Understanding Second-Order Theory of Mind

Laura M. Hiatt
Naval Research Laboratory
4555 Overlook Ave, SW
Washington, DC 20375
laura.hiatt@nrl.navy.mil

J. Gregory Trafton
Naval Research Laboratory
4555 Overlook Ave, SW
Washington, DC 20375
greg.trafton@nrl.navy.mil

ABSTRACT

Theory of mind is a key factor in the effectiveness of robots and humans working together as a team. Here, we further our understanding of theory of mind by extending a theory of mind model to account for a more complicated, second-order theory of mind task. Ultimately, this will provide robots with a deeper understanding of their human teammates, enabling them to better perform in human-robot teams.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General—*cognitive simulation*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*intelligent agents, coherence and coordination*

General Terms

Theory

Keywords

theory of mind; human-robot teams

1. INTRODUCTION

Theory of mind (ToM) is a critical capability for teams of agents working together, whether robots, humans, or both; without it, people can be extremely limited in their abilities to interact naturally and effectively with others [3, 7]. By better understanding how people achieve this core component of teamwork, we can develop robots that are able to more effectively coordinate with their human partners, both by enabling them to have theory of mind in the same way, and, just as importantly, by allowing them to understand what beliefs, desires and intentions a human partner may be ascribing to them or to someone else [4].

Since adult humans are typically very adept at theory of mind tasks, studies on theory of mind's underlying mechanisms are often performed on developing children. Additionally, studies are typically done on variations of *false-belief* (FB) tasks [7], a common setting in which to explore ToM where participants distinguish between a "true-belief" answer (e.g., where the participant thinks an object actually is) and a "false-belief" answer (e.g., where the participant believes someone else thinks an object is). More complicated variations include *avoidance* tasks, or predicting someone's behavior based on their identified false-belief [6].

In previous work, we developed a cognitively-plausible process model of how children develop the ability to correctly answer false-belief and avoidance queries [5]. The model's theory states that children simultaneously develop the abilities to answer these queries, including using cognitive simulation for the more complicated avoidance query, and learn to take advantage of their newly unlocked abilities. Here, we strengthen the position of our work by using the model's processes to also account for a new study on *second-order* false-belief queries, where participants are asked about what another believes about a third party [1]. We show that our model can account for the new data with only procedural changes, strengthening its theoretical claims.

2. THEORY OF MIND MODEL

Our existing theory of mind process model hypothesizes that, for very simple ToM queries, such as the first-order FB query, the model relies on simple reasoning mechanisms (e.g., inhibiting their own, true, belief so that they can identify another's belief [6]). The model also predicts that being able to select between beliefs in this way is a precursor for simulation, which allows people to use the beliefs and desires of others to predict and understand another's behavior; this is ultimately what provides full-fledged ToM. Simulations have access to the model's procedures and cognitive resources, and include the knowledge and beliefs needed to perform them. For the avoidance query, then, the model first predicts the other person's belief like it does in the first-order FB task; then, it uses the identified belief as the basis for simulating the others' actions to predict their behavior.

The model's development is based on the idea that, as children grow, they learn and mature simultaneously; e.g., as they develop, they learn to take advantage of their maturing ability to select between competing beliefs. This occurs during two developmental shifts: the first where the model develops the ability to inhibit beliefs, and the second where the model develops the ability to perform simulation. Maturation is modeled via a developmental parameter; learning occurs via procedural utility learning.

The model fit the data very well for both a first-order FB task meta-analysis, and for an avoidance query experiment.

HRI'15 Extended Abstracts, March 2–5, 2015, Portland, OR, USA.
ACM 978-1-4503-3318-4/15/03.
http://dx.doi.org/10.1145/2701973.2702030.

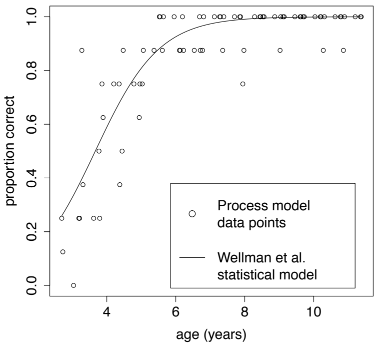

Figure 1: Scatterplot of the model's first-order false-belief results and the best-fit curve from [7].

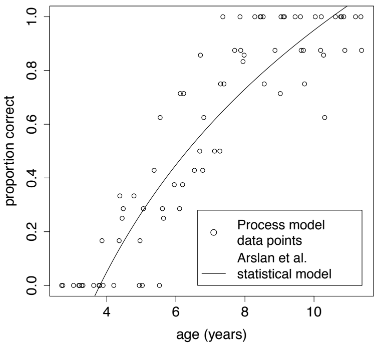

Figure 2: Scatterplot of the model's second-order false-belief results and the best-fit curve from [2].

See [5] for more details on the model mechanisms, development, and results.

2.1 Model of Second-Order False Belief

Here, we extend our prior model to account for second-order false-belief queries. The process is identical to how the model answers avoidance queries: the model first uses inhibition to select a second person's beliefs, and then uses that as the basis for simulation. The key difference is that the goal of simulation, instead of predicting another's actions, is to identify a third person's beliefs based on what the second person believes. For the second-order FB query, then, where the model is asked to identify someone else's belief about a third person's belief, the model identifies the second person's belief, then spawns a simulation of the second person that, in turn, inhibits the second person's belief to identify the third person's belief.

The model develops this ability as it develops the ability to answer the avoidance query: the ability to perform simulation matures as the model simultaneously learns to use it to answer these types of queries.

3. RESULTS AND DISCUSSION

We test our model against two datasets focusing on how children acquire the ability to perform first- and second-order ToM: the meta-analysis of first-order false-belief tasks [7] that was previously matched against the original model, and a study showing how children develop the ability to perform second-order false-belief tasks [1]. We simulated five developing models, querying each at fifteen different model "ages", to collect results on how the models develop theory of mind over time.

For the first-order FB task, the model matches the data well: $r^2 = 0.53$, which is significantly better than the meta-analysis' statistical model ($r^2 = 0.39$), and approaches their multi-variate model ($R^2 = 0.55$). For the second-order FB task, $r^2 = 0.83$, which is also an excellent fit.

A different approach explains second-order FB tasks using the activation of different strategies in memory [2]. Our approach, however, provides a stronger fit to the data than this work, which matches only limited data points. Our work also provides a more complete picture, since it explains why, even with exposure, children do not reliably answer these queries until they (and, thus, the involved cognitive mechanisms) reach an advanced enough developmental stage.

To summarize, we have shown that our previous model, which accounted for children developing the ability to correctly answer the first-order false-belief and avoidance queries, can naturally extend to account for children answering second-order false-belief queries. This suggests that both the avoidance query and the second-order false-belief query utilize the same mechanisms in the brain; here, we posit that that mechanism is simulation.

Overall, robots with theory of mind are viewed as more natural and intelligent teammates to their human partners [4]. In this work, we improve our understanding of how this phenomenon works, as well as expand our capabilities to also account for second-order ToM. Ultimately, this enables us to build robots that are able to be better partners to their human teammates because of their deeper knowledge of how humans use theory of mind in team situations.

4. 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 *Proceedings of Cognitive Science*, pages 1290–1295, 2012.

[2] B. Arslan, N. Taatgen, and R. Verbrugge. Modeling developmental transitions in reasoning about false beliefs of others. In *Proceedings of ICCM*, 2013.

[3] S. Baron-Cohen, A. M. Leslie, and U. Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21, 1985.

[4] L. M. Hiatt, A. M. Harrison, and J. G. Trafton. Accommodating human variability in human-robot teams through theory of mind. In *Proceedings of IJCAI*, 2011.

[5] L. M. Hiatt and J. G. Trafton. A cognitive model of theory of mind. In *Proceedings of ICCM*, 2010.

[6] A. M. Leslie, T. P. German, and P. Polizzi. Belief-desire reasoning as a process of selection. *Cognitive Psychology*, 50:45–85, 2005.

[7] H. W. Wellman, D. Cross, and J. Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3):655–684, 2001.