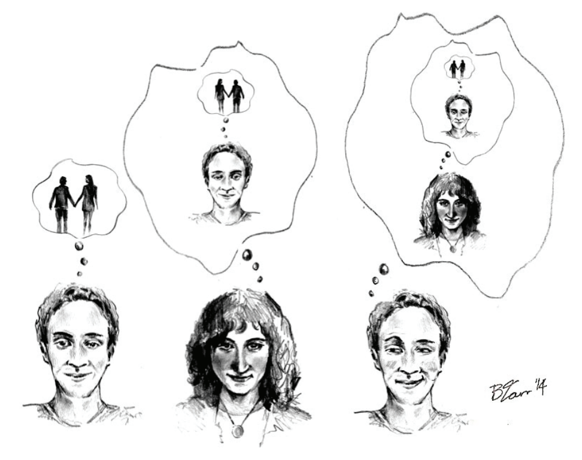
Modelling Common Ground for Theory of Mind

Project 2.13



If two agents collaborate on a task, they will need a shared understanding of the goal they are working towards. Implicitly or explicitly, they will negotiate about the steps that need to be taken on the way. This warrants a constant need for Theory of Mind (**ToM**), i.e., the ability to empathize with others and understand the situation from their perspective. Research on ToM modeling has focused on the "reasoning component". **Common ground** (or shared intentionality) is a necessary input of ToM reasoning, but has not yet been targeted in computational and hybrid modeling.

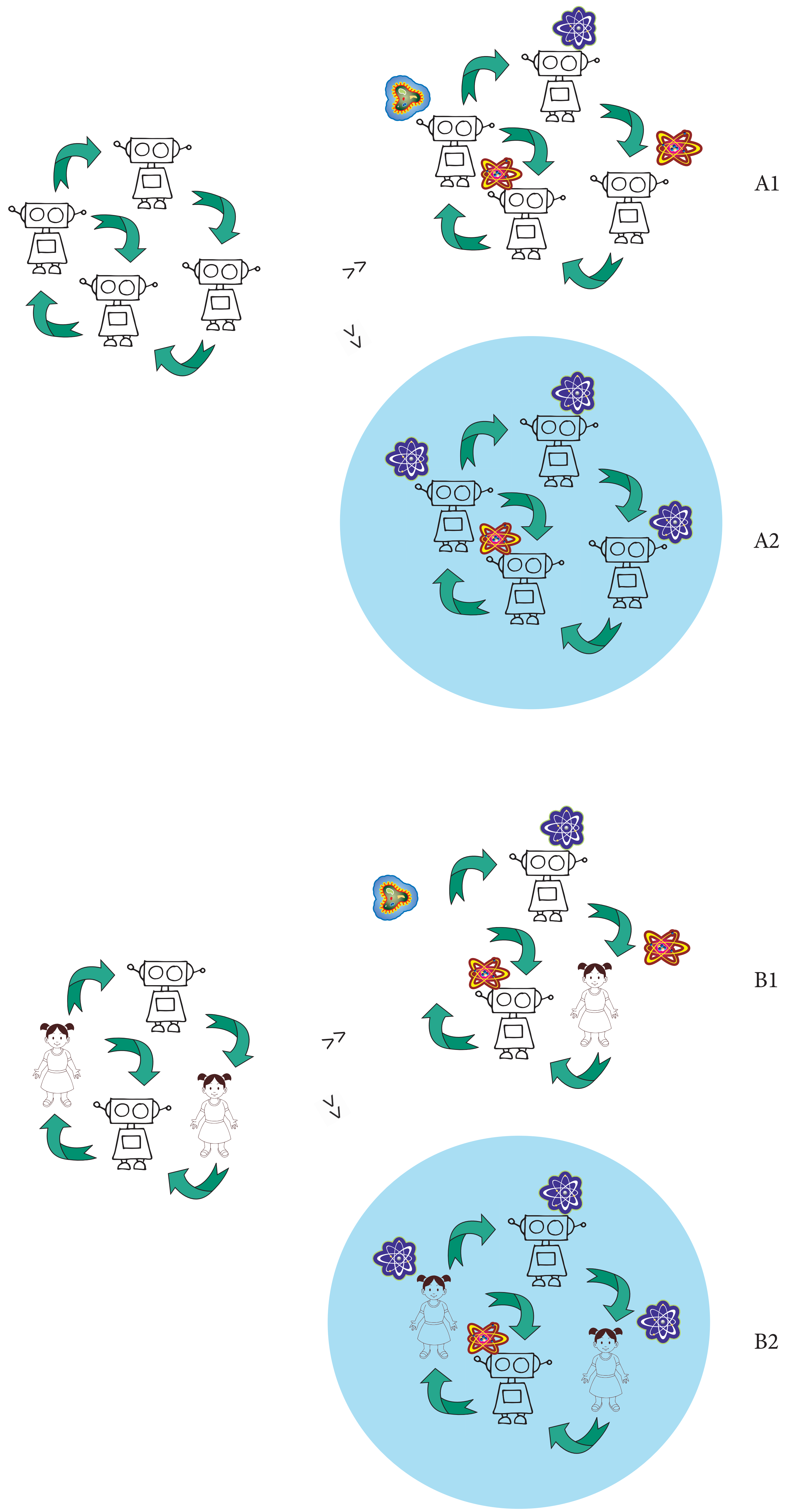
We use agent-based modeling and lab experiments in which humans and AI agents play the same game. The aim is to study when and to what degree common ground emerges, and how this is affected by factors such as memory, analogical reasoning, and communication using a shared code. Understanding this informs the creation of better ToM models for **agent-human collaboration**.

Start date	**Candidate**	**Supervisors**	**Location & Embedding**
1 October 2023	Ramira van der Meulen	Max van Duijn Rineke Verbrugge	Creative Intelligence Lab, LIACS, Leiden (~80%) Multi-agent Systems group, ALICE, Groningen (~20%)

In a first stage (A1), a diversity of agents perform various sequences of observable actions in a game that is performed with two agents at a time; one could think of adaptations of games such as the Tacit Communication Game or Colored Trails [1, 2], or a novel implementation inspired by **The Game**. Actions can be understood and predicted in terms of underlying "intentions". Agents build up more knowledge about action-intention patterns in other agents over the course of their interaction history. In addition, they partly adjust their own action-intention patterns towards those of others they are exposed to, meaning that the variation in patterns used within the population decreases.

In a second stage (A2), agents can also **communicate**. It is still possible to learn action-intention patterns directly from another agent's behaviour; however, in addition, these patterns are now also being made public through limited communication, so that a "communal knowledge pool" evolves (blue circle). It is expected that this improves performance on the collaborative task/game in at least two ways. Firstly, learning efficiency increases because agents do not need to encounter all action-intention patterns first-hand. Secondly, because the action-intention patterns are public and agents adjust their individual intention-action patterns towards those they are exposed to, the patterns will become shared between agents at a faster rate.

Next, **human participants** will interact with the agents on the same task/game via an interface (B1 & B2). The aim is to study whether humans and agents converge on a set of action-intention patterns that is similar or different from agent-only versions, how successful they are in predicting one another's behaviour, and how this is affected by adding possibilities for communicative interaction between humans and agents.



Bruce Eric Kaplan in *The New Yorker*

"Of course I care about how you imagined I thought you perceived I wanted you to feel."

Universiteit Leiden

rijksuniversiteit groningen

liacs

Creative Intelligence Lab

Hybrid Intelligence