Understanding Intention for Machine Theory of Mind: A Position Paper

Casey Kennington
Boise State University

Casey Kennington

Abstract—Theory of Mind is often characterized as the ability to recognize desires, beliefs, and intentions of others. In this position paper, I look at the literature on modeling Theory of Mind in machines and find that, to date, intention is not usually a focus. I define what I mean by intention—choice with commitment—following prior work. Intention has a long history of research in some communities, and I offer one theoretical framework for modeling intention as a starting point. I take inspiration from how children learn intention through joint attention with others and how that leads to Theory of Mind. I argue that though models of machine Theory of Mind need not follow the same learning progression as children, intention is an aspect of Theory of Mind that should be more explicit.

I. INTRODUCTION

Defined broadly, human Theory of Mind (ToM) refers to the capability that people can recognize, represent, and make inferences about the desires, beliefs, and intentions of other people [1]. ToM has been studied in child development, cognitive, and psychological literature (see an overview in [2]), and has recently been explored as an important aspect of interaction between people and machines. From the side of the humans, it is well known that humans mentally attribute anthropomorphic characteristics to machines—robots in particular—based on physical morphology and behavior in many different ways including sympathy and intelligence [3], emotional state [4], age [5], gender stereotypes [6], [7], and social group membership [8]. I infer from this that humans do, to some degree, apply ToM to robots when interacting with robots. The reported results in [5], for example, showed that participants interacting with a robot increased their frustrations with the robots when they had high expectations of the robots based on the robot’s morphology and spoken language production; both evidence that humans sometimes apply a ToM to robots.

However, regardless of whether or not humans apply ToM when interacting with robots, the question as to if, and how, robots could model ToM when interacting with humans is still an open question that has important implications for how humans and robots interact, and human-level artificial intelligence as a broad scientific goal. In this paper, I argue for intention as the driving force behind machine ToM (I adopt the terminology of [9] to distinguish human ToM from machine ToM; machine ToM is currently only an approximation of human ToM).

Existing work attempting to model machine ToM use the same definition as above where desires, beliefs, and intentions are integral parts of the ToM concept. However, recent work in machine ToM generally focuses on desires and beliefs. For example [10] showed how a simple ToM module improves multi-agent interaction. Other recent work for machine ToM [9] models ToM the degree that the model actually passes a version of the Sally-Anne test [11]. The Sally-Anne test has been used to determine the rough age when children develop a ToM, though it is unclear if passing that test for a model or robot is enough to prove existence of ToM. Furthermore, [12] also took inspiration from ToM to arrive at a model that improved a language interaction task. Moreover, in service of exploring machine ToM, [13] released a dataset that includes annotations for human (i.e., others’) beliefs. Taken together, prior work has shown promise in datasets and models for ToM, but even though the authors defined ToM as the desires, beliefs, and intentions of others, they model only considered beliefs over a very small set of possible intentions (though [10] does loosely define intention as being one of a complete game space, which is more about belief than intention to action).

If researchers are to arrive at a useful model of machine ToM, the field should first, I argue, arrive at definitions of what is meant by desires, beliefs, and intentions, and model them accordingly. The three are related, but it is important to make distinctions if they are to be computationally represented and computable. To clarify these terms, [14] argues that intention is related in many ways to desires and beliefs, but is not reducible to them—all three are required, though action is the crucial difference because intention does not just mean a mental state; rather, intention means a commitment to action [15], though as [16] pointed out, not all actions can be easily interpreted for intent so action alone is not enough—the actions must be interpretable as having intent. This paper focuses on intentions because there is ample literature for investigating what intention is, how to model intention computationally, and how it relates to machine ToM. Recent work explored in [17] is more in line with what we are advocating here.

In the following section, I explain what intentions are and for the rest of the paper sketch how existing formal models of intention could be applied to modeling machine ToM, including some of the requirements (e.g, joint attention) for a model that learns intention and ToM.

II. WHAT IS INTENTION?

Whereas belief and knowledge of an agent is often formalized as propositions or finite state spaces, intention in some fields is often encoded as actions. Following Bratman’s
philosophical basis for intention [14], [15] explained that intentions follow four functional roles:

1) intentions normally pose problems for the agent; the agent needs to determine a way to achieve them
2) intentions provide a “screen of a admissibility” for adopting other intentions
3) agents “track” the success of their attempts to achieve their intentions
4) agents must distinguish between possible and actual events

Furthermore, given the above functional roles, if an agent intends to achieve a possible outcome \( p \), then:
- the agent believes \( p \) is possible
- the agent does not believe they will bring about \( p \)
- under certain conditions, the agent believes they will bring about \( p \)
- agents need not intend all the expected side-effects of their intentions

To illustrate these functional roles, a common example is a task where a person asks an assistant robot to fetch an object in another room. In order for the robot to believe \( p \) (i.e., fetching the object) is possible, the robot must be able to identify the object, determine where the object is, plan how to navigate to the object, somehow retrieve the object with some kind of hand, then plan how to navigate to the person while carrying the object. The robot must “believe” that \( p \) is possible by planning the set of actions necessary to complete \( p \) and, having found at least one set of actions that would work, actually follow those actions to bring about \( p \). Along the way of following the actions, the robot may determine that, for example, the object is too big or oddly shaped for the robot to grasp it rendering \( p \) impossible. Moreover, as the robot follows through on its plan to bring about \( p \), it must track its progress, and as it tracks its progress it needs to determine which actions are already complete and which actions are yet to be completed; the set of actions may change as the robot changes its location or realizes that the object is in a different location than it remembered or if it finds other obstacles in its path.

Given the above functional roles (and as illustrated by the above example), [15] argues that intention is choice with commitment; that is, a decision and a coherent plan of action for following through on that decision. In the above example, the robot made a plan for the set of actions to bring about \( p \), and commitment meant following through on the plan to actually bring about \( p \). The authors of [15] offer a model-theoretic, possible worlds formalism for representing intentions. One of the main features of the formalism is that it is action-oriented; in that way it not only acts as a representation for a set of actions to be performed, it also functions as a references for other system components to bring about the actions. In subsequent work, Phil Cohen and colleagues have built on the proposed formal model to an impressive degree of success [18], [19], showing that the model is not just theoretical, but is useful practically, which is evidence to their claim that intention is choice with commitment, and to my claim that the field can pick up on prior work on modeling intention towards a machine ToM.

III. INTENTION AND THEORY OF MIND

How does that stance that intention is choice with commitment affect a machine ToM? The above explanation of intention and subsequent example illustrate only what intention means to an individual, but ToM is concerned with recognizing (beliefs, desires, and) intention of others. Suppose two agents, \( A \) and \( B \) both have the same implementation of the model outlined above and described in much greater detail in [15]. In a situation where \( A \) can fully observe \( B \) (i.e., access \( B \)’s intention state), \( A \) has the potential of recognizing \( B \)’s intention because \( A \) has the same model of intention, albeit a separate instantiation and history. The challenge in the real world is that no two agents can fully observe the other’s intention, and even if they could they would have difficulty interpreting the model because no two agents have exactly the same model (arguably, the same goes for humans). This problem is further exacerbated due to the extreme differences between how human intention is realized in the human brain and how automated agent intention is modeled.

Humans manage ToM because human \( A \) can assume that if agent \( B \) looks and acts like a human, then its internal cognitive capabilities can be assumed to be like human \( A \)’s capabilities through a kind of cognitive projection on \( A \)’s part (though this doesn’t work for everyone [11]), and though \( B \)’s internal cognitive capabilities are not fully observable to \( A \), \( A \) can still approximate \( B \)’s intention at any given moment by observing \( B \)’s behavior. Moreover, \( A \) can only approximate \( B \)’s intention through observation because \( A \) has produced behaviors in the past that \( A \) can draw from to interpret the intention of \( B \). To illustrate this further, [20] showed that human children cooperate with others (e.g., a child opens a door for someone whose hands are occupied carrying an object) because they can recognize the intention of others; i.e., children at a certain age have enough experience of their own intentions to recognize intentions in others [21]. This poses, I believe, one of the biggest challenges of machine ToM: a model of intention (i.e., of plans of actions and follow-through in that intentions are choices with commitment) necessarily needs to itself be able to intend to perform an action in order to overcome a problem, provide a screen of admissibility, track success, and distinguish between real and possible events (i.e., the four functional roles explained above). Furthermore, it must have ample experience actually choosing and carrying out actions in the past to give the agent learned experience of what it means to intend before it can successfully recognize intention in others.

Fortunately, choosing and planning actions has a long history in the robotics field, particularly in the domain of navigation because robots have to break down high-level goals to low-level sequences of actions (for an overview, see [22], [23]) which could be leveraged for recognition

---

1See section 1.5 of [15].
of intention in others. Some work has attempted to directly model shared intentions of humans and robots [24] that has direct implications for machine ToM, but more work is needed to explore learning intention. At the moment, existing models of machine ToM are distinct from modules that represent and plan actions, though [17] is an example of an action-oriented approach that is inspired by ToM.

IV. LEARNING THEORY OF MIND

Recognition of intention also has a long history in robotics; in particular, the field of developmental robotics. Interestingly, [25] eludes to ToM in many places, yet no published work cited therein claims that ToM is modeled directly. Rather, ample work has explored joint attention—the shared focus of two individuals on an object—a developmental precursor to intention, and therefore ToM. According to [26], in child development, joint attention develops through several stages: early identification with other persons (0-3 months), animate-inanimate distinction and physical vs. social causality distinction (6 months), first goal-oriented behavior (9 months), goal understanding behavior understood as goal directed (12 months), and intentional understanding of the same goal for different action plans (18 months). All of these steps must take place long before a child develops ToM at around age 4-5 [27]. Furthermore, joint attention and shared goals are important in spoken interaction as the building blocks of mutual understanding between people [28].

Bridging joint attention with ToM are two known ToM theoretical frameworks. The first is Leslie’s model that builds from theories of agency [29]. Leslie splits ToM into three modules: Theory of Body which deals with physical, perceptual events, Theory of Mind 1 which considers ToM intention within actions, and Theory of Mind 2 which focuses on attitudes and beliefs of agents.

The second theoretical framework is Baron-Cohen [2]. That work presents a framework for a “mindreading” (i.e., ToM) system that includes four parts: the intentionality detector, the eye direction detector, the shared attention mechanism and the theory-of-mind mechanism, which have parallels to recognition of intention, though the focus here is on the requirements for ToM (p.127). They review literature to show what capabilities children have and at what age. Importantly, the shared attention (i.e., joint attention as explained in the above paragraph) is a necessary precursor to the ToM mechanism in all cases. Furthermore, highly correlated with learning joint attention is the ability to observe the visual attention of others and draw inferences about what their visual attention means (i.e., children can triangulate the line of regard of others even when their eyes are not in view, which is quite remarkable). I find this intuitive because the eyes have only a narrow field of focused vision, so what someone is looking at must have something to do with what they are believing, desiring, and intending. Though it is not a complete picture of observing another mind, it is a developmental starting point.

What implications does an understanding of how children develop joint attention have for machines that can learn a model of ToM? Eyes seem to be important for early stages of learning joint/shared attention (at the very least, recognition of the line of regard in others), which is a precursor to recognizing that others have goals and intentions (though the intentions can only be recognized on the basis of committing to taking actions) which is a precursor to developing a ToM. This seems to suggest that embodiment is required; not just recognition of the embodiment of others, but the agent that develops a ToM must be embodied because it must have the capacity for joint attention and the ability to learn how to commit to taking actions, and that can only happen if the agent can actually enact those actions. [30], for example, claims that first-person play (body required!) is likely important in developing ToM (potentially due, I conjecture, to the fact that play is one way that children instantiate their intentions—they commit to and follow through on their decisions), though introspection is the key (p.130). Indeed, recent work has argued that embodiment is a requirement for cognition more generally [31], [32], [25], [33], though the claim of embodied cognition has its critics.

Most recent work in machine ToM seems to follow Leslie’s model, albeit only focusing on ToM 2 (i.e., beliefs and attitudes) instead of actions and intentions. An exception that is well worth mentioning is [34] that considered both models in human-robot interaction tasks. Their method followed a developmental-esque path that included visual attention, recognition of eyes and faces of humans, discriminating animate from inanimate objects (ToM only applies to animate objects), as well as following gaze and deictic gestures as precursors to joint attention. I suggest that this work be carefully considered by those who explore ToM either from the Leslie framework, the Baron-Cohen framework, or both, particularly if learning (including intention from joint attention) is the goal. Furthermore, I suggest that [34] helps fill in the joint attention requirement that would complement other, more recent models that tend to focus on Leslie’s ToM 2 model in their framework.

V. DISCUSSION

Intention is choice with commitment, and the learning intention of others involves joint attention, yet neither of these are considered in any of the recently proposed models of machine ToM. Focus has been so far on belief states (with the possible exception of [17]), which has a long and useful history in some fields (e.g., belief state tracking in spoken dialogue) and is an obvious place to start for considering ToM because modeling belief states could be a precursor to modeling intention, but my position is that any claim to modeling machine ToM must go beyond modeling belief states and include an understanding of joint intention and intention as defined above.

There are other open questions relating to ToM that have implications for machine ToM that I do not explore in depth here. For example [35] looks at how language plays a role in ToM development, the question is in what way and in
what degree. For instance, the understanding of terms at a semantic level like *want, think, misunderstand* and *lie* could play a role because when a child understands that people have wants, desires, and that they can tell untruths that child must understand that others might maintain a state of understanding and intention of the world that differs from that of the child. Moreover, [36], [37] have shown that when machines misunderstand humans or recognize the situation or intention of humans, a good way to observe the mind of a human is simply to ask them. Situated dialogue where an (embodied) agent interacts with humans is the same setting where joint attention and intention are learned in children. Other recent work [38], [39] explored how ToM relates to human-robot trust and showed that perceptions of higher ToM on the side of the robot led participants to trust the robot more in pricing and investment tasks. It’s not clear what degree trust is necessary for ToM, but if concepts like *misunderstand or lie* matter for ToM [35], then trust-related concepts might also be important.

Finally, I am not claiming that a model of machine ToM must be a model that learns in the same progression that human children do, nor am I claiming that a model of machine ToM needs to be a model trained with data, but I am suggesting that a full model of machine ToM would be more holistic if it understood joint attention and intention as defined above. I appreciate the trend that ToM has received increased attention in multiple research communities, but I am also cautious about what it means to claim that a machine has the capacity for ToM.

**References**