Non-recursive Approach for Mutual Understanding

Alexis Jacq[1,2], Wafa Johal[1], Pierre Dillenbourg[1]
[1]CHILI Lab, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2]INESC-ID & Instituto Superior Técnico, University of Lisbon, Portugal

Mutual understanding requires a theory of mind

A social robot is required to interact with humans. The quality of this interaction depends on its ability to behave in an acceptable and understandable manner by the user. Hence it is important for the robot to take care of its image: how much it is perceived as an automatic and repetitive agent, or contrarily as a surprising and intelligent character. If the robot is able to detect this perception of itself, it can adapt its behaviour in order to be understood: "*you think I am sad while I am happy, I want you to understand that I am happy*".

As humans, we have different strategies to exhibit understanding or to resolve a misunderstanding. As an example, if someone is talking about a visual object, we alternatively gaze between the object and the person to make sure he saw that we gazed at the object. Or if we detect that the other person has not understood a gesture (e.g. pointing at an object) we would probably exaggerate the gesture.

Developed by Baron-Cohen and Leslie [1], the Theory of Mind (ToM) describes the ability to attribute mental states and knowledge to others. In interaction, humans are permanently collecting and analysing huge quantity of information to stay aware of emotions, goals and understandings of their fellows. In this work, we focus on a generalization of this notion: **Mutual Modelling** (MM) is a the reciprocal ability to establish a mental model of the other [2].

However, HRI research has not, until now, explored the whole potential of mutual modelling. In [3], Scassellati supported the importance of Leslie's and Baron-Cohen's theory of mind to be implemented as an ability for robots. He focused his work on attention and perceptual processes (face detection or colour saliency detection). Thereafter, some works (including Breazeal [4], Trafton [5], Ros [6] and Lemaignan [7]) were conduced to implement Flavell's first level of perspective taking [8] ("*I see (you do not see the book)*"), ability that is still limited to visual perception.

Architecture enabling 2nd order of MM

A first intuition for MM is to assume that all agents have the same reasoning: given similar inputs they have similar behaviour. In [4], Breazeal presents a MM-based architecture where the robot takes the visual perspective of an human and uses its own reasoning to predict his behaviour. We can imagine higher orders of modelling where the robot recursively attribute to other agents the mutual modelling ability. We do not want to create an infinite recursive loop: the agent then models the robot that models the agent etc. Recursion must be stopped at a given depth. Such an architecture has limits: it is difficult to process in parallel the behaviour of the robot and the behaviour of the other agents, it becomes heavy in computation beyond second order of modelling, and different agents (the robot, the human) or perception of agents (the robot perceived by the human) may have different reasoning and may adopt different behaviours facing similar percepts.

We propose a different approach of modelling, where we define two orders of how agents are perceived: the **first-order agents** describe how the robot perceives agents (the human or the robot itself), while the **second-order agents** describe how the robot perceives the [*agents perceived by agents*] (the robot perceived by the human or another human perceived by the human). We could as well define $n^{\mathbf{th}}$-orders agents for higher levels of theory of mind. But taking into account such high levels would be difficult to process in real time and if a second order is prone to improve interactions (because it enables mutual understanding), it is not sure that higher levels will bring strong improvements.

With our approach, the cognitive architecture of the robot is not recursive: it attributes to each first-order and second-order agent its own separated reasoning. In other words, the robot has **one model of reasoning for itself**, **one for the human** and **one for itself-perceived-by-the-human**. **None of these models are performing mutual modelling**.

We explain how such an architecture would enable detection and repairing of different type of misunderstanding (when the robot does not understand the child, when it does not understand how the child is perceiving it, and when the child does not understand the robot). We believe that it will smooth and improve the quality oh human-robot interactions.

References

[1] S. Baron-Cohen, A. Leslie, and U. Frith, "Does the autistic child have a "theory of mind" ?" *Cognition*, 1985.
[2] S. Lemaignan and P. Dillenbourg, "Mutual modelling in robotics: Inspirations for the next steps," in *HRI*, 2015.
[3] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, 2002.
[4] C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. Thomaz, "Using perspective taking to learn from ambiguous demonstrations," *Robotics and Autonomous Systems*, 2006.
[5] J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *Systems, Man and Cybernetics*, 2005.
[6] R. Ros, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Solving ambiguities with perspective taking," in *HRI*, 2010.
[7] S. Lemaignan, "Grounding the interaction: Knowledge management for interactive robots," Ph.D. dissertation, CNRS - LASA, Technische Universitt Mnchen - Intelligent Autonomous Systems lab, 2012.
[8] J. H. Flavell, "The development of knowledge about visual perception." *Nebraska Symposium on Motivation*, 1977.
[9] H. H. Clark and S. E. Brennan, "Grounding in communication," *Perspectives on socially shared cognition*, 1991.