Socially smart software agents entice people to use higher-order theory of mind in the Mod game

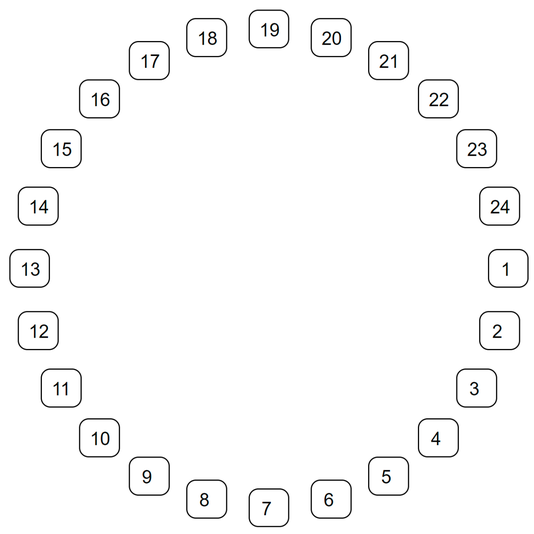
Kim Veltman[1], Harmen de Weerd[1,2], Rineke Verbrugge[1]
[1]Institute of Artificial Intelligence, University of Groningen
[2]Research Group User Centered Design, Hanze University of Applied Sciences

Introduction

- *Theory of mind* [3] allows people to reason about unobservable mental content of others, such as their beliefs, desires, or intentions.
- People are capable of using theory of mind recursively, and use higher-order theory of mind to reason about the theory of mind abilities of others [5].
- In strategic settings, people typically rely on zero-order or first-order theory of mind and are slow to engage in higher-order theory of mind [1].
- The best response to an opponent following kth-order theory of mind is to reason at $(k+1)$st-order theory of mind [6].

Experiment

The Mod game [2] is an extension of rock-paper-scissors. In our experiment, two players each choose a number between 1 and 24.



Players score a point if they chose the number that is exactly one higher than the number chosen by their opponent. In addition, players that choose the number 1 score a point if their opponent has chosen number 24.

Participants knowingly play Mod games against a ToM_1 agent, a ToM_2 agent, a ToM_3 agent, and a randomizing agent that randomly switches between these three options every round.

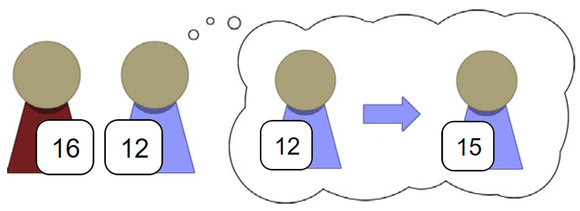
- Blocks of 20 rounds per opponent
- Each opponent appeared in two blocks

Random-effects Bayesian model selection

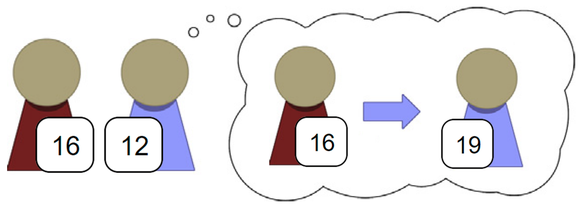
To classify participant behavior, we make use of random-effects Bayesian model selection [4]. In this analysis, we distinguish the following strategies to play the Mod game.

Behavior-based strategies

- The *k-self-regarding* strategy selects the number that is k higher than the number chosen in the last round with some fixed probability.



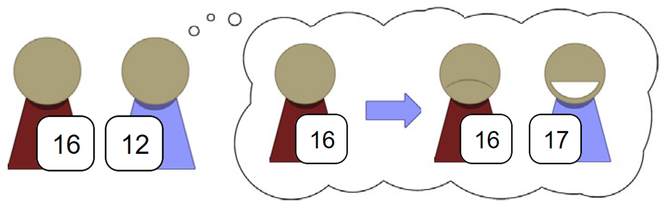
- The *k-other-regarding* strategy selects the number that is k higher than the number the opponent chose in the last round with some fixed probability.



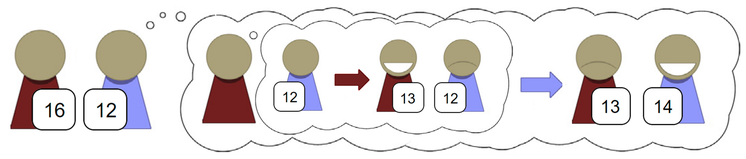
- The *win-stay lose-shift* strategy selects the same number as chosen in the last round if that number led to a victory, and otherwise randomly picks another number.

Theory of mind strategies [6]

- The *zero-order theory of mind* ToM_0 strategy predicts that if the opponent chooses number n, it is likely that the opponent will play number n again in the future.



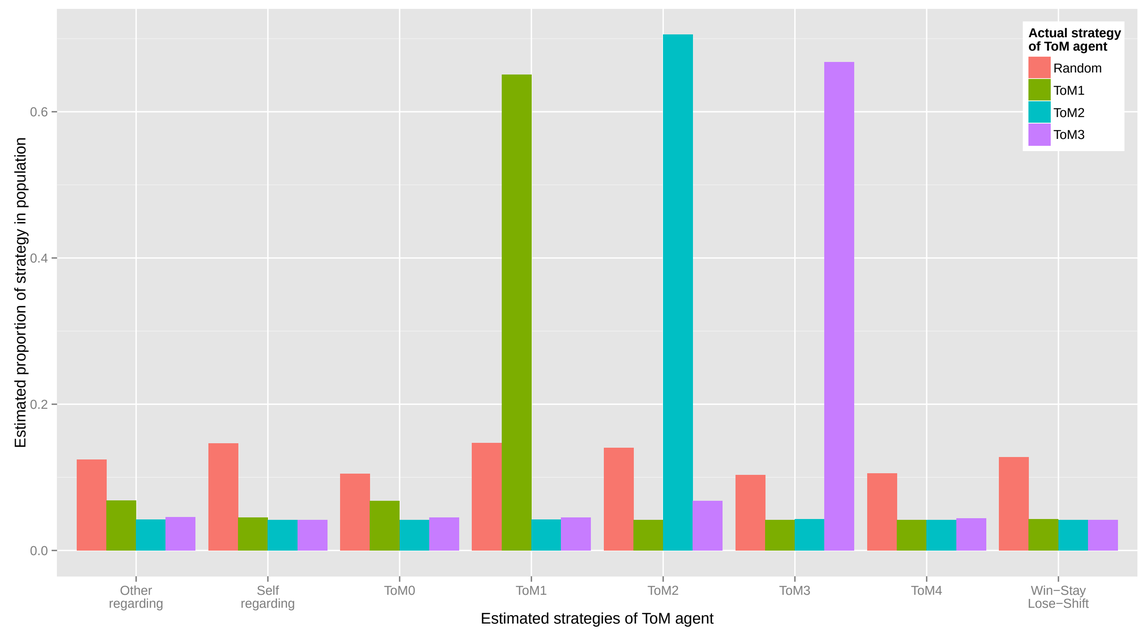
- The *first-order theory of mind* ToM_1 strategy extends the ToM_0 strategy with the possibility that the opponent follows a ToM_0 strategy.

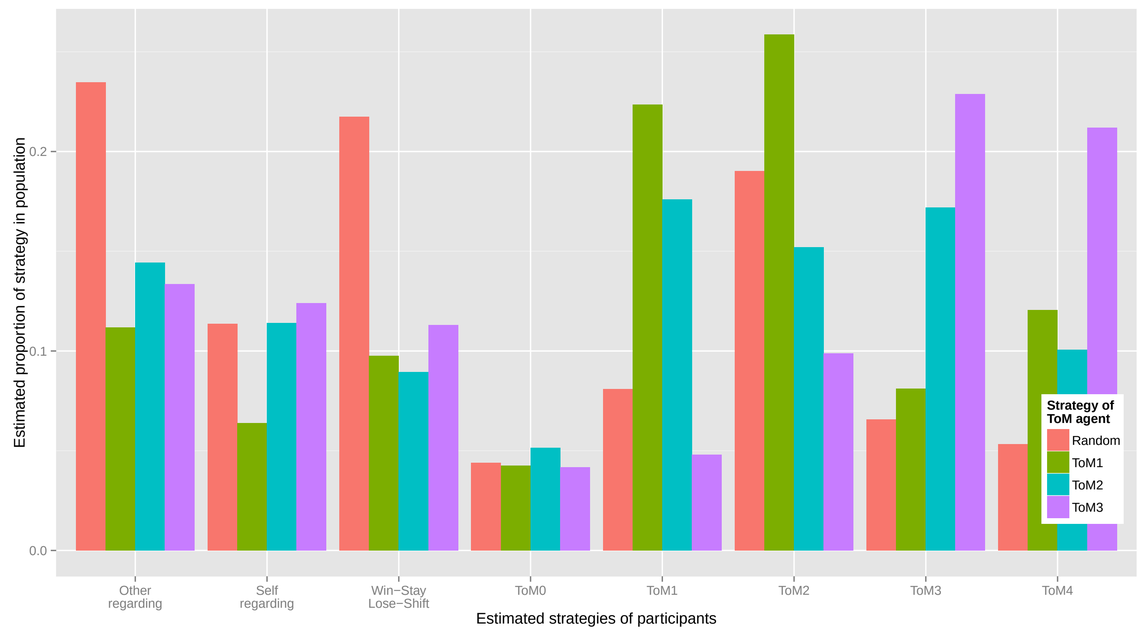


- The *kth-order theory of mind* ToM_k strategy attributes all lower order of theory of mind strategies to his opponent.

Results and discussion

We used random-effects Bayesian model selection (RFX-BMS, [4]) to determine the level of theory of mind reasoning of the participants playing the Mod game. The following figures show the estimated strategies for the artificial agents (left) and the participants (right).





- RFX-BMS accurately recovered the level of theory of mind reasoning of theory of mind agents (green, blue, and purple bars in left figure).
- RFX-BMS is unable to classify the randomizing agent (red bars in left figure).
- Participants adjust their level of theory of mind reasoning to their opponent:
  - When playing against a ToM_1 opponent, participants are best explained as using first-order or second-order theory of mind (green bars, right figure).
  - When playing against a ToM_3 opponent, participants rely on third-order or fourth-order theory of mind (purple bars, right figure).
  - Participants that play against the randomizing agent are better explained as using simple, behavior-based strategies (red bars, right figure).
- Surprisingly, participant behavior shows evidence of fourth-order theory of mind reasoning (purple bars, right figure). This is much higher than would be expected based on the literature.

References

[1] Goodie, A.S., Doshi, P., and Young, D.L.: Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, 25(1):95–108 (2012).
[2] Frey, S. and Goldstone, R.L.: Cyclic game dynamics driven by iterated reasoning. *PloS ONE*, 8(2):e56416 (2013).
[3] Premack, D. and Woodruff, G.: Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526 (1978).
[4] Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J.: Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017 (2009).
[5] Verbrugge, R.: Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic*, 38:649–680 (2009).
[6] de Weerd, H., Verbrugge, R., and Verheij, B.: How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence*, 199-200:67-92 (2013).

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