Knowing me, Knowing you: the Long Way to Build Recursive Mind Models

Jingze Zhang
Department of Automation
Tsinghua University
jz-zhang21@mails.tsinghua.edu.cn

Abstract

Can machines think? Can machines speculate about the behavior of intelligent agents? Can machines infer the inner psychological activities of other intelligent agents based on their behavior and provide reasonable explanations? With these questions in mind, we will explore the current technical approaches and developments in the computational theory of mind. We will particularly focus on the current progress in recursive mind inference, contemplating how to construct a unified and self-recursive model, which should not only predict the behavior of intelligent agents during the forward process but also engage in reasoning about the desires, beliefs, and intentions of other intelligent agents based on mental theories and observed phenomena.

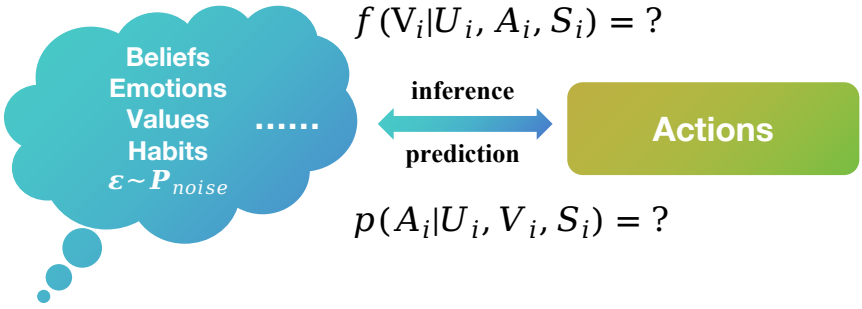

Figure 1: A illustration of ToM. The illustration is from probablistic perspective, where the agent predict another agents' actions via partial observation and inference the habit, value, beliefs, goals or emotions of another agent. The prediction and inference process is bi-directional, and we want a unified mind model to do the prediction and inference process mentioned above. We usually call the model *theory of mind, ToM*

1 Introduction

Each rational human individual possesses a unique model of the *theory of mind* (ToM), which refers to our human's ability to represent the mental states of others[5]. Consequently, it empowers us to comprehend and predict behavior based on these mental states and contextual information, facilitating rational decision-making across various scenarios. While individual's theory of minds may vary in aspects such as beliefs and desires, they exhibit a remarkable alignment in terms of social values, life activities, and other attributes[6]. These shared characteristics in the theory of mind enable

us to understand the behavioral traits and intentions of others, fostering smooth cooperation and contributing to the development of humans as intelligent beings with highly social attributes[4, 6, 7].

Upon further contemplation of this issue, we realize that the process by which humans infer the mental activities of others is a kind of magical endeavor. Unlike *the Trisolarans* depicted by Chinese author Xinci Liu, who possess the ability to perceive the thoughts of others, we lack the capacity to observe the information flow in others' brains, the intricate neural activities within the cerebral cortex, or the representation of all states and personalities in latent spaces. Nevertheless, within a short span of time, we can grasp the mental states of other intelligent entities, such as their desires, beliefs, and intentions. We can achieve this by simulating the inner worlds of others within our own brains[2]. This suggests that the model of theory of mind adopts a level of abstraction sufficient for intelligent agents. Understanding the form and implications of this abstraction and representing similar abstractions through computational methods becomes crucial.

For humans, possessing a Theory of Mind aligned with mainstream values is a prerequisite for integrating into societal life and establishing cooperative relationships. As artificial agents become part of the human world, the need to comprehend them is increasingly emphasized.

However, constructing a rich and flexible Machine Theory of Mind is an immensely challenging process. Currently, there is considerable research focus on the Machine Theory of Mind. These approaches primarily involve constructing probabilistic rational planner models through probability-based methods in a recursive way. And in the paragraph below, we will dive into these mind models and discuss the insights and limitations of these past methods.

2 A Computational Representation as a POMDP or I-POMDP

A *particial observed markov decision process* (POMDP) is similar to an MDP, whereas the agent in the POMDP can only observe part of the environment but the system dynamic is determined by both the observed and the unobserved part. A POMDP can be formulated as 7 parameters:

$$POMDP_i =< S, A, T, R, \Omega, O, \gamma > \tag{1}$$

In the equation above, $M_i =< S, A, T >$, is the set of states, actions, and conditional transition probabilities between states, which is the same as the definition of MDP. However, O means a set of conditional observation probabilities and Ω means the set of observations of the agent. Specifically, in many tasks, the reward in the same state is different among the agents, which also means that the reward function is bound with agents rather than the given environment. The intuition of the setting comes from the characteristic of the real world settings.

Due to the fact that agents can only observe partial features of the environment, it is crucial for them to develop a belief about the environment based on the relevant information they have acquired. Typically, we employ the following methods to model and iterate the beliefs of an intelligent agent. The belief of state s is denoted as $b(s)$, meaning the estimation of current state probability distribution from current observation. Here is the iteration process of the belief function:

$$b'(s') = \eta O(o|s', a) \sum_{s \in S} T(s'|s, a) b(s) \tag{2}$$

3 The Recursive Modeling Method

According to [3], in the context of modeling the theory of mind, *recursive modeling method* (RMM) refers to a computational approach that aims to simulate and understand how individuals attribute mental states, such as beliefs, desires, intentions, and emotions, to themselves and others.

Specifically, the RMM is a recursive process in inferring other beliefs about the environment. A illustrated demostration is shown in figure 2.

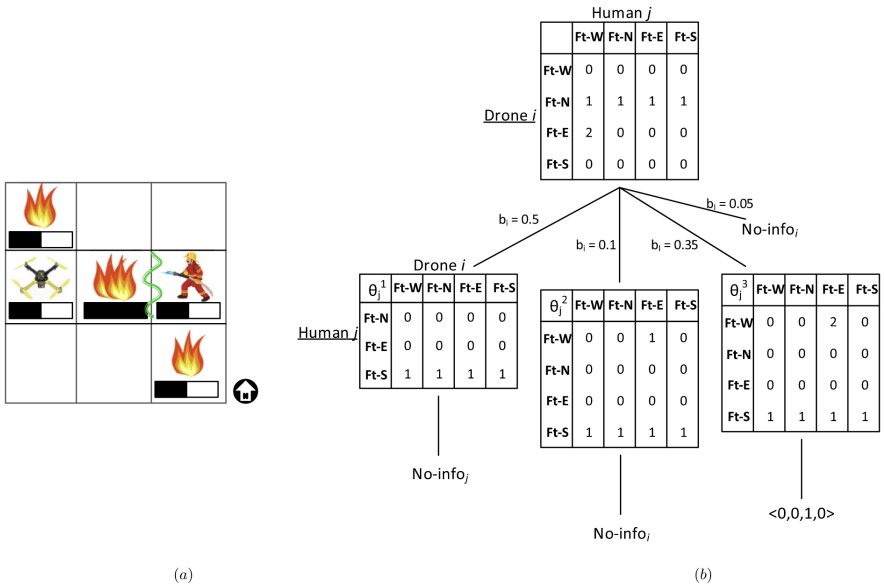

Figure 2: A illustration of recursive modeling method. The figure is from [1]

4 Machine Theory of Mind: Still in its Infancy

Based on our current literature review, most modeling approaches for theory of mind have been represented as a partially observed Markov process. However, this modeling is an excessive abstraction of real-world scenarios. Under this level of abstraction, current intelligent agents can achieve impressive results in game settings. However, significant challenges arise in the context of real human-machine interactions, where almost any real world problem cannot be totally depicted as a MDP. We are faced with a huge gap between the game scenarios and real world settings.

What's more, the process of human cognitive construction is not a completely recursive one. Although our theory of mind also involves speculating about what others think and perceive, as well as how others perceive us, the depth of this "recursive" thinking typically does not exceed two levels. Furthermore, our serial thinking patterns and the limited memory and computational resources of our brains make it difficult to sustain too many levels of recursive thinking calculations. So the computational cost limitations of the recursive modeling method is the barrier for its generalization in real-world settings.

5 Summary and Conclusion

In this essay, we conducted a literature review to explore the current approaches for modeling theory of mind using the partial observation Markov decision process. However, these methods have shown limited generalization capabilities, achieving decent results only in simple game settings. Additionally, we analyzed the intuition and limitations of recursive modeling for Theory of Mind. The problem with recursive modeling lies in the significant divergence between these models and human cognition, often resulting in substantial computational overhead. Based on the author's limited knowledge, it is suggested that future research on the psychological aspects of theory of mind is crucial for computational theory of mind.

References

[1] Prashant Doshi, Piotr Gmytrasiewicz, and Edmund Durfee. Recursively modeling other agents for decision making: A research perspective. *Artificial Intelligence*, 279:103202, 2020. 3

[2] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998. 2

[3] Piotr J Gmytrasiewicz and Edmund H Durfee. Rational coordination in multi-agent environments. *Autonomous Agents and Multi-Agent Systems*, 3:319–350, 2000. 2

[4] Esther Herrmann, Josep Call, María Victoria Hernández-Lloreda, Brian Hare, and Michael Tomasello. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, 317(5843):1360–1366, 2007. 2

[5] Olivier Houdé and Grégoire Borst. *The Cambridge handbook of cognitive development*. Cambridge Handbooks in Psychol, 2022. 1

[6] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018. 1, 2

[7] Michael Tomasello. Social cognition and metacognition in great apes: a theory. *Animal Cognition*, 26(1):25–35, 2023. 2