Sensitivity To Perceived Mutual Understanding In Human-Robot Collaborations

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.

1 Introduction

In contrast with virtual agents or any intelligent tool, a role played by a physical humanoid robot is known to promotes anthropomorphism [Kiesler *et al.*, 2008]. 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 [Tanaka and Matsuzoe, 2012] [Jacq *et al.*, 2016a]. Besides, another challenge of HRI is to design non-autistic robots by implementing ToM architectures [Lemaignan and Dillenbourg, 2015]. It is accepted that Human-Robot collaboration would be improved by an awarness of both intentions by sharing mental models [Shah and Breazeal, 2010]. 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 [Roschelle and Teasley, 1995]. 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 [Jacq *et al.*, 2016b]. 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. We found that, while the adversarial and random conditions were associated with a similar low level of measured mutual understanding (compared with the predictable condition), only the adversarial condition led to significantly lower appreciations of the robot. Besides, it seems that subjects perceived higher misunderstanding situations in the adversarial condition, which led us to hypothesize that a conscious low mutual understanding between humans and robots may degrade the appreciation of the robot and consequently, may reverse the benefits of the activity. Hence we invite, with this paper, to stress the design of the human-robot mutual understanding, especially in collaborative contexts.

2 Related work

Introduced by Premack and Woodruff [Premack and Woodruff, 1978] and developed by Baron-Cohen and Leslie [Baron-Cohen *et al.*, 1985], *Theory of Mind* (ToM) describes the ability to attribute mental states and knowledge to others. In interaction, humans are permanently collecting and analyzing huge quantities of information in order to stay aware of emotions, goals and understandings of their fellows. This process enables the maintenance of a common ground of knowledge [Clark and Brennan, 1991], which is essential for collaboration.

Robot architectures enabling first-order models have been developed within the HRI community, which led to solve basic ToM tests [Breazeal *et al.*, 2009][Warnier *et al.*, 2012]. More recent architectures extended such reasoning to plan execution for collaborative tasks [Devin and Alami, 2016]. Regarding mutual modeling, second order of ToM has been stepped by Nikolaidis, solving shared plan execution through visual perspective taking: in [Nikolaidis *et al.*, 2016], the robot is computing the most understandable trajectory in order to share a grabbing intention, rather than the most effective trajectory in terms of time and energy. Our model of reasoning is based on the same idea of playing with the estimated comprehension of the human, but is specialized to context-based story creation while gestural intentions are based on visual and physical computations. Since our activity concerns a sequential decision-making and does not need any visual reasoning, we moved to a simpler ToM approach.

First introduced in CSCL [Dillenbourg, 1999] and then borrowed by Human-Robot Interaction (HRI) community [Lemaignan and Dillenbourg, 2015], *mutual modeling* is a computational framework for ToM where agents are modeling each other's intentions, rather than knowledges and beliefs. As in other approaches, higher orders of mutual modeling are defined to express how humans can recursively attribute a model of ToM to others: in the first order agents only construct models of others without supposing that they may also perform mutual modeling, while in the second order they also infer how others model others, including themselves.

We wanted to place our study in the perspective of a pedagogical context, hence we adopted a mutual modeling approach. We focused on Mutual understanding, which involves a second order of modeling: more than simply understanding the other, an agent must take care of being understood. And trying to be understood requires an agent with the capacity to model itself through the eyes of the other.

3 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.

In order to choose an element, a subject must touch it on a touchable screen. For its part, the robot just vaguely points it with its finger and the element is in parallel selected on the screen. The robot is also provided with a face detector and alternates head movements, gazing at the screen or at the subject. Finally, when the robot performs hand gestures while speaking.

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.

4 Decision making

4.1 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.

4.2 Agent models

As suggested in [Jacq *et al.*, 2016b], 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, $\mathfrak{L}_\mathcal{H}(\textit{pirates})$ estimates the probability of the event "the human is going to pick an element in the *pirates* context", while $\mathfrak{L}_\mathcal{P}(\textit{pirates})$ 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 $\mathtt{C}^{max}_\mathcal{A} = \operatorname{argmax}_\mathtt{C} \mathfrak{L}_\mathcal{A}(\mathtt{C})$ and its least likely context $\mathtt{C}^{min}_\mathcal{A} = \operatorname{argmin}_\mathtt{C} \mathfrak{L}_\mathcal{A}(\mathtt{C})$.

4.3 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 ($W_\mathcal{P}$).

4.4 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 $\mathtt{C}_\mathcal{A}^{max}$:

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

4.5 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 $\mathtt{C}$:

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

4.6 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 $\mathtt{C}_\mathcal{P}^{max}$:

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

4.7 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 $\mathtt{e}_\mathcal{P} \in \mathtt{C}_\mathcal{P}^{max}$ **then**
  with prob. P=0.8, $\mathtt{e}_\mathcal{R} \in \mathtt{C}_\mathcal{P}^{min}$
  with prob. P=0.2, $\mathtt{e}_\mathcal{R} \in \mathtt{C}_\mathcal{P}^{max}$
else
  with prob. P=0.8, $\mathtt{e}_\mathcal{R} \in \mathtt{C}_\mathcal{P}^{max}$
  with prob. P=0.2, $\mathtt{e}_\mathcal{R} = \mathtt{e}_\mathcal{P}$
end

5 Experiment

We conducted an experiment in order to study the impact of the three behaviors of the robot on the interaction. The content of the activity was designed in English language. In order to make sure they had a good understanding of English, we invited undergrad students to be subjects for our experiment. However, this decision may have brought weaknesses regarding our possible results. First, this population is biased by the fact that a part of them have already been implied in a human-robot experiment. Then, this story co-creation activity aims to provide a support for children education, and results in adults population may never be generalized to children.

5.1 Groups

A total of 47 students (18f, 29m) accepted to participate to the study. The experiment was conducted in our laboratory. Subjects were aged between 18 and 34 (M 22.8, SD 3.9). We defined 3 groups in which subjects were randomly allocated: the random-behavior group (9f, 7m), the predictable-behavior group (5f, 11m) and the surprise-behavior group (4f, 11m). We used the random behavior as a control condition.

5.2 Design

Each subject was alone with the robot in the room during the whole interaction and the robot was fully autonomous. The spatial arrangement is detailed in figure 2 (top view) and 3 (camera view). The robot, standing on a support, is facing the human user and between them, a touchable screen is inclined for the subject. Also on the support, at the feet of the robot, a RGB-camera was tracking the user's face. We used face-tracking for attention estimation (see 5.3), but also in order to implement robot's head movements. The questionnaire was displayed on the touchable screen and required to scroll down with a mouse. For that purpose, subjects had a mouse available on the right of the screen. The experiment was designed in 4 phases:

1) Introduction (0.75 min exactly): At the beginning, the screen is empty. The robot introduces itself and the activity. All the speeches of the robot were scripted and can be found in our source code, available on GITHUB.

2)Turn by turn selection of elements (4.7 min on average): To start, the robot asks subjects to choose the first element: the place of the story (planet, forest, kingdom...). The interface appears on the screen, displaying a suggestion of possible elements the subject can choose. Figure 1 shows an example of screen capture of the interface for subject's turn. Then, elements that will be suggested to the robot are shown on the screen and subjects are asked to guess what the robot is going to choose. When a subject has made his prediction, the robot takes its turn and chooses, by pointing a button with its arm, the next element. During this turn, buttons to pick elements do not react to subjects' touch. Finally it is the subject's turn again, etc. In order to better feed user modeling algorithms, at two points the human had two consecutive turns, hence the human made more decisions than the robot (10 turns for the human and 8 turns for the robot).

3) Story-telling (3.6 min exactly): At the end, when all elements have been selected by the human and the robot, the

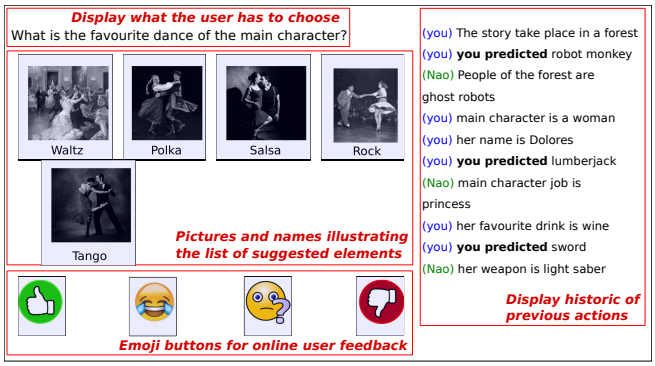

Figure 1: Screen capture of user interface. It contains 4 areas. Top-left: a question reminds what kind of element the user has to choose (for instance, the favorite dance of the main character). Center: the set of suggested elements the user can choose illustrated by pictures. Bottom: 4 emoji buttons the user can use, if he wants to, in order to share his feeling. Right: a column displays the historic of previous action in order to help the user to make prediction about robot's actions.

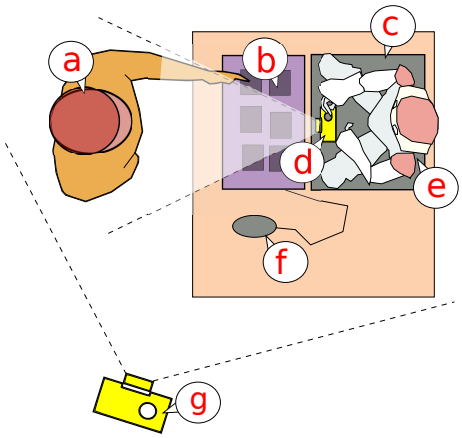

Figure 2: Spacial arrangement, top view: (a) subject, (b) touchable screen, (c) support for the robot, (d) rgb-camera for face-tracking, (e) robot, (f) mouse helping the subject to fill the questionnaire, (g) camera filming the interaction.

resulting story is generated, and the robot tells the story to the human. While the robot tells the story, the screen displays the told sentences. At any time during the whole interaction (including both co-creation and storytelling phases) four emoji buttons were displayed on the screen and could be used by subjects whenever they wanted to share feedback about their feelings. As in [Jacq *et al.*, 2016a] and [Johal *et al.*, 2016] we used thumbs up and down, plus two emoji buttons for "laugh" or "absurd" feeling.

4) Questionnaire (10.3 min on average) Finally, a questionnaire appeared on the screen, asking subject about their appreciation of the activity, their perception of the robot (Godspeed) and their perception of it's ToM abilities.

5.3 Measures

In order to measure our models accuracy, we counted the number of time the human was picking or predicting elements

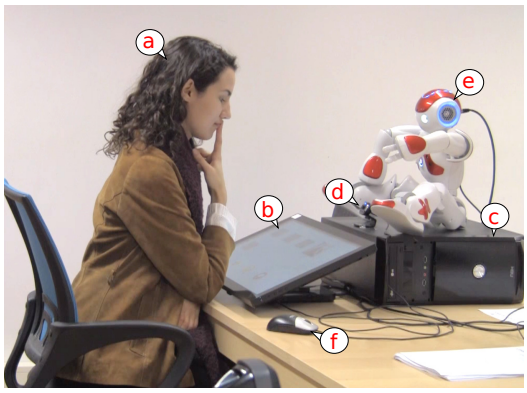

Figure 3: Spatial arrangement, camera view: (a) subject, (b) touchable screen, (c) support for the robot, (d) rgb-camera for face-tracking, (e) robot, (f) mouse helping the subject to fill the questionnaire.

in the expected most likely contexts: the number of time that $\mathtt{e}_{\mathcal{H}} \in \mathtt{C}_{\mathcal{H}}^{max}$ and the number of time that $\mathtt{e}_{\mathcal{P}} \in \mathtt{C}_{\mathcal{P}}^{max}$. We measured the actual mutual understanding as the number of time the human successfully predicted the robot. Emoji buttons were used to estimate on-line subjects appreciation of robot's decisions. In order to track the gaze direction of subjects, we used a system similar to Attention-tracker [Lemaignan *et al.*, 2016], improved with OpenFace Library [Amos *et al.*, 2016]. This system is available on GITHUB. As in [Lemaignan *et al.*, 2016], we measured an on-line estimation of *with-me-ness*. In our setup, *with-me-ness* was defined by the frequency a subject looks at the screen or at the head of the robot, over an exponential moving average:

$$W^t = 0.9 * W^{t-1} + 0.1 * \mathbb{1}^t_{targets}$$

In the above equation, W^t represents our estimated *with-me-ness* at time t. $\mathbb{1}^t_{targets}$ equals 1 if the subject is looking at the screen or the head of the robot at time t, otherwise it equals 0. This is a simplification of the original definition of *with-me-ness* where targets (robot's head and screen) are the same in all phases of the interaction.

The questionnaire was designed in three parts. The first part contained five questions regarding the appreciation of the subject: three about the resulting story (bad – good, not funny – funny, coherent – absurd) and two about feeling during the co-creation (negative – positive, bored – excited). The second part was a randomly shuffled Godspeed questionnaire [Bartneck *et al.*, 2009]. The last part contained four questions concerning the perception of mutual understanding:

- Do you think the robot took into account your choices?
- Do you think the robot took into account your predictions?
- Do you think the robot was predicting your choices?
- Were you able to predict the robot choices?

In all parts, subjects had to pick a number over a type-Likert scale between 1 and 6, in order to avoid middle points and to force them to settle between the two opposite answers.

	Measure	Condition	Observation	Comparison	Mann-Whitney rank test
M.U.	# successful predictions	pred.	M=2.38, SD=2.54	pred.>adv.	stat=62.5, p<.05 (*)
		adv.	M=1.31, SD=.71	pred.>rand.	stat=44, p<.05 (*)
		rand.	M=1.07, SD=.99	adv.≠rand.	N.S.
Perceived M.U.	Subject predicts robot	pred.	M=4.0, SD=1.5	pred.>adv.	stat=53, p<.01 (**)
		adv.	M=2.43, SD=1.49	pred.≠rand.	N.S.
		rand.	M=3.73, SD=.86	rand.>adv.	stat=56, p<.01 (**)
	Robot predicts subject	pred.	M=3.5, SD=1.12	pred.>adv.	stat=59, p<.01 (**)
		adv.	M=2.37, SD=.85	pred.>rand.	N.S.
		rand.	M=2.93, SD=1.52	rand.>adv.	N.S.
	Robot uses subject choices	pred.	M=4.93, SD=.88	pred.>adv.	stat=68, p<.05 (*)
		adv.	M=4.0, SD=1.6	pred.>rand.	N.S.
		rand.	M=4.6, SD=1.84	rand.>adv.	N.S.
	Robot uses subject predictions	pred.	M=3.62, SD=1.73	pred.>adv.	stat=62, p<.01 (**)
		adv.	M=2.37, SD=1.23	pred.>rand.	stat=69, p<.05 (*)
		rand.	M=2.6, SD=1.44	rand.>adv.	N.S.
Godspeed	Anthropomorphism	pred.	M=3.93, SD=1.18	pred.≠adv.	stat=2581, p<.05 (*)
		adv.	M=3.53, SD=1.42	pred.≠rand.	N.S.
		rand.	M=4.09, SD=1.29	rand.>adv.	stat=2212, p<.01 (**)
	Intelligence	pred.	M=4.52, SD=.73	pred.≠adv.	N.S.
		adv.	M=4.35, SD=.98	pred.≠rand.	N.S.
		rand.	M=4.71, SD=.99	rand.>adv.	stat=2217, p<.001 (***)
	Animacy	pred.	M=4.13, SD=1.19	pred.>adv.	stat=3550, p<.001 (***)
		adv.	M=3.70 ,SD=1.47	pred.≠rand.	N.S.
		rand.	M=4.20, SD=1.54	rand.>adv.	stat=3323, p<.01 (**)
	Likability	pred.	M=5.08,SD=.72	pred.>adv.	stat=2228, p<.001 (***)
		adv.	M=4.64,SD=.70	rand.>pred.	stat=2462, p<.05 (*)
		rand.	M=5.38,SD=.42	rand.>adv.	stat=1556, p<.001 (***)

Table 1: Statistical results. **First column**: observed means and standard deviations of different discrete measures. **Second column**: comparison of observed distributions using Mann-Whitney rank test with continuity correction. **M.U.**: measured mutual understanding (number of successful predictions). **Perceived M.U.**: answers to ToM part of the questionnaire. **Godspeed**: answers to Godspeed part of the questionnaire.

6 Results

6.1 Model accuracy

We compared the observed accuracies (frequency that $\mathtt{e}_{\mathcal{H}} \in \mathtt{C}_{\mathcal{H}}^{max}$ and that $\mathtt{e}_{\mathcal{P}} \in \mathtt{C}_{\mathcal{P}}^{max}$) with uniform distribution over the suggested set of element at each activity's step (figure 4). We observed frequencies significantly higher than random odds for rich context-depending steps (protagonist's name, favorite drink, job and weapon, 2nd character's type). Focusing on figure 4B, we could only predict subject's predictions better than randomly at the beginning of the interaction after which, in both random and adversarial conditions, it became too difficult for subjects to infer robot's intentions.

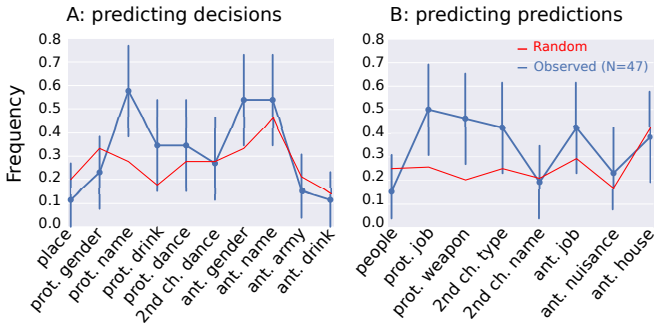

Figure 4: Model accuracy vs random probability. A: (blue) frequency that $\mathtt{e}_{\mathcal{H}} \in \mathtt{C}_{\mathcal{H}}^{max}$. B: (blue) frequency that $\mathtt{e}_{\mathcal{P}} \in \mathtt{C}_{\mathcal{P}}^{max}$. (red) probability of picking the most likely context from a random decision.

6.2 Actual vs perceived mutual understanding

As expected, choices of the robot in the predictable condition were more susceptible to be predicted by subjects. The number of successful predictions was higher in predictable condition than in adversarial and random conditions. We obtained similar results with the average intensity of answers (1=Not at all, 6=totally) to the question *"Were you able to predict the robot choices?"*, meaning subjects were aware of the difficulty to predict the robot in the adversarial and random conditions. However, to the questions *"Do you think the robot took into account your choices"* and *"Do you think the robot was predicting your choices"*, subjects gave higher scores in the predictable condition than in the adversarial condition, but no differences between predictable and control conditions were found. The robot took into account subjects predictions only in predictable and adversarial conditions. But when we asked subjects to answer the question *"Do you think the robot took*

into account your predictions", we found that answers intensity was significantly lower in the adversarial condition than in both predictable and random conditions. Observations and statistics are reported in the two first part of Table 1.

6.3 Appreciation

The First part of our questionnaire concerned the appreciation of the activity and the created story rather than the robot. However, answers of subjects were similar in the three conditions (*was the story good?*: M=5+/-0.12, *funny?*: M=5+/-0.2, *absurd?*: M=4+/-0.1, did you felt positive?: M=5+-0.1, excited?: 4.7+/-0.1). We used emoji buttons in order to capture on-line judgment of the robot by subjects. Unfortunately, the usage of these buttons (9.7 presses/subject) was too rare to obtain small enough standard deviations required for significant results. Despite this fact, we observed more presses in the adversarial condition (M=11.2, SD=7.9) than in predictable (M=8.15, SD=6.8) and random (M=9.38, SD=8.1) conditions. This higher usage of button in the adversarial condition is observed in all buttons separately, except for the "absurd" emoji button that was more used in the random condition. The Godspeed part of the questionnaire contains questions asking for a judgment of the robot. The difference with emoji buttons was the fact these judgments were not direct responses to particular choices of the robot, but rather global feelings about its aspect and behavior remaining after the interaction. These questions can be sorted into 4 groups: anthropomorphism, animacy, intelligence, and likability. We concatenated answers to questions belonging to the same group. We observed lower appreciations in the adversarial condition compared to predictable and random conditions in all other groups of questions. For anthropomorphism, answers from the adversarial condition were significantly lower than from predictable and random. A similar observation concerning animacy, with answers from the adversarial condition being lower than from predictable and random. For perceived intelligence, answers from the adversarial condition were lower than from random condition. The highest gap concerned answers to likability questions: answers from the adversarial condition were significantly lower than from predictable and random conditions. Interestingly, we also found a significant preference for the random condition compared to predictable condition. Godspeed measures and statistics are reported in the third part of Table 1.

6.4 Attention

Results obtained concerning attention will be discussed in detail in a longer version of this paper. Nevertheless, we can report that we obtained a set of time series representing evolution of *with-me-ness* for each condition. While measures in predictable and random conditions where correlated (Pearson's correlation between average curves: $r(540) = 0.75, p < .001$), the set of curves obtained in the adversarial condition deviated in average to stay at a lower level of measured *with-me-ness*.

7 Discussion

Regarding mutual modeling results, it seems that subjects were aware of their ability to predict the robot, but other questions of the last part of the questionnaire show how they perceived the adversarial condition as a lack of understanding in the robot. As expected, the adversarial condition generated a perception of the ToM reasoning of the robot significantly lower than in the predictable condition, but even lower than control condition concerning the impact of subjects predictions. Beside, it seems that the decision mechanism of the robot in the random condition was overestimated, being not differentiable from the predictable condition. We can associate these different perceptions of robot's decision making with tracked attention results, in which trajectories from predictable and random condition were similar while trajectories from adversarial condition were significantly lower during three phases of approximately 50s. We can also explain Godspeed results in which concerns robot's anthropomorphism, intelligence and animacy, for which, while no difference was observed between predictable and control conditions, robot's qualities were perceived significantly lower in the adversarial condition than in control and, except for intelligence, significantly lower than in predictable condition. However, an unexpected observation concerned answers to the Godspeed likability questions, according to which the robot was even more appreciated in the random condition than in the predictable condition. A possible interpretation could be that the random condition was least boring than the predictable condition. We could even suggest that in predictable and adversarial condition, subjects started to create a coherent story while in the random condition, they were directly tempted by the robot in making incoherent decisions, and perceived that this incoherence came from a mutual agreement with the robot. Another reason why the appreciation was lower in the adversarial condition can be the fact the robot starts by being coherent and so does the subject, and when suddenly the robots makes an unexpected decision the subject is disappointed or frustrated. All the code used in this experiment is open-source and available at <Site For Code To Be Available>. We hope our description of the experiment is detailed enough to ensure reproducibility of our results. However, we have to warn the fact we obtained these results in a biased population of engineering students and may not be observed in a different population, especially in children.

8 Conclusion

This experiment was a preliminary study for further explorations with the story co-writing interaction. We wanted to test our different conditions of ToM-behavior first with adults who would be more indulgent and least impacted by a robot's behavior. Thanks to these results, we know that different conditions of robot's ToM based behavior can strongly affect robot's appreciation and subjects attention. This also open the possibility to control the quality of interactions by seeking optimal 2nd-order ToM reasonings and behaviors. In future works, we will study pure human-agent interaction (without robot) through a large-scale experiment. For this we will deploy our activity's interface on a website. The goal will be to improve our ToM model by analyzing patterns in humans decision making. Then, we will use the improved model for real-world Child-Robot Interaction in pedagogical contexts.

References

[Amos *et al.*, 2016] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

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

[Bartneck *et al.*, 2009] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[Breazeal *et al.*, 2009] Cynthia Breazeal, Jesse Gray, and Matt Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *The International Journal of Robotics Research*, 28(5):656–680, 2009.

[Clark and Brennan, 1991] Herbert H Clark and Susan E Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991.

[Devin and Alami, 2016] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*, pages 319–326. IEEE, 2016.

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

[Jacq *et al.*, 2016a] 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.

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

[Johal *et al.*, 2016] Wafa Johal, Alexis Jacq, Ana Paiva, and Pierre Dillenbourg. Child-robot spatial arrangement in a learning by teaching activity. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 533–538. IEEE, 2016.

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

[Lemaignan and Dillenbourg, 2015] 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.

[Lemaignan *et al.*, 2016] 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.

[Nikolaidis *et al.*, 2016] Stefanos Nikolaidis, Anca Dragan, and Siddhartha Srinivasa. based legibility optimization. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*, pages 271–278. IEEE, 2016.

[Premack and Woodruff, 1978] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind. *Behavioral and Brain sciences*, 1(4):515–526, 1978.

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

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

[Tanaka and Matsuzoe, 2012] Fumihide Tanaka and Shizuko Matsuzoe. 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.

[Warnier *et al.*, 2012] M. Warnier, J. Guitton, S. Lemaignan, and R. Alami. When the robot puts itself in your shoes. managing and exploiting human and robot beliefs. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 948–954, 2012.