Theory of Mind: A Familiar Aspect of Humanity to Give Machines

Joel Michelson

Vanderbilt University Department of Computer Science
Nashville, Tennesse, USA
joel.p.michelson@vanderbilt.edu

Abstract

My research focuses on machine models of theory of mind, a set of skills that helps humans cooperate with each other. Because these skills present themselves in behavior, inference-based measurements must be carefully designed to rule out alternate hypotheses. Producing models that display these skills requires an extensive understanding of experiences and mechanisms sufficient for learning, and the models must have robust generalization to be effective in varied domains. To address these problems, I intend to evaluate computational models of ToM using a variety of tests.

Introduction

Theory of mind (ToM) is a set of social reasoning skills which involve making inferences about the mental states of other beings, e.g. intent, perception, beliefs, etc. Typically developing human toddlers acquire ToM skills at specific developmental stages. Certain groups of people, like those on the autism spectrum, are prone to delayed or impaired ToM development. Other animals, including chimpanzees, dogs, and corvids, have also had their ToM skills rigorously tested, but humans generally surpass these other species in ToM development at a young age.

In spite of the wealth of information regarding ToM, many details about these skills remain unclear, including how they arise, how their driving mechanisms operate, and why they present themselves differently across groups and individuals. The former topic—how ToM skills are learned on both the evolutionary and developmental timescales—is of particular importance for teaching these skills to both artificial agents and humans. My research is motivated by the belief that computational models of ToM skills will lead to valuable insights across these points.

Research Questions

The problem I address in my research is the lack of applicable knowledge about ToM skills. My core hypothesis is that this knowledge can be found through computational modeling, but only after key details about these skills are investigated, including how they can be evaluated, how they can be taught and learned, and how they differ across the various domains in which they find use.

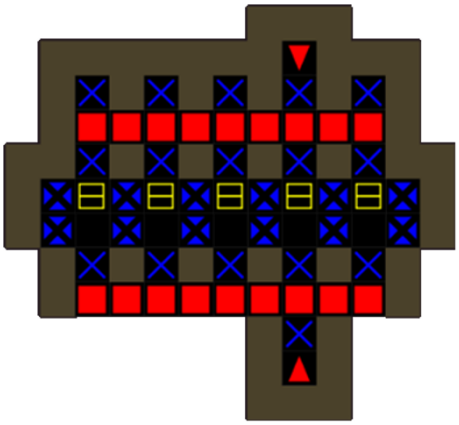

Figure 1: A screenshot of the Standoff environment as observed by one agent. Success depends on inference of the other agent's knowledge and future behavior.

Research Question 1: Measuring ToM Skills: *How can ToM skills be measured in a way that rules out false positives and alternate hypotheses?*

Tests for ToM should ideally involve careful control settings to rule out the null hypotheses, that a subject's behavior is governed by reasoning that does not use ToM. The space of these hypotheses is quite broad, as even the subtlest of behavioral cues could supply a wealth of task-relevant data. As long as measurements are based on observation of behaviors alone, the necessity of carefully designed tasks and rigorous control scenarios remains.

Research Question 2: Learning ToM Skills from Experience: *How do different knowledge representations and learning experiences interact to produce ToM skills?*

Numerous hypotheses attempt to explain the learning process of and the mechanisms behind ToM skills like simulation theory, which emphasizes the importance of pretend play. By developing and evaluating models, both novel and inspired by these hypotheses, I hope to uncover insights into their biological plausibility and computational functionality.

Research Question 3: ToM Skills' Generalization: *To what extent are ToM skills task- or domain-specific?*

ToM covers a variety of interrelated skills and presents itself in multiple domains, so one important aspect of ToM skills is robust generalization. I hope to investigate the extent to which ToM tasks of different kinds and in different domains pose fundamentally different challenges.

RQ1: Measuring ToM Skills

Related Work: The most well-known test for ToM skills is the Sally-Anne test, developed by Wimmer and Perner, which assesses a subject's ability to attribute false beliefs about object location (1983). Comparative and developmental psychologists have since designed many other ToM tests. Penn and Povinelli (2007) describe the issues inherent to common methods and propose solutions. A recent survey (Gurney and Pynadath 2022) highlights the need for standardized ToM measures in the AI community.

Progress: I have implemented and published the Standoff environment (https://github.com/aivaslab/standoff), a gridworld POMDP setting inspired by real-life tests for ToM skills. Originally built to imitate Competitive Feeding (CF), a test designed to test chimpanzees' ability to attribute "seeing" and "knowing" to other chimpanzees, Standoff uses a set of tasks described by Penn and Povinelli (2007) to meaningfully measure agents' ToM skills based on their actions.

Anticipated Progress: I intend to implement additional measures to quantify agents' behaviors in the Standoff environment, answering questions like: *does the subject avoid its conspecifics when it should?* or *does performance vary across different kinds of false beliefs?*. These measures will aid ablationary experiments by describing how and when agents display unsuccessful or unexpected behavior.

Future Work: Additional environments and improvements upon existing environments will serve as tools with concrete purposes. Building these tools will require a deep literature review of ToM and ToM-adjacent tasks, covering developmental and comparative psychology, neuroscience, cognitive science, and machine learning (ML). Evaluating human subjects will allow model performance to be compared with humans, in addition to allowing models to interact with human actors to evaluate ToM-for-humans abilities.

RQ2: Learning ToM Skills from Experience

Related Work: ToM skills have recently become a popular subject of study for multi-agent environments. Rabinowitz et al. developed ToMNet, a supervised learning system designed to predict the mental states of agents in a gridworld environment (Rabinowitz et al. 2018). To achieve this effect, their model makes use of explicit representation of an agent's inner states given sequences of past behaviors.

Progress: Ongoing research using Standoff involves generating a strong baseline for comparison between model architectures and training methods. I have thus far found success using standard off-the-shelf RL algorithms and models at a precursor navigation task, but none perform consistently on the tasks which measure ToM skills.

Anticipated Progress: Computational models of ToM skills, which range from tabula rasa ML algorithms to hard-coded decision makers, are primary candidates for a near-future goal of replication and comparison. Ablationary experiments will help understand models' performances.

Future Work: I intend to examine the psychological theories of ToM learning and mechanisms by implementing cognitive models and exposing them to the Standoff testbed. Based on experimental results, I will also design and test novel models of ToM skills.

RQ3: ToM Skills' Generalization

Related Work: ToM skills are varied, as are the tests designed to measure them. Rusch et al. (2020) provide a typology for classifying ToM and ToM-adjacent tasks in terms of interactivity and uncertainty, but they also present themselves in multiple domains. Verbal question-answering systems have been benchmarked recently using Sally-Anne scenarios, similarly to how they are used to test verbal human subjects (Grant, Nematzadeh, and Griffiths 2017).

Progress: As with skill acquisition, comparison between domains requires a strong baseline. Minor additions to the Standoff CF scenarios test for a variety of ToM skills.

Anticipated Progress: After establishing and publicizing a set of baseline CF models, I intend to explore the space of modifications to Standoff that allow for testing a variety of ToM skills, with a focus on transfer across task variants.

Future Work: Establishing and publicizing a set of baseline models for the CF task is an important step towards comparing performance across different domains. However, narrow test cases like Sally-Anne and CF do not necessarily capture the full range of ToM and ToM-adjacent skills. ToM tasks which bear little resemblance to CF or exist in different domains require their own specialized environments for training and evaluation.

References

Grant, E.; Nematzadeh, A.; and Griffiths, T. L. 2017. How Can Memory-Augmented Neural Networks Pass a False-Belief Task? In *CogSci*.

Gurney, N.; and Pynadath, D. V. 2022. Robots with Theory of Mind for Humans: A Survey. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 993–1000. IEEE.

Penn, D. C.; and Povinelli, D. J. 2007. On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480): 731–744.

Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. A.; and Botvinick, M. 2018. Machine theory of mind. In *International conference on machine learning*, 4218–4227. PMLR.

Rusch, T.; Steixner-Kumar, S.; Doshi, P.; Spezio, M.; and Gläscher, J. 2020. Theory of mind and decision science: towards a typology of tasks and computational models. *Neuropsychologia*, 146: 107488.

Wimmer, H.; and Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1): 103–128.