14th national conference of child and adolescent psychiatry

Theory of mind and AI

Can AI truly understand what's on our minds?

Katayoon Razjouyan

Child and Adolescent Psychiatrist

SBMU

20/NOV/2025

Can you guess that what is she doing?

We are inferring something about what is going on in that person's head. But it's more than just pattern matching.

It is not simply memorizing the pattern equals the problem.

It's the ability to attribute mute mental states to other people, like what their intentions are, or their desires, or their emotions, or what they know or don't know.

ToM is in everything we do!

Theory of mind

It is critical foundation for all of our meaningful social interactions because those require you to be able to simulate other people's intentions and emotions and beliefs.

Why are human brains so talented at making theories about other people's minds?

- In psychology, ToM has been extensively studied in a range of scenarios, such as studies of manipulation, secrecy, deception, lying, and misleading behavior.

- The job of intelligent brains is to predict the future.
- It does not come for free.

Main kinds of Theory of Mind

Belief-desire ToM

Around ages 3-4: Understanding that both beliefs and desires influence behavior.

Desired-Based ToM

Around ages 2–3: Early understanding that others have desires that guide their actions

False-belief ToM

Around age 4;:The realization that others can hold beliefs that are false, and that these beliefs will guide their actions.

Visual illustration of the classic Sally-Anne task

1

Sally places a marble in a box.

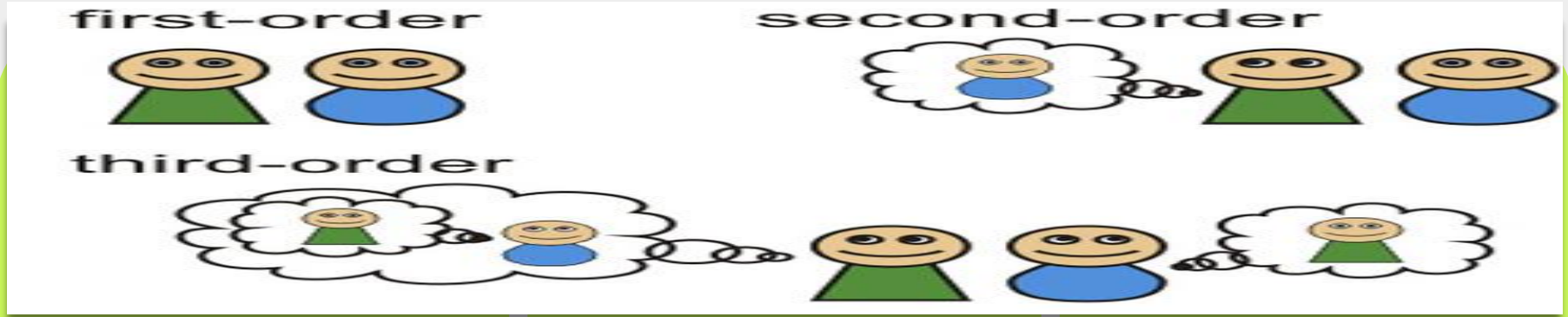
2

The marble is moved to a jar when Sally is not around.

3

Sally comes back. Where does Sally believe the marble is?

Main kinds of Theory of Mind



First-order ToM

- **Around ages 4–5 Understanding what another person thinks about the world (one level of mental state reasoning)**

Second-order ToM

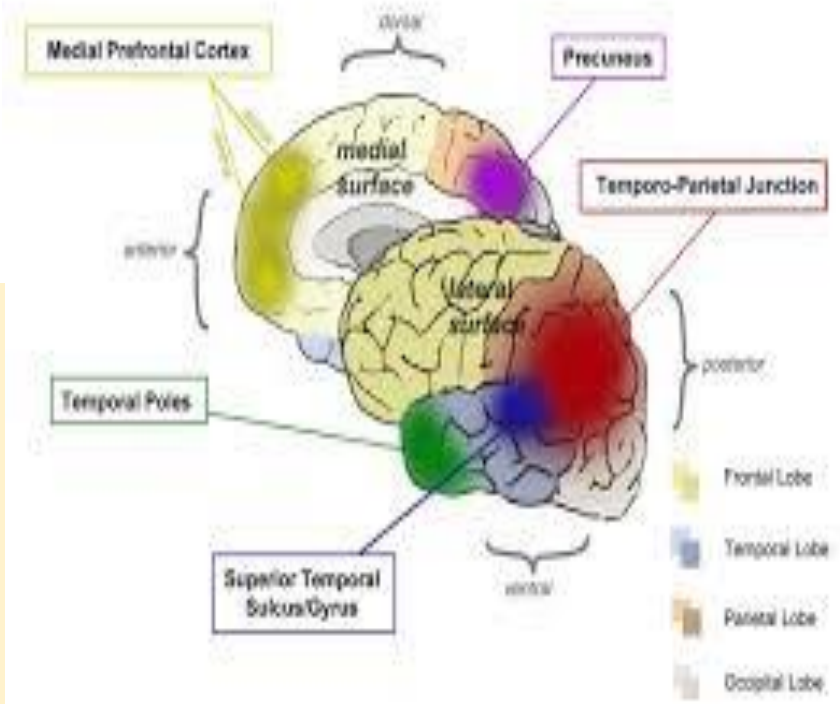
- **Around ages 6-7: Understanding what one person thinks about another person's thoughts (two levels of mental state reasoning).**

Higher-order ToM

- **Later childhood and adolescence; Even more complex reasoning about nested beliefs, sarcasm, or knowledge about someone's beliefs about someone else's beliefs**

T.o.M is a complex network in the brain

- **Medial prefrontal Cx.**: making social judgment
- **TP junction:** understanding perspectives
- **Superior temporal Cx.**: processing social information
- **Mirror Neuron System** (MNS)
- **Mentalizing**
- **Threat and emotion processing circuits**



Today question:

Have current large language models (LLM) come to solve this problem without no instruction?

AI in self-driving car

What is LLM?

- A type of artificial intelligence designed to understand, generate, and manipulate human language at scale.
- They are built by training deep neural networks on vast amounts of text data
- Scale: They typically have billions or trillions of parameters. The large size helps them learn nuanced language patterns, grammar, facts, and even some reasoning abilities.
- They can predict the next word in a sentence, generate coherent text, and perform a variety of language-related tasks.

Capabilities of (ALL)

1. Text generation
2. Understanding and summarization
3. Translation
4. Question answering
   Conversational agents
5. Writing assistance
6. Brainstorming and creative tasks

A Parallel Comparison Between Human Brain and AI Agents

HUMAN BRAIN

AI AGENTS

	HUMAN BRAIN	AI AGENTS
ENERGY EFFICIENCY	Extraordinarily energy-efficient	High energy consumption
CONSCIOUSNESS	Subjective experiences, self-awareness	No consciousness or emotions
LEARNING	Lifelong, experiential learning	Large-scale offline training Limited adaptability
CREATIVITY	Shaped by emotions, experiences	Statistical recombination

Limitations of T.o.M AI

1- Not true understanding mental state

2- **Ambiguity and uncertainty handling**

3- Cultural and individual variance

4-Context dependence and drift

5- Safety and manipulation risks

6- Generalization vs. overfitting

7- **Ethical and governance considerations: like autonomy, consent, potential harm and …**

8- Bias in emotion recognition datasets

Nice to know:

It seems that a large language AI is brilliant when it is just a good imitator.

People are prompting and programming their AI to sound and write like them, then imitating what the AI writes. They turn themselves into a imitation of an imitation of themselves.


Thank
You