Toward Artificial Theory of Mind

Theory of mind (ToM) is the ability to attribute mental states to other agents and use those states to explain and predict behavior (Carruthers et al., 1996). To reason about the mind, it would seem that an agent must experience the mind. But developments in AI and robotics are making it increasingly clear that an agent does not need a mind to play the game of mind reading.

In human-human interactions, having a strong theory of mind generates social intelligence. From relatively few social cues and without explicit communication, we can infer an agent's beliefs, intentions, and desires, and act on them accordingly. Peering into the far reaches of others' minds leads to better interpersonal understanding and cooperation. Recently, humans have been introduced to another agent that we interact with on a daily basis: artificial intelligence. But there is a major communication barrier limiting human-AI interaction. Namely, AI does not understand humans, and humans do not understand AI (Celikok et al., 2019). Anyone who has tried to have a conversation with Siri or Alexa knows just how problematic this can be. Fostering cooperation between both parties requires that AI have a theory of mind of humans and humans have a theory of mind of AI. The former is achieved by developing an AI that can form internal models of human agents. The latter is achieved by making the AI's algorithms transparent enough that the AI can explain its decision-making process to human agents. The described artificial ToM promises to transform human-AI, human-robot, and multi-agent AI interactions (Winfield, 2018).

How, then, can AI have a theory of mind? How can a series of logic gates and electrical signals learn to reason about and predict human behavior?

One clarification is important to mention here: artificial ToM is not looking to attain any real computational *understanding* of the human mind. It is accepted in the field that a computer will likely not comprehend the mind until artificial consciousness is realized. Instead, what is proposed is an AI that can infer human ideas and anticipate human actions. But this requires no awareness of the mind -- only an ability to replicate how humans move back and forth between action and conception (Sebastian, 2016).

Even so, the above questions remain relatively open ended. The complication that arises when considering how to implement an artificial ToM is that there is little empirical evidence supporting any one theory about ToM in humans. Some leading cognitive approaches like Simulation Theory, Theory Theory, Bayesian Theory of Mind, and the Intentional Stance have been applied in isolation or in combination to artificial ToM programs with limited success. Some are more computationally conducive than others, some succeed in depth but lack the ability to generalize, and some succeed in breadth but fail to produce precise results. Given this ambiguity, one may wonder whether pursuing artificial ToM is worth it until the field realizes an accurate cognitive theory. But what is particularly encouraging about artificial ToM is that it provides a testing ground for our cognitive theories. Implementing ToM allows us to falsify some approaches and computationally model others with clear and decisive data. So, in pursuing artificial theory of mind, we pursue genuine theory of mind in tandem (Winfield, 2018).

In the remainder of the paper, I will first run through a couple of the most promising artificial ToM experiments to date. Then, I will offer my two cents on potential improvements to existing programs and ideas for new experiments moving forwards.

In 2018, Google's DeepMind rolled out ToMnet -- a theory of mind neural network that adopts Simulation Theory as its conceptual basis. Simulation Theory proposes that humans understand others' minds by simulating what we would think or feel if we were in their shoes (Barlassina et al., 2017). But ToMnet does not interact with human agents; instead, it builds internal models of other AI agents from observations alone and makes rich predictions about their states. ToMnet comprises three neural nets: the first learns the tendencies of other AIs based on their past actions, the second forms a general concept of their current state of mind (beliefs and intentions at a particular moment), and the third uses the outputs of these networks to predict the AI agent's actions. In one experiment, ToMnet observed three AI agents maneuvering a room to collect colored boxes. One agent was programmed to be near-sighted, so when the layout of the room changed, they falsely believed that they were still navigating the old environment and stuck to their original paths. ToMnet identified this disability and accurately adjusted predictions for the agent's movement by simulating itself in a near-sighted state. In this way, ToMnet recognized that other agents can hold false beliefs about the world, passing the revered "false belief task" which is often used to demonstrate ToM in cognitive studies. ToMnet's primary drawback, however, is that its understanding is deeply entwined with its training context. ToMnet performs poorly when predicting behavior in radically new environments and would struggle to model a human agent (Rabinowitz et al., 2018).

MIT's Saxelab has recently introduced a novel approach to cognitive ToM: Bayesian Theory of Mind (BToM). BToM is a computationally realizable quantitative model, making it a promising candidate for use in artificial theory of mind. BToM is grounded in Daniel Dennett's ToM theory the Intentional Stance, which hypothesizes that humans treat others as having a mind as a way of making sense of their behavior. Dennett posits that we can reverse engineer human mental state inferences by treating minds as rational actors whose behavior comprises intentional actions, which is just what the BToM looks to replicate. As rational actors, agents are expected to choose the actions that achieve their desires most effectively, or in other words, maximize their expected utility (Dennett, 2016). By observing an agent's behavior within an environment and the utility function that they pursue, their beliefs and desires are inferred using Bayesian inference. Actions, utilities, beliefs, and desires are then filed into an agent's prior probability and used to predict future behavior. In an experiment where BToM was tasked with predicting an agent's food preferences and spatial beliefs based on their movement between food vendors, BToM made nearly identical inferences to human participants. Despite this success and BToM's versatility among problem spaces, BToM fails to generate as precise results as DeepMind's ToMnet (Baker et al., 2017).

Artificial ToM is only in its infancy and is evidently a ways away from perfection. The described approaches are pioneering the field but their empirical success may be jeopardized by omission of certain cognitive processes.

DeepMind's ToMnet may benefit from considering the distinction between factive and non-factive ToMs. Factive ToM is the general capacity to represent another agent's understanding of the world, while non-factive ToM is the ability to represent the way another

agent believes the world to be (Phillips et al., 2018). The experiment in which ToMnet simulated an agent's false belief was a demonstration of its capacity for non-factive ToM: ToMnet dissociated its understanding of the world from the agent's belief of the world and identified that inconsistency. I would be curious, then, to see how ToMnet performs in a test of factive ToM. Specifically, I question ToMnet's ability to simulate altercentric and egocentric ignorance. How might an AI simulate itself not knowing something it already knows? How might an AI simulate knowing something that it doesn't and cannot know?

If Saxelab's BToM is to be applied to artificial ToM, then it may benefit from revising the confidence that it places in Dennett's Intentional Stance. Recall that the Intentional Stance supposes that humans make mental state inferences by treating minds as rational actors that seek to maximize expected utility. Challenging this expected utility theory, Amos Tversky and Daniel Kahneman propose the prospect theory, which adjusts for cognitive constraints on decision-making. When it comes to risk taking, Tversky and Kahneman show that humans are largely irrational. We impose a series of heuristics and biases that, under certain risk-inducing circumstances, do not lead us to maximize expected utility (Tversky et al., 1992). In human-human interactions, this irrationality is neutralized because prospect theory is ubiquitous; when reasoning about another agent's mind, one will assume their own cognitive biases and therefore make an accurate inference of the other agent's cognitively biased mental state. But in human-BToM-based-AI interactions, we are faced with another communication barrier: a human agent infers cognitive biases in the AI that don't actually exist, and the AI assumes the human is rational in situations where cognitive biases prevail. To overcome this hurdle, I propose that a

BToM-based AI adopt prospect theory instead of expected utility theory so that cognitive biases are realized.

Lastly, I would be interested to see if artificial ToM approaches involving Simulation Theory and BToM can be reconciled. Since Simulation Theory is empirically strong in depth but not in breadth while BToM is empirically strong in breadth but not in depth, the two theories may find the balance of generalization and precision that current programs lack.

The road ahead for artificial ToM is long and winding, but progress is steady. As theories about cognitive ToM inform experiments in artificial ToM and vice versa, we near what will one day be a compatible and transparent relationship between humans and technology at large.

Works Cited

Baker, C., Jara-Ettinger, J., Saxe, R. *et al.* Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat Hum Behav* 1, 0064 (2017). https://doi.org/10.1038/s41562-017-0064

Barlassina, Luca, and Robert M. Gordon. (2017) "Folk Psychology as Mental Simulation." Stanford Encyclopedia of Philosophy, Stanford University, plato.stanford.edu/entries/folkpsych-simulation/.

Carruthers, P., and Smith, P. (1996). *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Celikok, Mustafa Mert, et al. (2019) "Interactive AI with a Theory of Mind." Helsinki Institute for Information Technology, arxiv.org/pdf/1912.05284.pdf.

Dennett, Daniel Clement. The Intentional Stance. The MIT Press, 2006.

Phillips, J., & Norby, A. (2018). Factive theory of mind. *Mind & Language*.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M., & Botvinick, M. (2018). Machine theory of mind. *arXiv preprint arXiv:1802.07740*.

Sebastian, M. A. (2016). Consciousness and theory of mind: a common theory? *Theoria Revista de Teoria, Historia y Fundamentos de la Ciencia* 31, 73–89.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297-323.

Winfield AFT (2018) Experiments in Artificial Theory of Mind: From Safety to Story-Telling. *Front. Robot. AI* 5:75. doi: 10.3389/frobt.2018.00075