Special Interest Group - Theory of Mind

Progress Report (2024), Future Plans (2025)

Max van Duijn (Leiden University), Rineke Verbrugge (University of Groningen), Ramira van der Meulen (Leiden University)

Who are SIG - ToM?

Theory of Mind (ToM; a.k.a. 'perspective taking', 'user/mental models') is the ability to attribute mental states to others, in order to make sense of their behaviour and perspectives on the world. It is one of the building blocks of an intelligent system: One needs to be able to consider others' perspectives to predict how they will respond to one's actions, which is beneficial both in cooperation and in competition. As a SIG we organise discussion meetings and provide an active platform to discuss what role ToM plays/could play in HI projects, with the aim of advancing collaboration and communication between agents (be they human, or AI) within the consortium.

What have we accomplished so far?

Our main goal is to connect the field to the concept of ToM. We have accomplish this by:

- Organising reading groups that highlight interesting papers in the ToM - AI intersection. This year, discussions were on:
  - Modern multi-modal ToM processing / Q&A
  - Computational models of (human) ToM evolution
  - Human perspectives on RL learning (combo with SIG-RL)
- Advising HI matrix elements who need advice on ToM.
- Establishing a common ground about ToM in AI. The leads wrote a position paper on the role ToM can play in AI, through the misconceptions the field has on ToM, which was presented at an AAAI workshop on ToM.

Intentions for the Future

In 2025, we aim to:

- Continue reading group meetings, with discussions on:
  - Summarised usage of ToM in HI (kick-starter for new projects)
  - Common Ground and ToM in developmental psychology
  - The position paper: Misconceptions in ToM for AI and CS
- Continue advisory role: still available as the matrix element that advises on (the use of) ToM in use cases, challenges, the field
- Investigate the discrepancy between the HI-wide interest in SIG-ToM and the SIG-ToM attendence rates through surveys

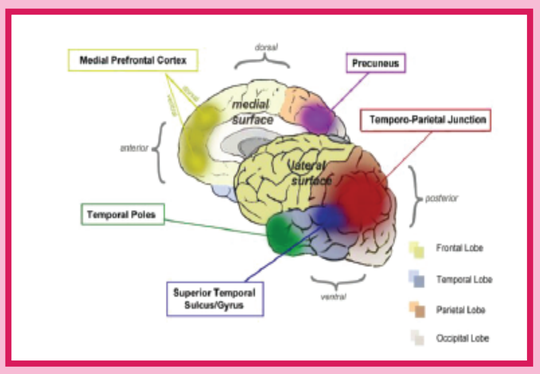

Figure 1a: There is (relative) consensus over regions showing high activity when doing ToM-branded tasks, most notably the temporo-parietal junction (TPJ) and medial prefrontal cortex (MPFC). Image by Bowman & Wellman (2014).
Figure 1b: Theory of Mind can be applied in many situations, but the context is generally social.

Position Paper: Four ToM Misconceptions in AI

ToM is a concept rooted in different research traditions across the cognitive and developmental sciences. We observe that researchers from AI & CS, ourselves included, often have difficulties finding their way in the ToM literature when working on interactive systems. We identify four hyperbolised misconceptions we believe one should account for when implementing ToM in an AI system.

1: "Humans and AI both use a ToM Module"

AI studies sometimes treat ToM as a single 'brain' module, but it is unclear if ToM is reducible to such a modular concept in humans, where ToM is often used to denote an interplay of socio-behavioural skills: social and spatial deduction, linguistic reasoning, knowledge retrieval, etc. This is supported in the neurosciences: Studies differ in underlying mechanisms they identify for ToM, and the neuro-psychological correlates of ToM are often task- dependent, activating different brain segments for different ToM tasks.

>> We cannot flatten perspective-taking to one type of process in hybrid interactions: human-ToM serves to take the AI perspective and AI-ToM to take the human perspective.

2: "Every Social Interaction needs (Advanced) ToM"

We address two issues: (1) When to use ToM to begin with, and (2) how advanced this ToM has to be. For example: Saying 'thanks, you too' after a waiter says 'enjoy your meal' is not ToM failing; it is because ToM usage is not the default in routine conversation. Research into ToM benefits show it to be less effective in predictable environments, removing a motivator for using it situations one has seen many times before. This is also shown in collaborative and competitive games associated with ToM, as they are solvable using learned experiences and probability. Examples include chess, Go, negotiation simulators (e.g. Coloured Trails), and cooperative card games (e.g. 'The Game').

>> AI agents need not always rely on ToM to make interactions work

3: "All ToM is the Same"

It is important to realise that every entity in an interaction is different. This is already the case in human-human interaction, especially if we account for cultural background and personality, but this difference is even more pronounced in human-AI interaction. We quite easily anthropomorphise such a system, even though it experiences the world from a machine perspective, and thus attributing human-like beliefs and perceptions to such a system is often in error. Two possible solutions:

- Humans should invest in building up an 'AI ToM'. Developers need to explain how an AI perceives the world, and how it maps its beliefs and intentions to behaviour.
- Alternatively, the system needs to be designed in a human-like way, so that humans' tendency to anthropomorphise can be properly leveraged and applied.

>> Instead of seeking a single, all-encompassing ToM module to implement, AI & CS researchers should allow ToM to emerge from an interplay between different functions

4: "Current AI Systems Already have ToM"

Computational models of ToM have a long tradition in agent-based modelling. Lively debate emerged around the question whether some AI models possess ToM, instead of simulating ToM-like reasoning processes. LLMs invoked bold claims about spontaneous emergence of ToM. Benchmarks were introduced, modalities and model architectures were integrated, theoretical reflection was added, and more. We highlight three aspects:

- While validity of ToM tests is also an issue in humans, for LLMs it is even more uncertain how performance translates to to real-world social settings.
- Most ToM tests use an 'observer' perspective, where LLMs benefit from a training set with ample descriptions of social life. Answering questions about social situations differs greatly from actually engaging socially from a 1st-person perspective.
- Model characteristics (size, training data, fine-tuning), test type, and test approach, greatly influence performance, raising robustness and flexibility concerns.

>> Without undervaluing ToM-achievements by LLMs, a clear view of the limitations is key: LLMs do not yet exhibit a form of ToM that is ready to support human-AI interaction in a robust and flexible manner.

**See the preprint here –
full version soon to be submitted >>**

Consortium Meeting 1-2 July, 2025

Hybrid Intelligence