The Case for More Comprehensive Evaluations of Machine Theory of Mind

Jason R Wilson[1], **Irina Rabkina**[2], **Zach Locher**[1], **Laura M. Hiatt**[3]

[1]Franklin & Marshall College
[2]Barnstorm Research
[3]Navy Research Lab
jrw@fandm.edu

1 Introduction

Theory of mind (ToM) is considered to be the ability to understand that others can have differing beliefs, knowledge, desires, and intentions. ToM reasoning, in turn, is the inference of any one (or more) of these aspects of cognition in another agent. In much current work, such reasoning is done based on observations of the agent's actions. The representation of the action observations is varied across the literature, and can include an agent's trajectory in 2D or 3D space, symbolic or predicate representations of their actions, or even a textual story describing what the agent is doing.

In this work, we instead focus on situations where such observations, alone, are not sufficient for ToM reasoning. There are many real-world analogs of this, such as inferring an agent's intent when they provide noisy or underspecified instructions (Ying et al. 2024), or when observations alone are not sufficient to disambiguate between different hypotheses of the agent's task.

To this end, we present a benchmark task and associated dataset involving Tangram puzzles. In such puzzles, different baseline polygons are arranged to form a larger, target shape. In part because of the intricacy of these puzzles, effective ToM reasoning is challenging, if not impossible, without knowledge of the task. In addition to supporting complex ToM reasoning, the accompanying dataset facilitates the investigation into how ToM reasoning connects to the subsequent actions of the ToM reasoner.

2 Current ToM Evaluations

Many machine ToM evaluations focus on predicting an agent's intentions based on observations of their actions, such as trajectories of their movement in a 2D grid world represented at either the pixel or lifted symbolic level (e.g., Shum et al. 2019; Rabkina and Forbus 2019; Rabinowitz et al. 2018; Nguyen and Gonzalez 2020, see Mao, Liu, Ni, Lin, and He (2024) for comprehensive list). For example, the stag-hunt game (Shum et al. 2019; Rabkina and Forbus 2019) involves predicting which target each observed agent intends to capture, and whether any agents intend to cooperate to capture a larger target. Observations consist of discrete agent movements (i.e., whether each agent moved up, down, left, or right). Due to the simplicity of the grid world and constraints of the task, the lower-level intentions that can be inferred from these movements (i.e., moving toward or away from a target) map directly to the higher-level intentions of cooperating or capturing a specific target. Similarly, the food truck problem (Baker, Saxe, and Tenenbaum 2011) involves agents attempting to purchase food from one of several food trucks parked on a 2D map. The associated ToM task is to predict which food truck each agent is targeting based on their movements. While movement in the food truck domain is continuous, and therefore less constrained, the task nonetheless primarily involves reasoning about which food truck(s) each agent can see and/or is moving toward. Such tasks capture the inferences that can be made based off of agents' movements with respect to their surroundings, but are not representative of broader ToM reasoning.

The recent popularity of large language models (LLMs) has also ushered in text-based ToM benchmarks. These largely fall into one of two categories: narrative descriptions of movements similar to the 2D world benchmarks described above (e.g., Jin et al. 2024; Shi et al. 2024; Verma, Bhambri, and Kambhampati 2024) or textual variations on the Sally-Anne false belief task (Baron-Cohen, Leslie, and Frith 1985). In the FANToM false belief benchmark (Kim et al. 2023), the LLM is told a story in which a character is not present for part of a conversation. The LLM is then asked questions about the character's knowledge of the conversation, including information that was shared while they were not present. Xu et al. (2024) similarly target the story false belief dimension of ToM reasoning in LLMs, while Nematzadeh et al. (2018) and Le, Boureau, and Nickel (2019) do so for other neural models and Hiatt and Trafton (2010) and Rabkina et al. (2017) do so for ToM models in cognitive architectures. Notably, all of these evaluations focus on reasoning based on descriptions of the agent's actions: where they moved to or whether they were present.

Intent recognition from actions has more explicitly been studied as goal and plan recognition (see Mirsky, Keren, and Geib 2021). However, these evaluations are not typically framed as ToM problems. One exception is the Minecraft dataset presented in Rabkina et al. (2020) and Rabkina et al. (2022). This dataset combines reasoning about an agent's movements and knowledge about the task to predict its goals. For example, most goals can be accomplished by the

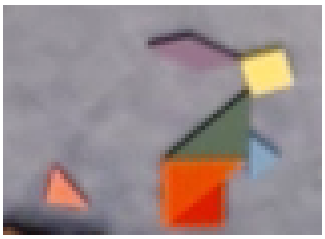

Figure 1: Screenshot of a nearly complete rabbit puzzle.

agent moving towards a farm, but information about harvested items must be combined with task knowledge to reason about which item the agent is trying to obtain. However, it does not require explicit reasoning about the observed agent's knowledge or beliefs.

3 Tangrams As Alternative Evaluation

We propose a new benchmark that requires reasoning over beliefs and knowledge. The benchmark uses a new Tangram dataset, which consists of observations of a child assembling tangram puzzles with the assistance of either a social robot or human instructor. Tangram puzzles are constructed from seven basic pieces that are assembled to resemble an object (e.g., cat, rabbit; see Fig. 1). The acting and observing agents need to know the structure of the resulting puzzle (i.e., where do all the puzzle pieces belong). The observing agent (i.e., the instructor) has this knowledge, but the acting agent (i.e., the child) may have incorrect beliefs regarding positions of the pieces. Additionally, the position of some pieces is ambiguous because there are pairs of pieces of the same size that may be swapped for each other. When the observing agent recognizes what the child is doing or any misconceptions they may have, it provides verbal assistance. The complexities of this task allows our dataset to support goal recognition, plan recognition, false belief reasoning, and observer intervention inference.

The dataset is based on videos of child-robot interactions collected as part of a previous study (redacted).For each interaction, we hand-coded the videos to record a series of observations. A single observation includes the ID of the interaction, a timestamp, the target puzzle, the structure of the puzzle, and the instruction provided (if any). The instruction is represented in three forms: a transcription of what the robot or human said, a symbolic representation of the instruction, and the intended change in the puzzle structure.

To represent the structure of the puzzle, we developed a representation designed to model the spatial relationships between 2D, non-overlapping, polygons. The representation is an extension of RCC-8 that expands upon the types of relationships defined in the "externally connected" relation (Randell, Cui, and Cohn 1992; Cohn et al. 1997). Polygons are represented as a set of edges, where each edge is a pair of vertices. Our representation uses these edges and vertices by describing all 9 possible relations between edges and vertices. For example, Fig. 1 includes a "vertex connection" between vertices of the yellow square and the green triangle. Since the set of representations are mutually exclusive and collectively exhaustive, we can safely assume that any undefined relation is not present.

This dataset and representation support benchmarking the following types of reasoning:

Goal recognition: The first application of this dataset involves inferring the goal of the task, where the goal is defined as which puzzle the child is assembling.

Plan recognition: The dataset can be used to infer next steps in the child's plan, based on the steps they have taken so far. This implicitly requires goal recognition.

False belief reasoning: The dataset may be used for reasoning about false beliefs. Each incorrectly placed piece suggest the child may have a false belief regarding the relation between that piece and the puzzle. Perhaps more interesting is that a child may be working with the false assumption that all puzzles pieces are initially provided. However, one piece is always hidden. The observing agent needs to recognize when the child is working with this false belief if it were to help in rectifying the belief.

Observer intervention: Decision-making for intervention requires at least reasoning about the child's goals and intentions. The child may also have some false beliefs regarding the puzzle or the pieces. For effective intervention that aligns with a students needs and preferences, the observer may need to reason about a child's desire for help. The agent may also want to reason about when to help or how much help to give, since helping too soon or too much may negatively impact the child's autonomy (Wilson et al. 2018; Wilson, Aung, and Boucher 2022).

4 Discussion

A significant amount of existing benchmarks and evaluation scenarios involve a common pattern of observing an agent's trajectory in relation to some goal locations or textual stories about what an agent is doing. Our dataset introduces new challenges by requiring knowledge-based reasoning about a task and the spatial arrangement of objects to infer the intentions of another agent. This dataset, featuring sequences of actions to assemble Tangram puzzles, also supports reasoning about a child's beliefs regarding the task. Like previous work that has used plan recognition to decide when or how to intervene (Weerawardhana, Whitley, and Roberts 2022; Freedman and Zilberstein 2017), our Tangram dataset allows for a connection between ToM and intervention.

Many real world applications of ToM involve observing an agent in a physical environment. While the Tangram dataset does not include data that may be used to reason about physical trajectories, there is a physical component to the observations that requires reasoning about spatial arrangements instead.

The Tangrams dataset is not without its limitations. Only two goal puzzles (rabbit or cat) are currently included. Additionally, the dataset includes data extracted from videos. We cannot provide the videos themselves due to the presence of minors in them.

Nonetheless, the Tangram dataset has qualities that will help push forward AI for ToM. The AI community has made incredible gains in modeling and simulating ToM, and there

is now a great opportunity to expand these capabilities to consider additional aspects of ToM. Most importantly, there is a need to recognize the role of the background or task knowledge held by the observer and the acting agent.

References

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 2: 37–46.

Cohn, A. G.; Bennett, B.; Gooday, J.; and Gotts, N. M. 1997. Qualitative spatial representation and reasoning with the region connection calculus. *geoinformatica*, 1: 275–316.

Freedman, R.; and Zilberstein, S. 2017. Integration of planning with recognition for responsive interaction using classical planners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Hiatt, L. M.; and Trafton, J. G. 2010. A cognitive model of theory of mind. In *Proceedings of the 10th international conference on cognitive modeling*, 91–96. Citeseer.

Jin, C.; Wu, Y.; Cao, J.; Xiang, J.; Kuo, Y.-L.; Hu, Z.; Ullman, T.; Torralba, A.; Tenenbaum, J. B.; and Shu, T. 2024. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R. L.; Kim, G.; Choi, Y.; and Sap, M. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

Le, M.; Boureau, Y.-L.; and Nickel, M. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5872–5877.

Mao, Y.; Liu, S.; Ni, Q.; Lin, X.; and He, L. 2024. A Review on Machine Theory of Mind. *IEEE Transactions on Computational Social Systems*.

Mirsky, R.; Keren, S.; and Geib, C. 2021. Introduction to Symbolic Plan and Goal Recognition, ser. *Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers*, 16(1).

Nematzadeh, A.; Burns, K.; Grant, E.; Gopnik, A.; and Griffiths, T. L. 2018. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*.

Nguyen, T. N.; and Gonzalez, C. 2020. Cognitive Machine Theory of Mind. In *CogSci*.

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.

Rabkina, I.; and Forbus, K. D. 2019. Analogical reasoning for intent recognition and action prediction in multi-agent systems. In *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*, 504–517. Cognitive Systems Foundation Cambridge.

Rabkina, I.; Kantharaju, P.; Wilson, J. R.; Roberts, M.; and Hiatt, L. M. 2022. Evaluation of goal recognition systems on unreliable data and uninspectable agents. *Frontiers in Artificial Intelligence*, 4: 734521.

Rabkina, I.; Kathnaraju, P.; Roberts, M.; Wilson, J.; Forbus, K.; and Hiatt, L. 2020. Recognizing the goals of uninspectable agents. *Advances in Cognitive Systems*.

Rabkina, I.; McFate, C.; Forbus, K. D.; and Hoyos, C. 2017. Towards a computational analogical theory of mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39.

Randell, D. A.; Cui, Z.; and Cohn, A. G. 1992. A spatial logic based on regions and connection. *KR*, 92: 165–176.

Shi, H.; Ye, S.; Fang, X.; Jin, C.; Isik, L.; Kuo, Y.-L.; and Shu, T. 2024. MuMA-ToM: Multi-modal Multi-Agent Theory of Mind. *arXiv preprint arXiv:2408.12574*.

Shum, M.; Kleiman-Weiner, M.; Littman, M. L.; and Tenenbaum, J. B. 2019. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6163–6170.

Verma, M.; Bhambri, S.; and Kambhampati, S. 2024. Theory of Mind abilities of Large Language Models in Human-Robot Interaction: An Illusion? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 36–45.

Weerawardhana, S.; Whitley, D.; and Roberts, M. 2022. Models of Intervention: Helping Agents and Human Users Avoid Undesirable Outcomes. *Frontiers in Artificial Intelligence*, 4: 723936.

Wilson, J. R.; Aung, P. T.; and Boucher, I. 2022. When to help? A multimodal architecture for recognizing when a user needs help from a social robot. In *International Conference on Social Robotics*, 253–266. Springer.

Wilson, J. R.; Lee, N. Y.; Saechao, A.; Tickle-Degnen, L.; and Scheutz, M. 2018. Supporting human autonomy in a robot-assisted medication sorting task. *International Journal of Social Robotics*, 10: 621–641.

Xu, H.; Zhao, R.; Zhu, L.; Du, J.; and He, Y. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. *arXiv preprint arXiv:2402.06044*.

Ying, L.; Liu, J. X.; Aarya, S.; Fang, Y.; Tellex, S.; Tenenbaum, J. B.; and Shu, T. 2024. SIFToM: Robust Spoken Instruction Following through Theory of Mind. In *Proceedings of the AAAI Fall Symposium on Unifying Representations for Robot Application Development (UR-RAD)*.