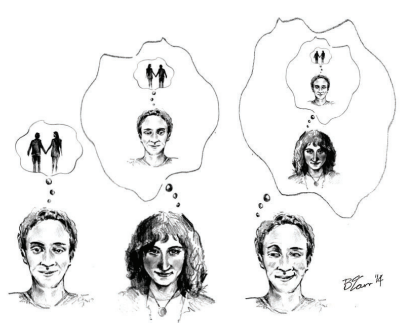
Modelling Common Ground for Theory of Mind

Ramira van der Meulen - Leiden University

Project 2.13

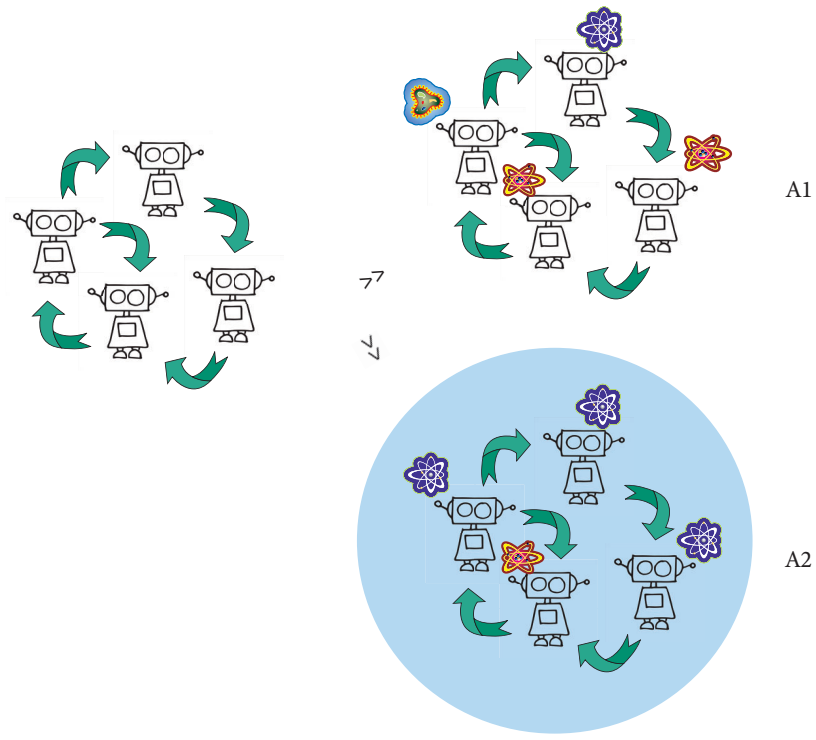


If two agents collaborate on a task, they will need a shared understanding of the goal they are working towards. Implicitly or explicitly, they will negotiate about the steps that need to be taken on the way. In many cases, this warrants a need for Theory of Mind (**ToM**), i.e., the ability take someone else's perspective and make estimations of their beliefs, desires and intentions, in order to make sense of their behaviour and attitudes towards the world. Research on human-agent collaboration often focuses on this ToM - despite its mental costliness. Our project focuses on researching the effectiveness of alternative and combined strategies, such as the use of **Common Ground**.

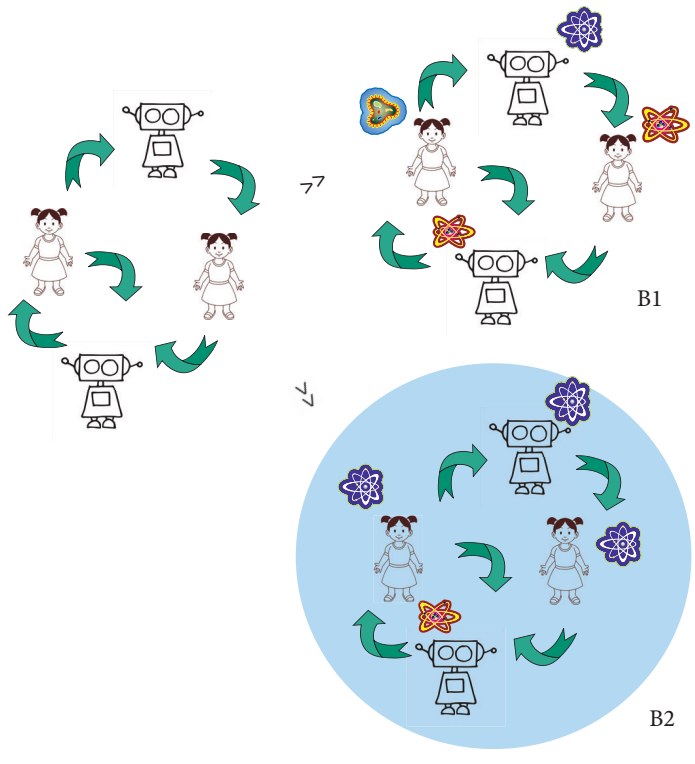
We use agent-based modeling and lab experiments in which humans and AI agents play the same game, traditionally through the use of ToM. The aim is to study when and to what degree mechanisms such as common ground emerge, and how communication affects their formation. Understanding this informs the creation of better mental models for **agent-human collaboration**.

In a first stage (A1), agents perform various sequences of observable actions in a game that is performed with two agents at a time, using an adaption of **The Game**. Through different models of social intelligence, we explore the nature of 'shortcut strategies' in a non-verbal collaborative setting. Agents build up more knowledge about action-intention patterns in other agents during their interactions. 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 receive the same task, but they are also allowed to **communicate verbally**. 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, to evolve a "communal knowledge pool" (blue circle). We expect 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.



In a final stage, **human participants** will interact with the agents on the same task/game via an interface (B1 & B2). We aim 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