Computational Theory of Mind for Human-Agent Collaboration

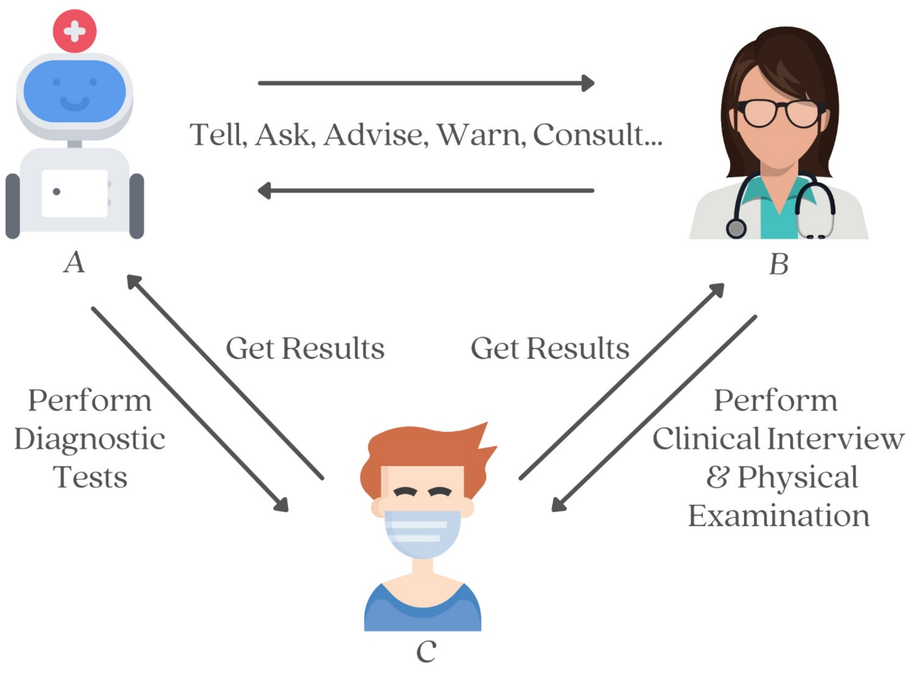
Emre Erdogan[1] Frank Dignum[1,2] Rineke Verbrugge[3] Pınar Yolum[1]

[1]Utrecht University, Utrecht, Netherlands {e.erdogan1,p.yolum}@uu.nl
[2]Umeå University, Umeå, Sweden dignum@cs.umu.se
[3]University of Groningen, Groningen, Netherlands l.c.verbrugge@rug.nl

Computational Theory of Mind

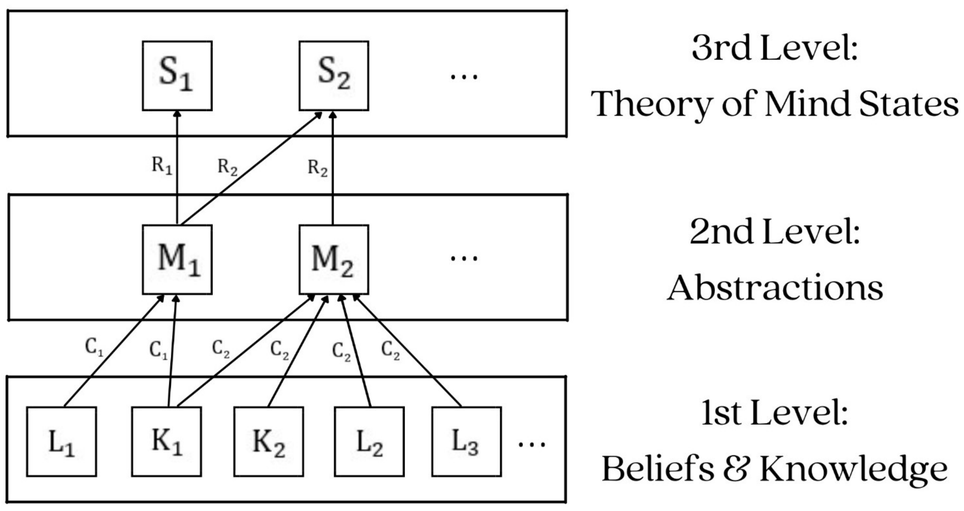
- Theory of Mind (ToM) is the ability of reasoning about the mental content of other people, such as their beliefs, desires, and goals.
- ToM reasoning helps people to understand and predict others' behaviour and is crucial to adapting to the complex dynamics of social life.
- An important area in which ToM can be useful is hybrid intelligence, where an agent can collaborate with a human towards a particular goal.

→ Example: Human-agent collaboration in medicine.



Abstraction and Theory of Mind

- For effective human-agent collaboration, we propose an abstraction framework for computational modeling of ToM reasoning.
- The idea is to employ an agent's beliefs and knowledge to produce more abstract, compact concepts for the agent to use when collaborating with humans.



ToM agents should:

- Take into account complex human notions such as emotions, feelings, and traits.
- Be capable of abstracting social frames of reference such as social roles, norms, and human values.
- Represent and reason on these notions to produce effective, explainable, and fair interactions with humans.

Human-Agent Collaboration in Medicine with Abstraction and Theory of Mind

- We use epistemic logic for formalization purposes: K for knowledge and L belief.
- With examples from the medical domain, we show how abstraction can be useful for an agent doctor when collaborating with a human doctor.

→ **Capturing role-induced abstractions:** An abstraction can be directly induced by a role.

$$\frac{\begin{array}{l}K_A(Doctor(B)) \\ K_A(MedicalCollaboration(A,B)) \\ K_A(Doctor(B) \wedge MedicalCollaboration(A,B) \rightarrow Respect(A,B))\end{array}}{\therefore K_A(Respect(A,B))}$$

→ **Capturing different perspectives:** The role of a role in abstraction can differ for others.

$$\frac{\begin{array}{l}L_A K_B(Doctor(A)) \\ L_A K_B(GoodCapabilities(A)) \\ L_A(K_B(Doctor(A)) \wedge K_B(GoodCapabilities(A)) \rightarrow Respect(B,A))\end{array}}{\therefore L_A(Respect(B,A))}$$

→ **Revising abstractions:** Detecting and reacting to lack of an abstraction can be crucial.

$$\frac{\begin{array}{l}L_A \neg K_B(Capabilites(A)) \rightarrow \textbf{Demonstrate}(B, Capabilites(A)) \\ L_A(\neg Respect(B,A)) \\ L_A(\neg Respect(B,A) \rightarrow \neg K_B(Capabilites(A)))\end{array}}{\begin{array}{l}\therefore L_A \neg K_B(Capabilites(A)) \\ \textbf{Demonstrate}(B, Capabilites(A))\end{array}}$$

→ **Monitoring inconsistencies:** Agents can use abstraction with computational ToM reasoning to monitor inconsistencies in beliefs.

$$\frac{\begin{array}{l}L_A L_B(Trust(C,B)) \wedge L_A(\neg Trust(C,B)) \rightarrow \textbf{Warn}(B, \neg Trust(C,B)) \\ L_A(Lie(C,B)) \\ L_A(Lie(C,B) \rightarrow \neg Trust(C,B)) \\ L_A L_B(Trust(C,B))\end{array}}{\begin{array}{l}\therefore L_A(\neg Trust(C,B)) \\ \textbf{Warn}(B, \neg Trust(C,B))\end{array}}$$

References

[1] Z. Akata, D. Balliet, M. De Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(08):18–28, 2020.

[2] E. Erdogan, F. Dignum, R. Verbrugge, and P. Yolum. Abstracting minds: Computational theory of mind for human-agent collaboration. In *HHAI2022: Augmenting Human Intellect*, pages 199–211. IOS Press, 2022.

[3] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[4] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157): 1124–1131, 1974.