Adaptive and Satisficing Cognition for Theory of Mind in Interaction

Jan Pöppel and Stefan Kopp

Social Cognitive Systems, CITEC, Bielefeld University, Bielefeld, Germany
{jpoeppel,skopp}@techfak.uni-bielefeld.de

When agents interact with each other, they often depend on some form of "understanding" of each other. For competitive interactions, insight into the intentions of the other agent can give one an edge. For cooperative settings, understanding the intentions and capabilities of team members is important to maintain coordination. In either case an agent needs to make suitable decisions based on beliefs about others' mental states. The concept of inferring others' mental states is a cognitive capacity that has been termed Theory of Mind (ToM). People develop increasingly sophisticated mentalizing capabilities from inferring others' likely goals to their (potentially false) beliefs, preferences and even emotions [9] which can even be considered recursively.

Nevertheless, it appears as if people do not always invest all the computational costs involved in these complex inferences. For example, evidence has suggested that people often employ heuristics, e.g., in the form of egocentric biases, in their decision making when not explicitly prompted to consider others' mental state. In our own work, we have found that people appear to "switch" between different modes of ToM depending on a wide range of factors, including priming a mental state consideration [7] as well as the observed agent's decision problem [8]. This view of mentalizing is in line with a resource-rational [4] perspective on human reasoning. Exact inferences of complex mental states are computationally very demanding [3]. At the same time, the inferred knowledge may not always be relevant for a given situation or the potential costs of making incorrect inferences may be negligible. In that sense people may be performing "satisficing" mentalizing where they try to only invest as much mental effort as required for the current situation.

For artificial agents to be able to interact with others in socially intelligent ways, they also need to be equipped with those capabilities. The Bayesian Theory of Mind (BToM) framework already provides promising results when it comes to explaining the inferential capabilities of people [1]. However, BToM is inherently computationally very demanding and relies on strong assumptions and heuristics to be applicable [2]. These computational costs are more severe in interactions where people expect timely reactions of others. Further, in scenarios with independent agents, long computation times may lead to reasoning about outdated states and thus unsuitable actions. In [5] we have shown how a minimal BToM model coupled with a predictive procession hierarchy can enable low-level coordination with real-time capabilities in a situated interaction task.

For these reasons we present first steps towards a computational approach to adaptive and satisficing ToM. Such models can, on the one hand, better account for how humans mentalize about others ecologically and, on the other hand, enable artificial agents to efficiently draw the other-related inferences needed in competitive or cooperative interaction. In earlier work, we successfully explored the use of a "switching" approach that employs different inference models that make different assumptions (or heuristics) [6]. The agent sticks with simpler models as long as they are capable of explaining its observations, and only switches to more complex ones when needed. This switching model was able to outperform any single specialized model as well as a full BToM model that was not restricted to specific assumptions while being orders of magnitude more efficient than

the full model. However, the model repertoire was limited to a specific scenario and the switching strategy rather simplistic. Work is underway to improve on these strategies in an actual interaction setting, where an agent may also want to adapt its strategies for planning as well as mentalizing based on what the interaction partners do.

References

1. Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1:0064, 2017.
2. Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.
3. Mark Blokpoel, Johan Kwisthout, Theo P van der Weide, Todd Wareham, and Iris van Rooij. A computational-level explanation of the speed of goal inference. *Journal of Mathematical Psychology*, 57(3-4):117–133, 2013.
4. Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, page 1–85, 2019.
5. Jan Pöppel, Sebastian Kahl, and Stefan Kopp. Resonating minds – emergent multi-agent collaboration throughhierarchical active inference. *Cognitive Computation*, submitted.
6. Jan Pöppel and Stefan Kopp. Satisficing models of bayesian theory of mind for explaining behavior of differently uncertain agents: Socially interactive agents track. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 470–478. International Foundation for Autonomous Agents and Multiagent Systems, 2018.
7. Jan Pöppel and Stefan Kopp. Egocentric tendencies in theory of mind reasoning: An empirical and computational analysis. In *Proc. CogSci*, volume 41, pages 2585–2591, 2019.
8. Jan Pöppel, Stacy Marsella, and Stefan Kopp. Less egocentric biases in theory of mind when observing agents in unbalanced decision problems. In *Proc. CogSci*, volume 43, 2021.
9. Henry M Wellman and David Liu. Scaling of theory-of-mind tasks. *Child development*, 75(2):523–541, 2004.