Cognitive Architecture for Mutual Modelling

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

***Abstract*—In a collaborative learning scenario, learner(s) and eventual teacher(s) must be able to understand each other. For instance, learners share their knowledge and they must adapt their behaviour in order to make sure they are understood by others. This ability requires a model of their own mental states perceived by others. In this paper, we discuss of the importance of a cognitive architecture enabling second-order Mutual Modelling for Human-Robot Interaction in educative contexts.**

I. Introduction

Using robots in educative scenario is an important challenge of HRI, especially when the role of the robot can not be easily taken by humans. Such a scenario has been developed within the CoWriter Project [1] [2] the activity introduces a new approach to help children with difficulties to learn handwriting. Based on *learning by teaching* paradigm, it aims to repair self-confidence and motivation of the child rather than his handwriting performance alone.

Learning by teaching is a technique that engages students to conduct the activity in the role of the teachers in order to support their learning process. This paradigm is known to produce motivational, meta-cognitive and educational benefits in a range of disciplines [3]. The CoWriter project is the first application of learning by teaching approach to handwriting.

The effectiveness of this learning by teaching activity is built on the "protégé effect": the teacher feels responsible for his student, commits to the student's success and possibly experiences student's failure as his own failure to teach. The main idea is to promote in the child an extrinsic motivation to write letters (he will do it in order to help his "protégé" robot) and to reinforce the self-esteem of the child (he plays the teacher and the robot actually progress).

In that context, we need the robot to be able to pretend enough difficulties to motivate the child to help it. This pretending strongly depends on the perception of the robot by the child: the micro-behaviours (gestures, gazes and sounds), initial level and learning speed of the robot must match with what the child imagines of a "robot in difficulty".

Developed by Baron-Cohen and Leslie [4], 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 characterizes the effort of one agent to model the mental state of another one [5].

Until now, the work conduced by the Human-Robot Interaction (HRI) community to develop mutual modelling abilities in robots was limited to a first level of modelling. Higher levels require the ability to recursively attribute a theory of mind to other agent (*I think that you think that* ...) and their application to HRI remains unexplored.

We believe that, at least in a context of educational HRI, we must promote first and second-order mutual modelling abilities in cognitive architectures. In the following sections, we first present the different frameworks that have been developed for Mutual Modelling and their previous adaptations in HRI research. Then we describe in the possible ways to infer both first and second-order models in real time and to make decisions based on this information.

II. Related works

In several contexts, a large amount of fields have introduced their own framework to describe mutual modelling ability [6]. In Developmental psychology, Flavell [7] denotes two different levels of perspective taking: The *cognitive connection* (I see, I hear, I want, I like...) and *mental representation* (what other agents feel, hear, want...).

In Psycholinguistics and collaborative learning, and more precisely in *computer supported collaborative learning* (CSCL), Roschelle and Teasley [8] suggested that collaborative learning requires a process of constructing and maintaining a *shared understanding* of the task at hand. The term "mutual modelling" was introduced in this context, and focused on knowledge states of agents [5]. Dillenbourg developed in [9] a computational framework to represent mutual modelling situations.

Epistemic logic describes knowledges and beliefs shared by agents. This framework enables consideration of infinite-level of mutual modelling. It defines a *shared-knowledge* (all the agents of a group know **X**) and a *common-knowledge* (all the agents of a group know **X**, and know that all the agent know **X**, and know that all the agents know that all the agents know **X**..., etc.) [10].

However, HRI research has not, until now, explored the whole potential of mutual modelling. In [11], Scassellati supported the importance of Leslie's and Baron-Cohen's theory of mind to be implemented as an ability for robots. He limited his work to perceptual processes (face detection or colour saliency detection). Thereafter, some work (including Breazeal [12], Trafton [13], Ros [14] and Lemaignan [15]) has been conduced to implement Flavell's first level of perspective taking [16] ("*I see (you do not see the book)*"), ability that is still limited to visual perception. In [17], Lemaignan implemented a system that compute in real-time the visual field of agents and estimate which objects are looked at. This time, the robot is not just aware of what *can be seen* by agents, but it perceives what *is currently looked at*. Lemaignan used this system to measure Sharma's *with-me-ness* [18], visual commitment based on expected focus of attention in an activity.

III. Next steps for MM-based cognitive architecture

A. Representation of models

A first intuition for mutual modelling is to assume that all agents have the same cognitive abilities. This assumption leads to recursive structures: the robot is projecting its own cognitive abilities in other agents to build their models. Following this direction, a second level of mutual modelling would be encoded by a second step of recursion: the robot also projects its faculty to make such a projection. But high computational capacities are required and modelling operating structures is not sufficient: the contents of models (physical and mental states) must be memorized, as well as contents of second-order models.

We propose a different approach, where we define two order of agents: the *first-order-agents* concerns direct representations of agents by the robot (for example the child), while the *second-order-agents* concerns the representation of agents by agents (for example the robot-perceived-by-the-child). To model second-order-agent like the robot-perceived-by-the-child is crucial if we want to play with the perception of the robot, e.g. to make sure the child understand that the robot is learning from his demonstrations. We can as well define n^{th}*-order* agents to play with higher level of theory of mind. But taking into account high levels of mutual modelling would be difficult to process on-live. Unlike the epistemic logic, this framework will not take into account infinite regress [19] of mutual modelling.

The sensitive devices (for example cameras, micros, robot's sensors ... and in the case of CoWriter the tablet's inputs) are used to measure a quantity of *perceived variables* (the position in space of agents, the gaze targets, quality of demonstrations, the time to respond, the evaluation of the robot by the child...). Each *perceived variable* is associated with the model of one agent (in CoWriter, the positions in space of the child is associated with the model of the child, while the evaluation of the robot is associated with the model of the robot-perceived-by-the-child).

Some variables can't be directly measured by the sensitive devices and are deduced from other variables of models (see an example in the section [expected causality]). We call them *abstract variables*.

A model of an agent is the set of all the values of the perceived or abstract variables associated to this model. Since the values of variables are likely to change with the time, the models are dynamic.

B. Reasoning

In order to deduce the values of abstract variables (that can't be obtained from direct perception), We can build a Bayesian model based on the knowledge from perceived variables. The choice of a probabilist approach instead of a symbolic approach comes from the weakness in the perception of the robot: knowledge and other mental states of agents can not be directly perceived from the observation of behaviours. They must be inferred, hence a probabilistic model enables richer predictions.

This Bayesian network would contain the probabilities that abstract variables take values given the values of perceived variables.

For example, if the robot points an object and detects that a child saw the movement, then we expect that the child will gaze the target object. In other terms, If the child looks at the hand of the robot but does not look at the target object in the following instants, the probability that the child understood the pointing movement is expected to be small. Then, the robot and can make the decision to exaggerate its movement of pointing.

C. Global description of the architecture

Our cognitive architecture for mutual modelling can be designed in three main parts. A **perception part** will regroup all the modules that measure the values of the perceived variables using sensitive devices (camera, sensors, micros...). An example of module of the sensitive part could be the system developed by Lemaignan [17] that uses camera to measure a value of *with-me-ness*. These values are sent to the **mutual modelling part** that updates in real-time models with measured values of perceived variables and deducts the value of abstract variables. Finally, the **decision part** contains all the modules associated to the control of the robot (and other active devices like tablets in CoWriter). These modules can read values given by mutual models in order to compute decisions. In the example of the CoWriter activity, these modules are given by the system that learns and generate letters, but we can add a module that generates micro-behaviour, another that decides to switch to a new activity (e.g. drawing with the robot),... etc. The following subsections explain in detail the content and operation of the three parts of the architecture.

D. Perception modules

The sensitive modules contain all the module able to use sensitive devices to measure value of relevant perceived variables. Some of these modules are associated with the content of the interaction activity. In CoWriter, the learning module takes inputs of tablets to compute the new state of robot's writing. It defines a sensitive module, and the value of the new state of the robot to write a letter defines a perceived variable. As well, the evaluation of the robot by the child via the feedback buttons defines another perceived variable provided by the modules of the activity. Other modules are independent of the activity: the system that estimates the target objects looked by the child provides additive information not directly used by the modules of the activity.

E. Mutual modelling modules

Each perceived value measured by sensitive modules are associated with the model of an agent (or a n^{th}-order agent). Each mutual model can be designed as a module that knows the list of its associated perceived variables and watch if the value of one of these variables has been changed. An additive module knows all the expected causalities and computes the values of abstract variables. Some expected causalities can be empirically learned and other set by hands.

F. Decision making

We believe that the values of variables provided by mutual modelling will provide rich and useful information for decision making. Taking in account these values to elaborate decision should improve the realism and the efficiency of the robot in the interaction. Just like the sensitive ones, the modules that take decision can be directly associated to the activity (in CoWriter, the choice of a new learning curve or the decision to suddenly make a big mistake), or can govern independent behaviour (for example a module that generate the micro-behaviour of the robot).

Some decisions, especially the one that will govern the micro-behaviours of the robot can be autonomously made by the robot. It will not strongly affect the content and objectives of the activity (in CoWriter the main objective is to provide the child with a new extrinsic motivation to write in helping the robot). But other decisions can have a high impact on the progress of the interaction: to stop an activity and to switch to a new one can frustrate a child that was committed into the activity. The conditions to make such a decision are not directly assessable, but must be learned by the robot. In order to make these decision cautiously, we propose to start by a Wizard-of-Oz approach and to move towards an autonomous approach following these steps:

1) **Wizard-of-Oz**: A human takes decisions; the robot learns
2) **Mixed-initiative**: The robot makes suggestions; a human agrees or disagrees
3) **Autonomous**: The robot makes decisions

The picture 1 visually summarize the global design of our architecture.

IV. CONCLUSION

Educational HRI based on learning by teaching approach needs robot able to perform second-level mutual modelling. We introduced a new approach to implement mutual modelling into a cognitive architecture. We used the CoWriter activity as an example of application, but our architecture could be easily generalised for any kind of interaction. We believe that this step must be reached in other context of HRI, in order to develop higher realism of behaviours and to improve the quality of interactions.

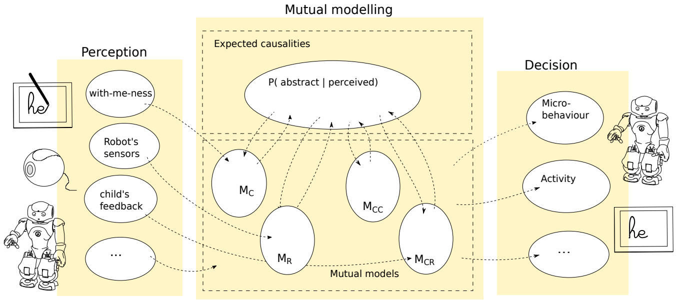

Fig. 1. **Overview of the cognitive architecture**. Yellow squares represent different main parts. White ellipses represent modules. We named in example some content of modules and illustrated possible devices used for perception and decisions.

REFERENCES

[1] D. Hood, S. Lemaignan, and P. Dillenbourg, "When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: ACM, 2015, pp. 83–90.

[2] A. Jacq, S. Lemaignan, F. Garcia, P. Dillenbourg, and A. Paiva, "Building successful long child-robot interactions in a learning context," in *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference*, 2016.

[3] C. A. Rohrbeck, M. D. Ginsburg-Block, J. W. Fantuzzo, and T. R. Miller, "Peer-assisted learning interventions with elementary school students: A meta-analytic review," *Journal of Educational Psychology*, vol. 95, no. 2, pp. 240–257, 2003. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0663.95.2.240

[4] S. Baron-Cohen, A. Leslie, and U. Frith, "Does the autistic child have a "theory of mind" ?" *Cognition*, 1985.

[5] P. Dillenbourg, "What do you mean by collaborative learning?" *Collaborative-learning: Cognitive and Computational Approaches.*, pp. 1–19, 1999.

[6] S. Lemaignan and P. Dillenbourg, "Mutual modelling in robotics: Inspirations for the next steps," in *Proceedings of the 2015 ACM/IEEE Human-Robot Interaction Conference*, 2015.

[7] J. H. Flavell, F. L. Green, and E. R. Flavell, "Developmental changes in young children's knowledge about the mind," *Cognitive Development*, vol. 5, no. 1, pp. 1–27, 1990.

[8] J. Roschelle and S. D. Teasley, "The construction of shared knowledge in collaborative problem solving," in *Computer supported collaborative learning*. Springer, 1995, pp. 69–97.

[9] M. Sangin, N. Nova, G. Molinari, and P. Dillenbourg, "Partner modeling is mutual," in *Proceedings of the 8th iternational conference on Computer Supported Collaborative Learning*. International Society of the Learning Sciences, 2007, pp. 625–632.

[10] V. Hendricks and J. Symons, "Epistemic logic," in *Stanford Encyclopedia of Philosophy*, 2008.

[11] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.

[12] C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. Thomaz, "Using perspective taking to learn from ambiguous demonstrations," *Robotics and Autonomous Systems*, pp. 385–393, 2006.

[13] J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 35, no. 4, pp. 460–470, 2005.

[14] R. Ros, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Solving ambiguities with perspective taking," in *5th ACM/IEEE International Conference on Human-Robot Interaction*, 2010.

[15] S. Lemaignan, "Grounding the interaction: Knowledge management for interactive robots," Ph.D. dissertation, CNRS - Laboratoire d'Analyse et d'Architecture des Systmes, Technische Universitt Mnchen - Intelligent Autonomous Systems lab, 2012.

[16] J. H. Flavell, “The development of knowledge about visual perception.” *Nebraska Symposium on Motivation*, vol. 25, pp. 43–76, 1977.

[17] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, “From real-time attention assessment to with-me-ness in human-robot interaction,” in *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference*, 2016.

[18] K. Sharma, P. Jermann, and P. Dillenbourg, “with-me-ness: A gaze-measure for students attention in moocs,” in *International conference of the learning sciences*, no. EPFL-CONF-201918, 2014.

[19] H. H. Clark and S. E. Brennan, “Grounding in communication,” *Perspectives on socially shared cognition*, vol. 13, no. 1991, pp. 127–149, 1991.