Cognitive Models as a Computational Correlate of Theory of Mind for Human-Machine Teaming

Leslie M. Blaha (leslie.blaha@us.af.mil)
Air Force Research Laboratory, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

I delve into an initial discussion on the nature of the theories of mind needed to support effective human-machine teaming. Effective human-machine teaming will require humans to have a theory of mind about machine intelligence and for machine intelligence to have a theory of mind about human teammates. The latter will require a machine to be able to make inferences about the cognitive states related to observable behaviors by the human and to predict future states and actions consistent with the human's beliefs, goals, and desires. This paper proposes that cognitive models can provide the computational correlates to enable a machine theory of mind to reason about its human counterparts.

Keywords: Human-machine teaming; Computational cognition; Cognitive models; Human information processing; Theory of mind

Introduction

The purpose of this paper is to spark an exploration around the nature of the theory of mind required to support human-machine intelligence teaming. I begin with the claim that a theory of mind is necessary for humans and machine intelligence to work together in collaborative teaming situations. These are situations in which a machine has autonomous capability, meaning it can act alone without human supervision or direct intervention, can take direction or feedback from a human, can give direction or feedback to a human teammate, and leverages some form of artificial intelligence to process information, learn and adapt to complete tasks and achieve the team's goals.[1]

Human-machine teaming of this type is predicated on the assumption that humans and machine intelligence understand each other. We can see this in claims that increasing transparency of automation will allow humans to properly calibrate their trust and reliance on the technology (Lee & See, 2004). Or it is similarly implied in the claims that artificial intelligence endowed with the ability to explain its decisions (so called explainable AI or XAI) will aid human users to reason about the correctness and sources of error in the machine's output (Hoffman, Klein, & Mueller, 2018). The push for real-time state assessment in humans is partially driven by the goal of representing the human in ways that can be interpreted and adapted to by machine systems (e.g., Borghetti & Rusnock, 2016). Across these research topics and engineering endeavors, there is a common theme of measuring, identifying, and representing the unobservable states of agents to make them understandable to the other team members, particularly between heterospecific team members.

We have been implicitly demanding a theory of mind to support effective human-machine teaming.

Theory of Mind Defined for Human-Machine Teams

Theory of mind (ToM) is the term ascribed to the processes an agent uses to impute the internal "mental" states of itself and other agents (c.f. Fodor, 1992; Mahy, Moses, & Pfeifer, 2014; Premack & Woodruff, 1978). Note that herein, I am using the term mental state both for humans and machines to refer to the internal information processing mechanisms and representations that are only indirectly observable by the other agent. In various social and developmental lines of ToM research, this inference process is usually considered conceptually from the perspective of an exemplar human or primate, the "subject" of the study. The social interactions, and therefore relevant ToM, is about the subject's ability to reason about itself and one or a small number of other agents, usually other humans.

One level of reasoning within ToM emphasizes the subject's ability to interpret observed actions of the other as goal-directed behaviors. That is, the ToM must support the interpretation of a sequence of actions as representing a trajectory through a state space toward a goal state. Any time the agent is seeking the same goal state, it is likely to exhibit similar sequences of behaviors. A subject could reason over these trajectories to abstract a degree of meaning about the goals driving the observed behaviors. However, ToM is usually invoked at a deeper level: the inferences by the subject should be representing the intentions, emotions, prior experiences, mental state, awareness, and goals of the other agent. That is, we hypothesize that a subject capable of full ToM is attempting to represent to him or herself the latent factors within another agent that contextualize the goal-oriented behaviors.

The dominant theories about ToM generally argue that either people rely on their own mental mechanisms to simulate the experiences of other agents (e.g., Scholl & Leslie, 1999), or they rely on their ability to reason over internal conceptual representations of cognitive mechanisms (e.g., Gopnik & Wellman, 1994). A key commonality across theories is the reliance on an internal representation of the mechanisms of mind. This brings us to the crux of the challenges in defining a ToM for human-machine teaming, which can be summarized in three questions:

1. What are the mechanisms of mind for machine intelli-

[1] At this junction, I am agnostic to whether that intelligence is embodied in a robotic form and to the specific nature of the interactions and communications between the human and machine intelligence. These details not change the present argument, though are critical for engineering actual systems.

gence?

2. How do we represent machine mechanisms of mind in humans to be reasoned over?

3. How do we represent human mechanisms of mind in machines to be computed about?

The nature of human-machine teaming and the fundamental differences between human cognition and computational processes require that we expand the concept of ToM to include multiple types of ToM models and mechanisms. In spite of our often-useful analogy of cognition as computation, the nature of the ToM for machines reasoning about machines, machines reasoning about humans, and humans reasoning about machines must be different than human ToM about other humans. Elucidating the nature of these new theories of mind is a hard problem. Indeed, I note that developing an artificial theory of mind to support human-robot interaction was listed as one of the top grand challenges in humanoid robots today (Yang et al., 2018).

The human ToM within a human-machine team will likely operate as a classical ToM: introspection about self and introspection about other people (particularly for multi-human, multi-agent team configurations) will continue to engage processes of simulating and theorizing about mental states based on our own experiences with self and interacting with other people. But now human ToM must also provide introspection about machine intelligence. Properly supporting such heterospecific introspection will require the development of appropriate mental models for machine intelligence capabilities. Deeper discussion about human mental models of machine intelligence is beyond the current scope.

Let us make the working assumption that a machine ToM parallels human ToM. It must enable a machine intelligence to "introspect" about itself.[2] It must enable a machine intelligence to introspect about other machine agents. In some cases, the other agents may employ similar artificial intelligence algorithms, but machine learning, which is sensitive to input data and conditions, may have produced deviating internal representations of the world. In other cases, other machine agents may have completely different algorithms, chip architectures, and system structure. It could potentially take a complex set of representations and savvy abstractions to enable machines to reason about other agents. Recently, Rabinowitz and colleagues (2018) have made headway in developing machine ToM that abstracts all agent behaviors into state-action trajectories and engages pattern recognition for inferences between agents (see also Winfield, 2018, for a candidate abstraction in robots).

Finally, a machine ToM for human-machine teaming must enable the machine to reason about human teammates. I argue that it will not be enough to abstract a human into a simple, observable state-action sequence for pattern recognition. Analogous to human ToM, the machine intelligence will need to make inferences about the mechanisms of mind, the emotions, intentions, beliefs, and goals of the humans. There may also be cases where the machine must make inferences about physical states and capabilities, too.

The reason we must go beyond simple state-action pattern recognition is that our intentions for human-machine teaming capabilities entail intelligent machines that anticipate and adapt to their human teammates in addition to adapting to dynamic task environments and data. This will require that machines can predict *future* human states and likely actions (and sometimes likely consequences).[3] For machine ToM about humans to achieve prediction or anticipation, it must incorporate a representation of the internal states, intention, beliefs, and goals of the human. It is not enough for the machine intelligence to be reactive to the behavior or action of the human, which may facilitate pattern recognition but not prediction of future actions contextualized by the mental state of the human teammate. It is here that cognitive models of the mental mental mechanisms and processes supporting the human states have a critical role to play.

Cognitive Models in the Machine ToM

We now come to a primary question for consideration by the cognitive modeling community: can cognitive models provide the algorithmic framework(s)—computational correlates, if you will—to enable machine intelligence to have a ToM about human teammates? A limitation of the few current artificial theories of mind is that they do not offer a human-specific representation that differentiates human teammates from other environment variables or computational agents, though the need for such representations to support effective interactions is recognized within social robotics at least (Yang et al., 2018). Winfield (2018) states that the artificial ToM for robots based on a consequence engine is most effective for conspecific agents; that is, reasoning about another agent is most effective when the agent is the same type as the robot. Scassellati (2002) had demonstrable success integrating models of fundamental perceptual skills into humanoid robots to encourage behaviors consistent with the emergence of higher level ToM-related behaviors (e.g., gaze tracking). While behavior consistent with a machine ToM about human teammates is promising, we can go further by not only leveraging models of elements of perception and cognition but leveraging models instantiating full decision-action processes and information processing systems or even full architectures of cognition *and* conceptualizing them as the machine's ToM about the human teammate. In this way, the cognitive models provide a computationally tractable representation of human mental mechanisms, states, beliefs, intentions, and goals—

[2] I use the term introspection here loosely and without proper definition at the present time. This definition will need to delve into the nature of computational inference and state assessment of computational algorithms, which is beyond the present scope.

[3] I note for completion that there is an analogous need for humans to predict the future states, likely actions, and likely consequences of machine activity in the human-machine team. This is related to the need to examine the nature of human mental models about machine intelligence and is left to future exploration.

all those elements critical for deeper introspection within a ToM. And because computational model implementations are in computational languages, they can be integrated into system architectures and intelligent processes.

We must ask then, if cognitive models are to be thought of as a correlate for machine ToM about humans, do they provide the same support to machines that neural correlates of human ToM provide to humans? Within their review of neural correlates of ToM from the social and developmental psychology perspectives, Mahy et al. (2014) offer some initial criteria we can use to evaluate conceptual consistency.

A correlate for ToM should support mental simulation. One key hypothesis for ToM is that people simulate themselves in novel situations and then project inferences about what will happen onto other people (Fodor, 1992; Scholl & Leslie, 1999). Such simulation relies on people having direct access to their own mental states and past experiences. Cognitive models, whether computational cognition formalized in cognitive architectures or mathematical models instantiated in computational algorithms, can simulate human behavior. While the "mental states" of a specific model depend on the mechanisms instantiated in it, generative cognitive models are theoretically grounded in known cognitive mechanisms. In this way, cognitive models might provide machine intelligence teammate direct access to the internal model/mechanism states. Traces of the model history or direct representations of memory, such as declarative memory in ACT-R, provide access to past experiences. The simulated representation of a human (or multiple simulations), can then be compared to observed human behavior to further inform the machine ToM.

A correlate for ToM should be modular in nature. Multiple theories of mind postulate the existence of dedicated, even innate, neural correlates and cognitive mechanisms supporting reasoning about self and others. Modularity of mechanisms is important for the reasoner to keep the inferences about self separate from inferences about others. In our case, then we want to construct human-machine teaming systems where the cognitive models constitute their own module that keeps the representation of human teammates unique from the representations of the task, environment, data or machine's own capabilities. It is not inconsistent to consider the cognitive models within the machine intelligence in a modular way. Designing machine intelligence-based systems in a modular way would enable the system to access its ToM about human teammates when operating with those teammates and to operate autonomously when the human teammates are not present. The representation of the human remains consistent even as the structure or mission of the team changes.

A correlate for ToM supports reasoning over multiple perspectives. A mature ToM is able to hold multiple perspectives in working memory and reason over them independently. This helps someone to differentiate inferences about themselves from inferences about each other individual. Cognitive models have been used as independent agent representations within larger systems. One example is the use of model to support human-robot interaction using ACT-R to simulate human predictions to inform robot planning (Lebiere, Jentsch, & Ososky, 2013). This system enables reasoning about potential human states together with computation about the robot itself. Another example is the development of cognitive-model based synthetic teammates for training (Ball et al., 2010) where the system tracks the synthetic agent and models human learning behavior simultaneously. As long as they are incorporated into systems with adequate processing resources, cognitive models are capable of being used in a modular way in parallel with all other relevant machine intelligence algorithms and artificial ToM about other machine agents.

A correlate for ToM should support theoretically grounded conceptual learning. Human ToM evolves over time, as people learn about themselves and others. They move from simpler to more complex conceptual representations. They evolve to account for observations about other that are inconsistent with currently held conceptions. It is argued that relevant conceptual knowledge must reside in theory-like structures that support the human ToM (Gopnik & Wellman, 1994).

Cognitive models are theoretically grounded in the mechanisms of cognition. As such, they can provide the theoretical structures needed for evolution of conceptual understanding about the human within the machine ToM. Cognitive models can further be equipped with human-like learning mechanisms that enable the model representations to evolve in human-like ways. Consistent with the assumptions of ToM development, this concept learning can be captured through experiential changes in the model and age-related changes in a model operating at a longitudinal scale. This is critical for the machine to have conceptual, or theoretically grounded, representations of how the human's mental state is or could be changing, even if the machine is not learning or reasoning in a human-like way.

Open Questions

Conceptually, cognitive models are capable of supporting machine ToM about human teammates. As we are early in the process of exploring ToM for human-machine teaming, there are a number of open questions that must be debated, including but not limited to:

- Do we need full computational cognitive architectures or unified theories of cognition instantiated in machine intelligence to make useful inferences?
- How detailed must a human's mental model of the machine be for useful inferences?
- What are the critical tests that a cognitive-model based machine ToM is, in fact, a full theory of mind?

- When would it disadvantage a human-machine team to require a full ToM in the system?

As we evolve our vision for the capabilities of human-machine intelligence teams, and as the evolution of such teams changes the way we even conceive of what be capable, the need for the human and machine to understand each other will remain a constant system requirement. It behooves us to consider now what it means for humans and machines to understand each other and how establishing human-machine teaming theories of mind will inform that understanding. Cognitive models have an important role to play in meeting the grand challenge of developing an artificial theory of mind and a critical role to play when those artificial minds interact with our own.

Acknowledgments

The author wishes to thank Christian Lebiere and John Anderson for stimulating discussions on this topic. This work was sponsored by a seedling grant from the 711th Human Performance Wing Chief Scientist. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

Ball, J., Myers, C., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., et al. (2010). The synthetic teammate project. *Computational and Mathematical Organization Theory*, *16*(3), 271–299.

Borghetti, B. J., & Rusnock, C. F. (2016). Introduction to real-time state assessment. In *International conference on augmented cognition* (pp. 311–321). Springer.

Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition*, *44*(3), 283–296.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. *Mapping the mind: Domain specificity in cognition and culture*, 257.

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining explanation for Explainable AI. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, pp. 197–201).

Lebiere, C., Jentsch, F., & Ososky, S. (2013). Cognitive models of decision making processes for human-robot interaction. In *International conference on virtual, augmented and mixed reality* (pp. 285–294).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Mahy, C. E., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience*, *9*, 68–81.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

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

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, *12*(1), 13–24.

Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and theory of mind. *Mind & Language*, *14*(1), 131–153.

Winfield, A. F. T. (2018). Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI*, *5*, 75.

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., et al. (2018). The grand challenges of science robotics. *Science Robotics*, *3*(14), eaar7650.