# Theory of Mind:from Philosophical Theory to Computational Model

**Jinfan He**
Peking University
2200012979@stu.pku.edu.cn

## Abstract

We delves into Theory of Mind (ToM) as the ability to understand and predict one's and others' mental states. We introduce two main hypotheses: the Simulation-Theory and the Theory-Theory. Simulation-theory emphasizes the use of cognitive or sensory mechanisms to simulate other people's mental states. Conversely, the Theory-Theory is based on comprehending a theory of mind and inferring others' mental states through reasoning. Finally, we propose an AI ToM model framework based on Bayesian learning model that takes into account the effects of different beliefs.

## 1 Introduction

ToM is the ability to understand and predict one's own and other people's mental states, including emotions, intentions, desires, beliefs, and to use this information to predict and explain the actions of others. In some cases, it is also known as mindreading or folk psychology.Humans have been able to create vast social organizations because they have ToM.

Over the past century, there has been a desire to use research on ToM to explain the reasons behind certain human decisions, and to apply this ability to AI, enabling AI to better predict human intentions and serve humans more effectively. Currently, there are two dominant hypotheses within ToM: Simulation-Theory and Theory-Theory. We introduce both hypotheses in the following sections, and present our own thoughts on constructing artificial intelligence models based on them.

## 2 Two Mainstream Hypotheses

### 2.1 Simulation-Theory

According to common sense, in many cases we understand the mental state of others by empathizing with them or putting ourselves in their shoes. The most straightforward idea is that people understand the mental states of others through mental simulations. To understand the mental state of others, we feed their relevant beliefs into our decision-making mechanism, creating a simulated decision and projected it onto them. This capacity for simulation not only appears in decision-making mechanisms, but can also be applied to our sensory mechanisms.

For example, if Lisa sees a banana, but there is no banana around, I cannot have a similar visual experience. However, I can still imagine Lisa seeing the banana, which is a mental simulation of her visual experience by reusing my visual mechanism. This simulation process is called high-level simulation because it is consciously simulated, independently of the stimulus[1]. On the other hand, a low-level simulation process refers to an unconscious, stimulus-driven simulation process discovered through the discovery of mirror neurons.Neuroscientists found that monkeys reacted similarly to seeing an experimenter grab food as they did to grabbing food themselves[2]. This led to the proposal of the mirror mechanism and the discovery of a similar mechanism in humans, similar to the disgust mirror mechanism.When A sees B's disgusted facial expression, A may also experience disgust unconsciously.By attributing this experience, A deduces that B has experienced something disgusting. This process of mental simulation is not under conscious control.

### 2.2 Theory-Theory

Another mainstream theory is theory-theory, which states that mindreading is based on a fundamental theory of mind.For example, if Lisa still sees a banana,but now my visual image is generated by an informative cognitive process that uses detailed knowledge of vision and incredibly powerful reasoning mechanisms. This process cannot be considered a mental simulation, as it does not reuse the visual mechanism.This process seems to be much more complicated than mental simulation theory.Why do we support the existence of theory-theory instead of attributing all mindreading process to simulation-theory? Imagine you are a tree. In this case, my imagination depends entirely on my understanding of a tree because I obviously can't have the same structure as a tree. these imaginative processes are entirely theory-driven.

The main disagreement among theory-theory proponents is how theory of mind is acquired. There are two main hypotheses. According to Child Scientist Theory-Theory, theory of mind is acquired through hypothesis testing and revision.This process is guided by Bayesian learning mechanism[3]. In contrast, nativist theory-theory suggests that A substantial part of the theory of the mind is innate rather than learned. Alternatively, it posits that the core of the theory of mind is the result of a specialized cognitive module for representing maturing mental states.

## 3 A New ToM Model for AI

By exploring how humans use ToM, we can instruct the direction of constructing artificial intelligence.Based on simulation-theory and theory-theory, we attempt to construct a theory of mind for artificial intelligence using a Bayesian learning mechanism, and use this mechanism to make predictions about the mental states of others from their action sequences and environments.But there are still some problems.

First, the different beliefs of different people may cause them to have vastly different states of mind and make different decisions in the same situation. If we do not know the beliefs of others, the probabilistic model we learn is influenced by the distribution of each belief in the population, rather than the actual probability distribution of other people's decisions.

Secondly, the beliefs of an individual evolve continuously. So we have to evaluate the newly acquired beliefs of others to correct any previously predicted erroneous beliefs.

In fact, when we start to predict others, there is a tendency to assume that they share our beliefs. Therefore, when constructing the prediction model, we can initially add some shared beliefs and subsequently introduce new beliefs or discard erroneous ones during the prediction process.

In summary, we train a Bayesian predictive model influenced by the beliefs.It considers the environment and actions as variables. At each stage, it predicts both beliefs and goals. We update the agent's belief repository with the predicted beliefs and the observations of the newly acquired beliefs of the agent. We can include some initial beliefs in the agent's belief repository and iterate through the agent's belief repository via an inference mechanism after each belief update.

## 4 Conclusion

In this article, we introduced Theory of Mind (ToM) as an ability to understand and predict one's own and others' mental states. We discuss two main hypotheses related to ToM: Simulation-Theory and Theory-Theory. Simulation-theory revolves around the reuse of cognitive or sensory mechanisms to model the mental states of others. It encompasses both high-level and low-level simulation processes.The main difference is whether the simulation process is controlled by consciousness or influenced by stimuli. On the other hand, the Theory-Theory is based on understanding a theory of mind and uses reasoning to infer the mental states of others. The main disagreement lies in whether the theory of mind is innate or acquired through continuous hypothesis testing and refinement.

Finally, we propose a new framework for artificial intelligence ToM models, based on a Bayesian learning model. This model is influenced by different beliefs, predicts both goals and beliefs based on action sequences and the environment, and maintains a repository of beliefs through newly predicted beliefs and inference mechanisms, thus incorporating the potential impact of various beliefs on the predicted objectives.

## References

[1] L Barlassina. Folk psychology as mental simulation. *Engineering*, 2017. 1

[2] V Gallese. Mirror neurons and the simulation theory of mind-reading. 1998. 1

[3] A Gopnik. Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory-theory. 2012. 2