Sensitivity To Perceived Mutual Understanding In Human-Robot Collaborations

Socially Interactive Agents Track

Alexis David Jacq
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
alexis.jacq@gaips.inesc-id.pt

Julien Magnan
Méroé films
Paris, France
julien.magnan@gmail.com

Maria Jose Ferreira
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
maria.jose.ferreira@gaips.inesc-id.pt

Pierre Dillenbourg
Ecole Polytechnique Federale de Lausanne
Lausanne, Switzerland
pierre.dillenbourg@epfl.ch

Ana Paiva
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
ana.paiva@gaips.inesc-id.pt

ABSTRACT

In order to collaborate with humans, robots are often provided with a Theory of Mind (ToM) architecture. Such architectures can be evaluated by humans perception of the robot's adaptations. However, humans sensitivities to these adaptations are not the one expected. In this paper, we introduce an interaction involving a robot with a human who design, element by element, the content of a short story. A second-order ToM reasoning aims at estimating user's perception of robot's intentions. We describe and compare three behaviors that rule the robot's decisions about the content of the story: the robot makes random decisions, the robot makes predictable decisions, and the robot makes adversarial decisions. The random condition involves no ToM, while the two others are involving 2nd-order ToM. We evaluate the ToM model with the ability to predict human decisions and compare the ability of the human to predict the robot given the different implemented behaviors. We then estimate the appreciation of the robot by the human, the visual attention of the human and his perceived mutual understanding with the robot. We found that our implementation of the adversarial behavior degraded the estimated interaction's quality. We link this observation with the lower perceived mutual understanding caused by the behavior. We also found that in this activity of story co-creation, subjects showed preferences for the random behavior.

Long version: https://alexis-jacq.github.io/papers/sensitivity.pdf

ACM Reference Format:
Alexis David Jacq, Julien Magnan, Maria Jose Ferreira, Pierre Dillenbourg, and Ana Paiva. 2018. Sensitivity To Perceived Mutual Understanding In Human-Robot Collaborations. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

INTRODUCTION

In contrast with virtual agents or any intelligent tool, a role played by a physical humanoid robot is known to promotes anthropomorphism [3]. This effect is often presented as an adventage in Human-Robot Interaction (HRI) community since it may reinforce subjects engagement in activities. A well known example of such a fenomena is called "protégé" effect, where subjects create an attachement as they feel responsible of the robot. This is usually desired in therapeutic and pedagogical contextes [7] [2]. Besides, another challenge of HRI is to design non-autistic robots by implementing ToM architectures [4]. It is accepted that Human-Robot collaboration would be improved by an awarness of both intentions by sharing mental models [6]. Especially in educative perspectives, where researchers in the field of *Computer-Supported Collaborative Learning* (CSCL) explain how a shared understanding helps in collaborative resolutions of problems [5]. The question we want to raise through this study concerns the impact of a ToM implementation on the human sensitivity during a collaborative task with a humanoid robot.

In this paper, we define *mutual understanding* by the ability of agents to predict others and to be predicted by others. We implemented a reasoning model for mutual understanding based on a three-agents architecture: *self; other; self-view-by-other*, introduced in [1]. We used it to implement two robot's behaviors: making predictable decisions or making adversarial decisions. These behaviors are designed within an activity where the robot chooses, turn by turn with a human, elements that construct a short story. Our predictable behavior is built in order to facilitate the mutual understanding, while our adversarial behavior lets the subject believe he understands the robot and suddenly surprises him with the least predictable decision. As a control condition, we also implemented a random behavior, in which the robot only makes random decisions.

We conducted a study involving 47 subjects, not aware of the robot's behavior condition. The experiment's design and results are described and discussed in the long version of this paper.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

STORY CO-CREATION BY SELECTING ELEMENTS

The activity consist in choosing, turn by turn with the robot, a specific element of the story. Such an element can be the place of the story (planet? kingdom? island?) or the job of the protagonist (space pioneer? knight? pirate?). Once all elements have been selected by the subject and the robot, the resulting story is generated, based on the human-robot collaborative selection of contents. Actually, the story is rather “filled" than generated: at the beginning, a sentence has a fixed structure but each word that is – or depends on – a selectable element is replaced by a symbolic variable. For example, our story could start with the two following sentences:

Once upon a time, in a **Place** *far away populated by* **People***, was living a wild* **Main_Char_Job** *named* **Main_Char_Name**.
Personal_Pronoun(Main_Char_Gender) *was very brave.*

In this text, variables are the bold terms. The variable “Place" is a selectable element, that can be replaced by any possible geographical place (planet, kingdom, island, ...). The personal pronoun related to the main character depends on the selectable element “Main_Char_Gender". Some whole sentences can also depend on a variable in order to avoid inconsistencies.

Before each robot’s turn, subjects are asked to predict what will be the robot’s decision. The sequence of successive triples (*subject’s decision*; *subject’s prediction of the robot*; *robot’s decision*) was feeding our two decision making algorithms based on 2nd order ToM.

DECISION MAKING

Contexts

We define a context as a set of selectable elements belonging to a same semantic field. For example, the context *science fiction* contains the elements *planet*, *alien*, *lazer gun*, etc. We arbitrary set 8 contexts: *science fiction*, *pirates*, *middle-ages*, *forest*, *science*, *army*, *robots*, *magic*. Since an element can be associated to several contexts, contexts are not disjoint.

Agent models

As suggested in [1], we define three agents: the robot ($\mathcal{R}$), the human ($\mathcal{H}$), the robot predicted by the human ($\mathcal{P}$). Each agent $\mathcal{A}$ is modeled by a log-probability distribution over contexts, $\mathfrak{L}_{\mathcal{A}}$, estimating the odds that it is going to pick elements from this context. For example, $\mathfrak{L}_{\mathcal{H}}(pirates)$ estimates the probability of the event “the human is going to pick an element in the *pirates* context", while $\mathfrak{L}_{\mathcal{P}}(pirates)$ estimates the probability of the event “the human predicts that the robot is going to pick an element in the *pirates* context". From these distributions, we can define, for each agent $\mathcal{A}$, its most likely context $\mathtt{C}_{\mathcal{A}}^{max} = \mathrm{argmax}_{\mathtt{C}}\, \mathfrak{L}_{\mathcal{A}}(\mathtt{C})$ and its least likely context $\mathtt{C}_{\mathcal{A}}^{min} = \mathrm{argmin}_{\mathtt{C}}\, \mathfrak{L}_{\mathcal{A}}(\mathtt{C})$.

Agent weights

Each agent $\mathcal{A}$ is given a weight $W_{\mathcal{A}}$ representing the human inclination to establish its predictions, rather based on the robot’s decisions ($W_{\mathcal{R}}$), on his own decisions ($W_{\mathcal{H}}$) or on his own predictions of the robot ($W_{\mathcal{P}}$).

Weights updates

At each step of the element-selection activity, we receive a new triple $(\mathtt{e}_{\mathcal{H}}; \mathtt{e}_{\mathcal{P}}; \mathtt{e}_{\mathcal{R}})$ where $\mathtt{e}_{\mathcal{H}}$ is the element picked by the human, $\mathtt{e}_{\mathcal{P}}$ is the human prediction of the element picked by the robot, and $\mathtt{e}_{\mathcal{R}}$ is the element actually picked by the robot. An agent’s weight $W_{\mathcal{A}}$ is incremented if its last picked element $\mathtt{e}_{\mathcal{A}}$ belongs to its most likely context $\mathtt{C}_{\mathcal{A}}^{max}$:

$$W_{\mathcal{A}} \leftarrow W_{\mathcal{A}} + \mathbb{1}\{\mathtt{e}_{\mathcal{A}} \in \mathtt{C}_{\mathcal{A}}^{max}\}\ \forall\ agent\ \mathcal{A}$$

Probabilities updates

Then, agents log-probability distributions $\mathfrak{L}_{\mathcal{H}}$ and $\mathfrak{L}_{\mathcal{R}}$ are both updated in a similar way, for all context $\mathtt{C}$:

$$\mathfrak{L}_{\mathcal{H}}(\mathtt{C}) \leftarrow \mathfrak{L}_{\mathcal{H}}(\mathtt{C}) + \mathbb{1}\{\mathtt{e}_{\mathcal{H}} \in \mathtt{C}\}$$
$$\mathfrak{L}_{\mathcal{R}}(\mathtt{C}) \leftarrow \mathfrak{L}_{\mathcal{R}}(\mathtt{C}) + \mathbb{1}\{\mathtt{e}_{\mathcal{R}} \in \mathtt{C}\}$$

While $\mathfrak{L}_{\mathcal{P}}$ is updated using weights $W_{\mathcal{R}}$, $W_{\mathcal{H}}$ and $W_{\mathcal{P}}$, for all context $\mathtt{C}$:

$$\mathfrak{L}_{\mathcal{P}}(\mathtt{C}) \leftarrow \mathfrak{L}_{\mathcal{P}}(\mathtt{C}) + \sum_{\mathcal{A} \in \{\mathcal{R},\mathcal{H},\mathcal{P}\}} W_{\mathcal{A}} * \mathbb{1}\{\mathtt{e}_{\mathcal{A}} \in \mathtt{C}\}$$

Predictable behavior

Our predictable behavior aims at making decisions that are easily predicted by the subject. In that purpose, the robot always pick elements from $\mathcal{P}$’s most likely context $\mathtt{C}_{\mathcal{P}}^{max}$:

$$\mathtt{e}_{\mathcal{R}} \in \mathtt{C}_{\mathcal{P}}^{max}$$

adversarial behavior

The adversarial behavior is more complex. We use the predictable behavior, waiting for the human to make good predictions (predicting an element $\mathtt{e}_{\mathcal{P}}$ belonging to $\mathtt{C}_{\mathcal{P}}^{max}$). Then, we suddenly move to the opposite: picking $\mathtt{e}_{\mathcal{R}}$ in the least likely context $\mathtt{C}_{\mathcal{P}}^{min}$. However, we wanted to make this behavior the least understandable. Therefore we add, with a low probability, the possibility to pick $\mathtt{e}_{\mathcal{R}}$ from $\mathtt{C}_{\mathcal{P}}^{max}$ while the humanis making a good prediction, or the possibility to pick exactly the element predicted by the subject while the human did not predict an element from $\mathtt{C}_{\mathcal{P}}^{max}$. Algorithm 1 summarizes this behavior.

Algorithm 1: adversarial behavior

```
if e_P ∈ C_P^max then
    with prob. P=0.8, e_R ∈ C_P^min
    with prob. P=0.2, e_R ∈ C_P^max
else
    with prob. P=0.8, e_R ∈ C_P^max
    with prob. P=0.2, e_R = e_P
end
```

ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

REFERENCES

[1] Alexis Jacq, Wafa Johal, Pierre Dillenbourg, and Ana Paiva. 2016. Cognitive architecture for mutual modelling. *arXiv preprint arXiv:1602.06703* (2016).

[2] A. Jacq, S. Lemaignan, F. Garcia, P. Dillenbourg, and A. Paiva. 2016. Building Successful Long Child-Robot Interactions in a Learning Context. In *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference*.

[3] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition* 26, 2 (2008), 169–181.

[4] S. Lemaignan and P. Dillenbourg. 2015. Mutual Modelling in Robotics: Inspirations for the Next Steps. In *Proceedings of the 2015 ACM/IEEE Human-Robot Interaction Conference*.

[5] Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*. Springer, 69–97.

[6] Julie Shah and Cynthia Breazeal. 2010. An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming. *Human factors* 52, 2 (2010), 234–245.

[7] Fumihide Tanaka and Shizuko Matsuzoe. 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction* 1, 1 (2012).