# Clarifying the MM architecture

Alexis Jacq

## I. Remind of the framework

### A. Justification of the approach

A first intuition for mutual modelling is to assume that all agents have the same basic architecture. In [1], Breazeal show a MM-based reasoning where the robot uses its own architecture to model other agents. We can imagine a second level of modelling where the robot recursively attribute to other agents the mutual modelling ability. But that creates an infinite recursive loop: the agent then models the robot that models the agent etc. Another reason to avoid a recursive approach is that different agent must have different behaviours: in similar situations, they will not necessary take similar decisions.

We propose a different approach of modelling, where we define two orders of agent's model: the *first-order-agents* concern direct representations of agents by the robot (for example the child), while the *second-order-agents* concern 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 will help to model how the child perceives 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 with higher level of theory of mind. But taking into account high levels of mutual modelling would be difficult to process in real time. Unlike the epistemic logic, our proposed framework will not take into account infinite regress [2] of mutual modelling.

### B. Notations

$M_R[A,B]$ stands for the model (built by the robot) of the agent B perceived by the agent A. As said above, we limit our approach to 1st and 2nd order of modelling. In a 2-agents interaction (the child and the robot) we will focus on 3 agent's models: $M_R[C]$ (the model of the child), $M_R[R]$ (the model of the robot) and $M_R[C,R]$ (the robot perceived by the child). I would be also interesting to study $M_R[C,C]$, the model of the child perceived by himself in order to play with his self-confidence. But detecting differences between $M_R[C]$ and $M_R[C,C]$ seems difficult with the current abilities of the robot.

As models are dynamic, $M_R^t[A]$ represent the model of an agent A at time $t$.

## II. Mutual understanding

Given those three models ($M_R[R]$, $M_R[C]$ and $M_R[C,R]$) the robot must be able to detect misunderstandings. A misunderstanding of an agent $A$ by the robot can be formalised as a error between what is actually in the mind of the agent (we can call it $\Phi[A]$) and the model built by the robot:

$\Delta(\Phi[A]; M_R[A])$. But if $A$ is human, $\Phi[A]$ is inaccessible to the robot. In order to maintain a mutual understanding, Humans [3] (and monkeys [4]), use prediction of other's behaviours. A bio-inspired approach would be to make, at time $t$, a prediction $P_R^{t+1}[A]$ of the model. Then, at time $t+1$, the robot can compute a prediction error $\Delta\left(M_R^{t+1}[A]; P_R^{t+1}[A]\right)$ in order to detect such a misunderstanding. This idea rely on the assumption that the better are the predictions of a model, the better the model fits the reality. Then, the dynamic of the model can be updated according to the resulting error of prediction. This rule can be used with $M_R[C]$ and $M_R[C,R]$.

Another type of misunderstanding concerns the comprehension of the robot by the child: using the same formalism, it is an error between the actual perception of the robot by the child (we can call it $\Phi[C,R]$) and the robot itself: $\Delta(\Phi[C,R]; M_R[R])$. Again, the robot does not have access to $\Phi[C,R]$, but it approximates it with $M_R[C,R]$. Finally we define the child's perception error at time $t$ as $\Delta(M_R^t[C,R]; M_R^t[R])$.

## III. Structure and dynamic of the models

Each model contains the knowledge and possible actions of an agent. The dynamic between actions and knowledges must be encoded. Some causality can be assumed in advance (if the child gives a good feedback to the robot, it means that the robot-perceived-by-the-child makes progresses) while other can be learnt (e.g. if each time the robot looks at the head of the child the child stops to write). Each modelled agent has its own goals, encoded as a reinforcement learning process. As an example, in CoWriter the robot-perceived-by-the-child ($M_R[C,R]$) is a learner and has the simple goal to make progresses. The child ($M_R[C]$) is modelled as a teacher and is assumed to have the goal to improve the robot's writing. Then, the goal of the robot ($M_R[R]$) is to keep a mutual understanding: its knowledges and actions must correspond to itself perceived by the child, and it must be able to predict the behaviour of the child. The schema on Figure 1 visually summarizes this example.

The goals of $M_R[C]$ and $M_R[C,R]$ depend on the activity and must be programmed each time we move to a new activity. They can be improved to better fit the child's behaviour by learning from experience. Contrariwise, the goal of $M_R[R]$ to keep a mutual understanding is independent and must be automatic.

## IV. Repairing misunderstandings

If an important misunderstanding is detected (according to $M_R(CR)$ the robot does not know $K$ while it actually knows
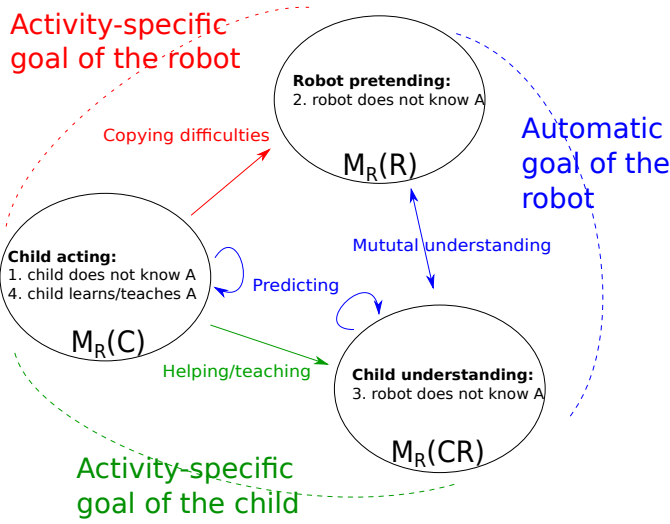
Fig. 1.

it according to $M_R(R)$) the robot can "turn the tide" of the dynamics between models to figure out what action of the robot leads the child to think that the robot does know $K$, and then exaggerate the action.

## REFERENCES

[1] 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.

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

[3] S. Suzuki, R. Adachi, S. Dunne, P. Bossaerts, and J. P. ODoherty, "Neural mechanisms underlying human consensus decision-making," *Neuron*, vol. 86, no. 2, pp. 591–602, 2015.

[4] K. Haroush and Z. M. Williams, "Neuronal prediction of opponents behavior during cooperative social interchange in primates," *Cell*, vol. 160, no. 6, pp. 1233–1245, 2015.