in simple games such as matching pennies, but also that many participants are better described by simpler models of behavior. In future research, we aim to determine whether this lack of theory of mind reasoning is due to difficulties participants may experience in classifying the behavior of their opponent. Alternatively, human players may not use their theory of mind abilities primarily to compete with others. Bayesian RFX-BMS estimation could provide more insight in whether participants are more likely to use theory of mind in more cooperative settings, or in settings in which both cooperation and competition play a role.

References

[1] Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence*, 199-200:67–92, 2013.

[2] Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. Higher-order theory of mind in the Tacit Communication Game. *Biologically Inspired Cognitive Architectures*, 11:10–21, 2015.

[3] Marie Devaine, Guillaume Hollard, and Jean Daunizeau. The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology*, 10(12):e1003992, 2014.

[4] Adam S. Goodie, Prashant Doshi, and Diana L. Young. Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, 25(1):95–108, 2012.

[5] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(04):515–526, 1978.

[6] Itai Sher, Melissa Koenig, and Aldo Rustichini. Children's strategic theory of mind. *Proceedings of the National Academy of Sciences*, 111(37):13307–13312, 2014.

[7] Klaas E. Stephan, Will D. Penny, Jean Daunizeau, Rosalyn J. Moran, and Karl J. Friston. Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017, 2009.