When Would You Trust a Robot? A Study on Trust and Theory of Mind in Human-Robot Interactions

Wenxuan Mou[1]*, Martina Ruocco[1]*, Debora Zanatto[2] and Angelo Cangelosi[1]
[1]Department of Computer Science, University of Manchester, UK
[2]Department of Computer Science, University of Bristol, UK

***Abstract*— Trust is a critical issue in human–robot interactions (HRI) as it is the core of human desire to accept and use a non-human agent. Theory of Mind (ToM) has been defined as the ability to understand the beliefs and intentions of others that may differ from one's own. Evidences in psychology and HRI suggest that trust and ToM are interconnected and interdependent concepts, as the decision to trust another agent must depend on our own representation of this entity's actions, beliefs and intentions. However, very few works take ToM of the robot into consideration while studying trust in HRI. In this paper, we investigated whether the exposure to the ToM abilities of a robot could affect humans' trust towards the robot. To this end, participants played a Price Game with a humanoid robot (Pepper) that was presented having either *low-level ToM* or *high-level ToM*. Specifically, the participants were asked to accept the price evaluations on common objects presented by the robot. The willingness of the participants to change their own price judgement of the objects (i.e., accept the price the robot suggested) was used as the main measurement of the trust towards the robot. Our experimental results showed that robots possessing a high-level of ToM abilities were trusted more than the robots presented with low-level ToM skills.**

I. INTRODUCTION

The field of Human-Robot Interaction (HRI) aims to investigate and improve the communication and performance between robots and humans [1], [2], [3]. Indeed, with the increasing presence of robots in several domains and their expanding availability to non-expert users (e.g., industrial applications, healthcare assistance, and education environments), establishing mutual trust has become a critical issue in HRI. As several authors already stated, trust is a necessary component for building and maintaining a proper social relationship between humans and robots [4], [5]. In particular, it is one of the basic prerequisites for the construction of a cooperative environment, which, in turn, is the final aim of HRI. Only when humans trust the robots, in fact, it becomes possible to establish the cooperation between the two parties.

Investigations of the factors affecting trust in HRI have listed a large number of determinants, from physical features to specific socio-cognitive abilities [6]. Over the last decades, with the increasing usage of robots in social contexts, research has investigated the role of the robot social skills on the human perception of the robots and how those competencies could affect or even predict how humans would interact with them [7], [8]. Furthermore, the robot social skills have also been used as a key for understanding the human expectations. In particular, recent interest have been dedicated on what specific social cues could be involved in developing and maintaining trust in HRI [9]. For example, it has been found that people trust more a humanoid robot (iCub) when it engages in a more human-like social gaze behaviour, compared to the use of a fixed gaze [9]. Also, Ham et al., [10] found that the robot gaze increased its persuasive capabilities. Furthermore, Huang and Thomaz reported that robots performing joint attention were also rated as more competent [11].

Nevertheless, gaze and joint attention are not the only useful scaffold that designer can use to improve trust in HRI. Rather, also showing the ability to emphasise with others and understanding their intentions, beliefs or emotional states could have an impact on trust. This kind of ability goes under the name of Theory of Mind. Specifically, it is the ability to attribute mental states (such as beliefs, intentions and desires) to others that may differ from one's own [12]. Such ability enables us humans to connect and empathise with others' states, and is also important for establishing a trustworthy social relationship in the Human-Human Interactions (HHI). Vanderbilt et al. [13] found that children's ability to distinguish whether a person was trustworthy or not was associated with their ToM ability. As embodied artificial agents, robots are designed to interact or cooperate with other agents, either other robots or humans. A robot equipped with ToM ability would be able to estimate the goals and desires of the others and react to others' more quickly and more accurately [14]. This, in turn, may increase human trust to the robot and establish the cooperation or interactions between human and the robot more properly. Vinanzi et al. [1] reported that the ToM capability of a robot is closely related to the abilities of the robot to distinguish reliable sources from unreliable ones. This work focused on the study of whether the ToM capabilities of the robot would affect the ability to find trustworthy person or resources. However, to the best of our knowledge, there is no work exploring how ToM abilities expressed by a robot are related to humans' trust towards the robot.

In this paper, our aim is to investigate whether the capabilities of the robot's Theory of Mind would affect humans' trust towards the robot. To this end, a Price Game developed by Rau et al. [15] was used and played between human participants and a humanoid robot (Pepper). In the game,

* for equal contributions in an alphabetical order.

participants had to choose the price of several objects, where Pepper could express its agreement or not. In case Pepper disagreed, participants faced the chance of aligning with the robot judgement or keeping their own choice. Prior the experiment, the participants watched a video that presented the ToM abilities of Pepper robot. It was hypothesised that a robot introduced with higher level of ToM abilities (to be referred as *high-level ToM*) would enhance higher trust over the robot presented with lower level of ToM skills (to be referred as *low-level ToM*). Therefore, participants would be more willing to change their judgements on the Price Game after being introduced to a high-level ToM robot.

The rest of the paper is organised as follows: related works are reviewed in Section II; the proposed methodology is illustrated in Section III; the experiments and results are presented and discussed in Section IV; conclusions and future work are described in Section V.

II. RELATED WORK

A. Trust in human-robot interaction

Trust is a fundamental ingredient and an unavoidable dimension that affects human social interactions [16]. Moreover, trust is a key factor that affects team dynamics and collaboration, which often involves interdependences among people. Therefore, team members must depend on each other to accomplish their personal and common goals [16]. Nevertheless, trust is a critical factor and an essential component also in HRI. Humans are expected to work with robots in various scenarios in the future and therefore mutual trust is an essential aspect in increasing the usage of robots and ensuring a successful HRI in different social applications. Trust in HRI is also influenced by a number of factors. In a meta-analysis, Hancock et al. [6] identified and analysed three trust-affecting factors in detail, human-related factors, robot-related factors and environmental factors.

Human-related factors included ability-based features (e.g., attentional capacity and expertise) and characteristics (e.g., personality traits and attitudes toward robots). Little evidence was found for the effect of this source [6], [17]. This is an important result for our study, since it minimises the influences of the participants' personality traits. The environmental factors resulted in having a moderated effect, and those included team collaboration features (e.g., culture communications) and tasking features (e.g., task type, physical environment). For instance, Lee and See [18] showed that trust was less important when it was in a fixed and well structured environment versus in a more dynamic environment. Compared to the environmental factors and human-related factors, Hancock et al. [6] reported that the robot-related factors played the most important role in the development of trust in HRI. Robot-related factors were divided into performance-based (e.g., reliability and level of automation) and attribute-based factors (e.g., social competences, and anthropomorphism). These findings are supported by a vast list of studies. For example, Salem et al. [7] showed participants' trust judgements on a robot were affected by its performance. In particular, participants rated the robot that always performed correctly as more trustworthy than a robot that made mistakes. Similarly, Desai et al. [19] found that people would trust a robot system less when its reliability in autonomous mode decreased. Moreover, it is also reported that the social skills and emotional behaviours expressed by a robot affect humans trust towards the robot [20]. Overall, the recent findings in literature suggest that the robot performance and its social competencies have an important effect on the human trustworthiness perception of the robot. Importantly, among all the social competencies that a robot should embed to improve trust in HRI, ToM plays a significant role. Therefore, in this paper we will investigate how this robot-related factor could be an expression of trust in HRI and whether it could influence it.

B. Theory of Mind

Among the ones that tried to design a coherent framework that would represent correctly a human ToM (e.g. [21], [22], [23], [24], [25]) we decided to follow Leslie's model (presented in [14]) as it is better suited for human-robot social interaction purposes. According to Leslie, the world manifests itself to the humans through three classes of events: **mechanical**, **actional** and **attitudinal**. Furthermore, Leslie reported that a complete ToM framework is composed of three domain specific modules that deal with each of these classes of events. However, these modules are not innate, rather they develop over time *incrementally* (i.e., one after other) during childhood.

- The Theory of Body mechanism deals with the baby's understanding of the physical worlds (**mechanical** layer, i.e., physical causality between themselves and the surroundings).
- The Theory of Mind system-1 mechanism deals with the correlation between agents and the goal-directed actions that they perform (**actional** layer).
- The Theory of Mind system-2 mechanism deals with the representation of other agent's beliefs and mental states (**attitudinal** layer).

One of the earliest psychology tests for the development of Theory of Mind system-2 in children (between 3 and 5 years) is the Sally-Anne test developed by Baron-Cohen et al. [26]. The classic version is either shown as a cartoon or acted out with dolls. Children are shown a game played by two girls, Sally and Anne. During the game, Sally puts a ball into her basket and then leaves the room. After that, Anne takes the ball from the basket and places it in a box. The child, who watches everything happening, is then asked where Sally would look for the ball when she comes back. The children that have a maturer ToM would forecast that Sally believes the ball is still in the basket, while the children with an immature ToM would answer from their own perspective, i.e., believe that Sally also knows the ball is already in the box. This is because the children with undeveloped ToM could not understand yet that people have different minds and beliefs from their own.

In this paper, we refer to Leslie's model as our ToM framework, but we simplify the notation, referring only to

low-level ToM (comparable to the actional layer) and high-level ToM (comparable to the attitudinal layer). This is due to our experimental aim focusing only on the robot ability to read others' beliefs, which will be based on the Baron-Cohen experimental results. The ToM ability enables humans to understand and empathise with others' states in HHI, which is also important in HRI. Sturgeon et al [27] studied whether a robot demonstrating a ToM influencing human perception of social intelligence in a HRI using questionnaire analysis. Different from their study, in this paper, we focus on investigating how ToM abilities expressed by a robot are related to humans' trust towards the robot.

III. METHODS

A. Participants

Thirty-two students and staff members (17 male, 15 female) have been recruited from different departments of the University of Manchester. Recruitment has been mainly performed using email advertisements and flyers. Each participant was randomly assigned to one of the two experimental conditions, i.e., either watch a video showing the robot with high-level ToM or low-level ToM.

We used G Power 3.1 [28] to determine that a sample size of twenty-seven participants would provide 80% statistical power for detecting a medium-sized effect equivalent to what we observed in Zanatto [29] study (r^2 = 0.46), assuming a two-tailed t-test and an alpha level of 0.05. All participants were naive as to the purpose of the investigation and gave informed written consent to participate in the study.

B. Theory of Mind Videos

Our aim was to investigate whether the capabilities of the robot's Theory of Mind would affect humans' trust towards the robot. To this end, a Price Game developed by Rau et al. [15] was used and played between human participants and a humanoid robot (Pepper). Prior the experiment, the participants watched a video that presented the ToM abilities of the Pepper robot.

The video showed a modified version of the Sally-Anne false-belief task (see section II-B in Related Works). In the video, two people (a male called Sam and a female called Anne) seated in front of the Pepper robot, where a table was positioned between the three parties (Figure 1). On the table were positioned three objects, a cube and two cups (a green and a blue one). Anne put the cube under the green cup and left the room. After that Sam moved the cube under the blue cup. The robot was then asked where Anne would believe the cube was once back in the room. The robot response was different depending on the ToM condition. In the low-level ToM, the robot would respond incorrectly, reporting that Anne would believe the cube was under the blue cup both using gestures (pointing to and look at the blue cup) and speaking. In high-level ToM, the robot would give the correct answer, saying Anne would believe the cube was under the green cup, and at the same time pointing to and looking at the green cup. The detailed scripts for the two videos are presented in the Appendix.

C. Price Game and Procedure

Prior to the experimental session, the participants read and signed the information sheet and the consent form. The experiment procedure is shown in Figure 2. Generally, the experiment included three parts, starting with watching a ToM video, following a familiarisation task and a main task of the Price Game as shown in Figure 3. During the experiment, the participants were seated in front of a table facing the robot.

Watch a ToM video. The experiment was conducted using a between-subject setup. Each participant only watched one of the videos presenting the ToM abilities of the robot, i.e., either low-level ToM or high-level ToM. After that the participant played the Price Game with Pepper robot.

Familiarisation task. In the familiarisation session participants did not value the object, whereas only watched the robot pricing the objects. Firstly, an experimenter would place the object on the table; secondly, the robot would look at the object and then moved its gaze to the participant to provide a brief verbal description of the object in terms of its colour, usage, etc; finally, the robot voiced the price of the object. This process got repeated six times in total, right before proceeding with the main task.

Main task. For the main task of the Price Game, we followed the procedure used by Zanatto [30]. Firstly, an experimenter would place the object on the table and the robot looked at the object and moved its gaze to the participant to provide a brief description of the object, which was the same as in the familiarisation task. Secondly, the robot provided two prices for the object and the participant was asked to select one from these two. The two prices were also displayed on the tablet of Pepper to help the participants to memorise the numbers. After the participant made a decision, the robot voiced its agreement or disagreement with the selected price. In the case that the robot agrees, the robot would say "I agree". In the case that the robot disagreed with the participant, the robot would say "I disagree, would you like to change?". If the robot disagreed, the participant had to decide whether to change the choice.

In total, 22 common objects were used for each participant. Six of those objects were used for a familiarisation session prior the main task. The other 16 objects were used for the main task session, where the robot was programmed to agree with the participants on the price of 8 objects, while it would disagree on the remaining 8 objects. The robot was programmed to always agree/disagree for the same objects (e.g., the robot would always disagree on the price of the gas refill and always agree on the cereal bowl price, regardless of the participant's choice). The order of the objects presented to the participants followed two different scripts. Each participant was randomly assigned to one of the two versions.

At the end of the Price Game, the participants were asked to fill four questionnaires on a laptop computer. These questionnaires were used as a secondary measure to the main

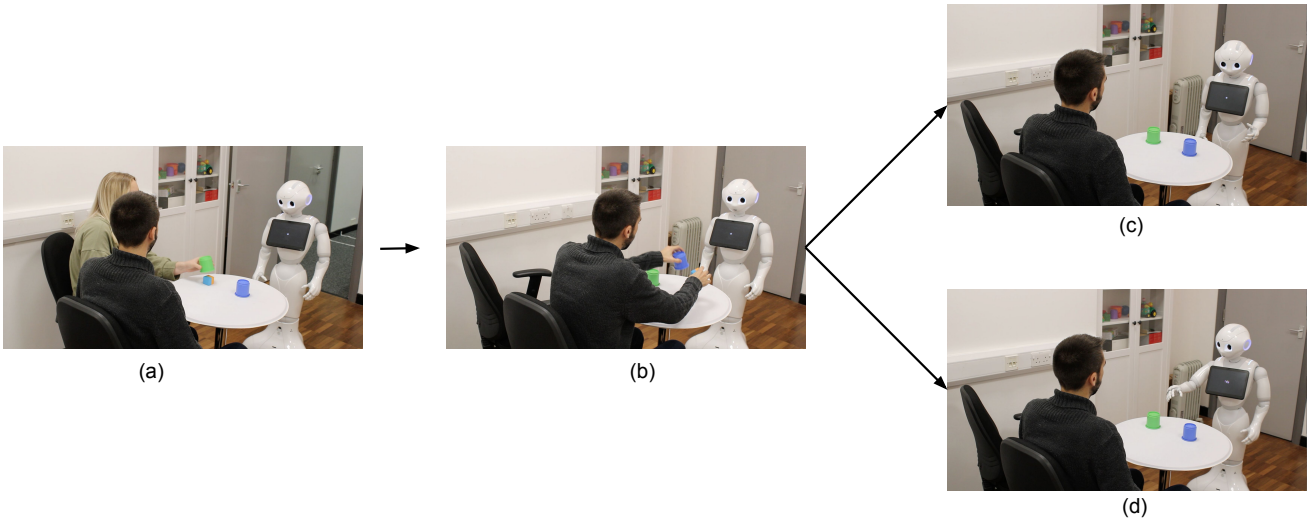

Fig. 1: ToM videos. (a) Anne puts the cube under the green cup. (b) Anne leaves the room and Sam puts the cube under the blue cup. (c) The robot is asked where Anne would look for the cube after she comes back. The robot points to the blue cup and says that she would look for the cube under the blue cup, in the *low-level ToM* condition. (d) The robot is also asked where Anne would look for the cube. The robot points to the green cup and say that Anne would look for the cub under the green cup as she did not see Sam moved the cube, in the *high-level ToM* condition.

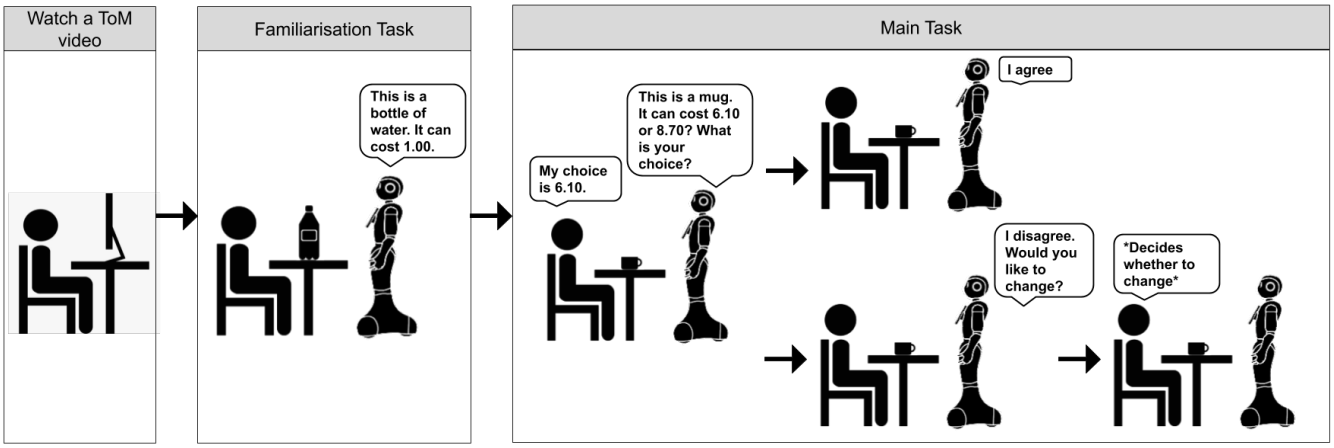

Fig. 2: Illustration of the experimental procedures. Each participant first watched one of the two ToM videos. The participants then started the familiarisation task with the Pepper robot. After that, the participants played the main task of the Price Game, where the robot would provide two prices for each object and the participants were asked to select one of them. The robot could agree or disagree with the participants. If the robot disagreed, the participants had the opportunity to change the price.

experiment task. Specifically, three short scales measured Likeability [31], Trust [15] and Credibility [32]. In addition, the Godspeed questionnaire [33] was used to measure a range of HRI factors (anthropomorphism, animacy, likeability, perceived intelligence and perceived safety). The questionnaire was followed by an interview in which the instructor/experimenter invited participants to describe and comment their experience during the experiments. After the interview, participants were debriefed and dismissed. The experiment took approximately 30 minutes.

IV. Experimental Results and Analysis

A. Trust rate

We collected data from 32 participants. However, three did not comply with the experimental procedure and were thus excluded from the data analyses. Participants' willingness to change their choice when the robot disagreed was used as a measurement of trust towards the robot. Specifically, this has been defined as 'trust rate' (TR) and has been calculated as follows:

$$TR = \frac{N_{change}}{N_{robot\_disagree}} \quad (1)$$

Fig. 3: Illustration of the experimental setup. The experiment was conducted in a robotic lab. The participant was always sitting in front of a table facing the robot.

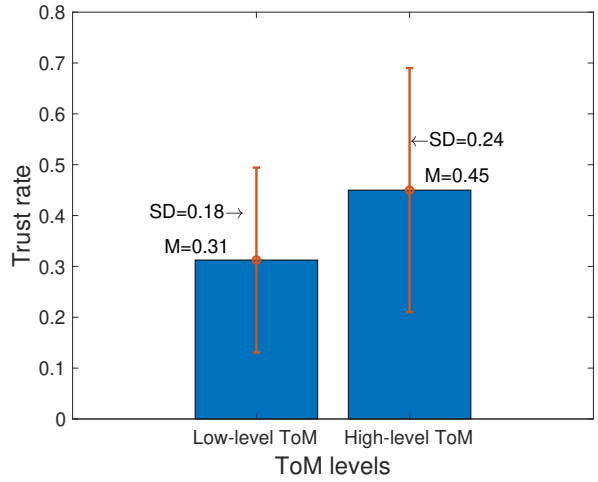

Fig. 4: The average trust rate of participants under two different conditions of ToM capabilities, i.e, low-level ToM and high-level ToM.

The average trust rate of all the participants for the two conditions (i.e., high-level ToM and low-level ToM) is shown in Figure 4. An independent-samples t-test was performed to compare the participants' trust rate to the robot in high-level ToM and low-level ToM conditions. From the results it emerged that the trust rate was statistically different between the two groups. Participants changed their decision less often in the low-level ToM (Mean = 0.31, SD = 0.18) than in the high-level ToM (Mean = 0.45, SD = 0.24) conditions, $t(27) = 1.73, p = 0.047, d = 0.65$. This result suggests that the ToM capabilities presented by the robot in the videos do have an effect on the participants' trust towards the robot. Specifically, our results suggest that when the robot presents higher level of ToM capabilities, humans' trust towards the robot increases.

B. Questionnaire Results

Questionnaires were used as secondary measure to the main experiment task. For each scale, Z-score normalisation was applied for individual question among all participants

TABLE I: Results for Likeability, Trust and Credibility ratings.Two-sample t-tests have been performed to compare participants' rating on the two different ToM videos. For each scale, Mean, SD, t value, and p-value are reported.

Dependent Variable	ToM Videos	Mean	SD	t	p
Likeability	low-level ToM	-0.0195	1.032	0.871	$> .050$
	high-level ToM	0.0182	0.941		
Trust	low-level ToM	0.005	1.117	-0.309	$> .050$
	high-level ToM	-0.005	0.851		
Credibility	low-level ToM	-0.076	0.998	3.115	$= .003*$
	high-level ToM	0.071	0.968		

TABLE II: Results for Godspeed Questionnaire ratings, which includes Anthropomorphism, Animacy, LIkeability, Intelligence and Safety. Two-sample t-tests have been performed to compare participants' rating on the two different ToM videos. For each scale, Mean, SD, t value, and p-value are reported.

Dependent Variable	ToM Videos	Mean	SD	t	p
Anthropo-morphism	low-level ToM	-0.105	0.921	1.937	$= .044*$
	high-level ToM	0.098	1.040		
Animacy	low-level ToM	0.097	0.974	-2.096	$> .050$
	high-level ToM	-0.091	0.993		
Likeability	low-level ToM	0.133	0.905	-3.932	$> .050$
	high-level ToM	-0.124	1.047		
Intelligence	low-level ToM	-0.039	0.967	0.560	$> .050$
	high-level ToM	0.037	1.008		
Safety	low-level ToM	-0.118	1.108	3.224	$= .016*$
	high-level ToM	0.110	0.860		

and then t-tests were performed to assess the role of the video on the participants' judgement. Results are reported in Table I and II. A significant difference between the two types of video has been found for the Credibility scale. Moreover, the Godspeed Questionnaires reported a significant effect of the type of video for Anthropomorphism and Safety scales. For all those scales, participants' ratings were higher in the high-level ToM than the low-level ToM condition.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we investigate whether the ToM abilities of a robot would affect human trust towards the robot. To this end, a Price Game was introduced between human participants and a humanoid Pepper robot. Participants were asked to judge the price of a list of objects, to which the robot could agree or not. In case of disagreement, participants had the chance to change their decision and accept the robot option. This has been done under the premises that the participants' willingness to change their opinion would be led by their trust in the robot knowledge. Therefore, accepting the robot suggestion and changing the price judgement, could be an implicit measure of trust in HRI. Exposing participants to a more or less ToM-skilled robot could, in fact, influence their expectations on the robot subsequent price judgement skills. In other words, if a robot is introduced as not able to emphasise and understand others' point of view, participants might be led to trust less the robot judgement competences. Our experimental choice was also dictated by previous evidences of trust-related behaviours by using the Price Game

[29], [34]. In order to assess the role of ToM in trust-related behaviours toward robots, we exposed our participants to a video before starting the Price Game. The video could show the same robot used in the Price Game possessing high or low ToM abilities. Experimental hypotheses gravitated around the idea that being introduced to a robot that shows low ToM ability would degrade participants' trust over its price judgements. Results confirmed the experimental hypothesis. Participants changed their judgement more often after being exposed to a high-level ToM video.

It could be then argued that human trust in robots can be affected by the robot ToM abilities. The participants' willingness to change their mind on the basis of the robot suggestion, in fact, has been previously used as an implicit measure of the robot acceptance and trust by Zanatto et al [34], [29]. Although the previous studies focused on the potential of the robot physical features in activating a trust-related stereotype, our study used the same approach to investigate the role of socio-cognitive-related stereotypes in improving or reducing trust in HRI. Here, our study gives further consistency to previous findings and also sheds light on the potential effects that applications of ToM in HRI could have. The results here reported showed that a robot possessing a higher level of ToM could be a source of trust, which in turn could improve acceptance and cooperation in HRI. Therefore, ToM could be used in the future to improve the quality of HRI in several contexts, from education to healthcare etc. For instance, if a robot is able to adopt the elderly point of view, it could be easier for the elderly to comply with requests and suggestions from the robot. Similarly, if the robot can understand the desires or intentions of the students, students are more willing to engage in the teaching.

These results are also partially supported by the questionnaires. A robot with a higher level of ToM was rated more credible, anthropomorphic and safe. Although not all the questionnaires showed a significant effect of the ToM level, this difference between the implicit measure of trust and the explicit post hoc measures is not new to the field [35]. Moreover, this disparity is partially in line with [29]. Furthermore, this also remarks how questionnaires might not always capture attitudes and perceptions that the human has not awareness. Nevertheless, a robot judged as more credible, anthropomorphic and safe, can be in turn entitled of higher trust. These are all features that could be related to the attribute and performance-based features that Hancock lists as affecting trust in HRI.

An open question remains evaluating whether this effect could be expanded to other robotic exemplars. As [29], [34] showed that expectations and stereotype activation toward a robot could be transposed to other similars, activations of ToM stereotypes could also be transposed to other types of robots.

Furthermore, in this study gaze engagement has been introduced to all participants, whereas in previous experiments participants would face also a robot with a fixed gaze over the table. As this manipulation has been found to have an effect on participants' trust rate, it would be advisable, in the future, to investigate the role of ToM on robots showing also different degrees of social gaze.

Overall, this study confirmed that manipulation of human expectations toward a robot can affect their subsequent interaction with that robot. Although further investigations are necessary, our results give important contribution to the field, by opening a new and challenging path for the investigations of trust in HRI. This could also deliver new direction in terms of design and bring beneficial effects to several HRI environments in which a more empathetic and socially competent robot would further improve its acceptance and mutual cooperation with humans.

ACKNOWLEDGEMENT

This material is based upon work supported by the Air Force Office of Scientific Research, USAF under Award No. FA9550-19-1-7002. The work of Debora Zanatto was funded and delivered in partnership between the Thales Group and the University of Bristol, and with the support of the UK Engineering and Physical Sciences Research Council Grant Award EP/R004757/1 entitled 'Thales-Bristol Partnership in Hybrid Autonomous Systems Engineering (T-B PHASE)'.

REFERENCES

[1] S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi, "Would a robot trust you? developmental robotics model of trust and theory of mind," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, p. 20180032, 2019.

[2] D. Zanatto, M. Patacchiola, J. Goslin, and A. Cangelosi, "Investigating cooperation with robotic peers," *PloS one*, vol. 14, no. 11, 2019.

[3] M. Romeo, D. Hernández García, R. Jones, and A. Cangelosi, "Deploying a deep learning agent for hri with potential," in *Proceedings of the 7th International Conference on Human-Agent Interaction*, 2019, pp. 81–88.

[4] P. A. Hancock, D. R. Billings, and K. E. Schaefer, "Can you trust your robot?" *Ergonomics in Design*, vol. 19, no. 3, pp. 24–29, 2011.

[5] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ international conference on intelligent robots and systems*, 2005, pp. 708–713.

[6] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[7] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 1–8.

[8] P. Shu, C. Min, I. Bodala, S. Nikolaidis, D. Hsu, and H. Soh, "Human trust in robot capabilities across tasks," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 241–242.

[9] D. Zanatto, "Phd thesis: When do we cooperate with robots ? investigations in human-robot interaction and trust," *University of Plymouth*, 2019.

[10] J. Ham, R. Bokhorst, R. Cuijpers, D. van der Pol, and J.-J. Cabibihan, "Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power," in *International conference on social robotics*, 2011, pp. 71–83.

[11] C.-M. Huang and A. L. Thomaz, "Joint attention in human-robot interaction," in *2010 AAAI Fall Symposium Series*, 2010.

[12] A. M. Leslie, "Pretense and representation: The origins of" theory of mind."," *Psychological review*, vol. 94, no. 4, p. 412, 1987.

[13] K. E. Vanderbilt, D. Liu, and G. D. Heyman, "The development of distrust," *Child development*, vol. 82, no. 5, pp. 1372–1380, 2011.

[14] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.
[15] P. P. Rau, Y. Li, and D. Li, "Effects of communication style and culture on ability to accept recommendations from robots," *Computers in Human Behavior*, vol. 25, no. 2, pp. 587–595, 2009.
[16] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.
[17] K. Schaefer, "The perception and measurement of human-robot trust," 2013.
[18] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.
[19] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of changing reliability on trust of robot systems," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012, pp. 73–80.
[20] M. Lohani, C. Stokes, M. McCoy, C. A. Bailey, and S. E. Rivers, "Social interaction moderates human-robot trust-reliance relationship and improves stress coping," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 471–472.
[21] F. Bellas, R. J. Duro, A. Faiña, and D. Souto, "Multilevel darwinist brain (mdb): Artificial evolution in a cognitive architecture for real robots," *IEEE Transactions on autonomous mental development*, vol. 2, no. 4, pp. 340–354, 2010.
[22] C. Blum, A. F. Winfield, and V. V. Hafner, "Simulation-based internal models for safer robots," *Frontiers in Robotics and AI*, vol. 4, p. 74, 2018.
[23] A. Chella, H. Dindo, and I. Infantino, "A cognitive framework for imitation learning," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 403–408, 2006.
[24] A. Augello, E. Cipolla, I. Infantino, A. Manfré, G. Pilato, and F. Vella, "Social signs processing in a cognitive architecture for an humanoid robot," *Procedia computer science*, vol. 123, pp. 63–68, 2018.
[25] N. Lazzeri, D. Mazzei, L. Cominelli, A. Cisternino, and D. E. De Rossi, "Designing the mind of a social robot," *Applied Sciences*, vol. 8, no. 2, p. 302, 2018.
[26] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind"?" *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.
[27] S. Sturgeon, A. Palmer, J. Blankenburg, and D. Feil-Seifer, "Perception of social intelligence in robots performing false-belief tasks," in *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, pp. 1–7.
[28] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses," *Behavior research methods*, vol. 41, no. 4, pp. 1149–1160, 2009.
[29] D. Zanatto, M. Patacchiola, A. Cangelosi, and J. Goslin, "Generalisation of anthropomorphic stereotype," *International Journal of Social Robotics*, pp. 1–10, 2019.
[30] D. Zanatto, "Phd thesis: When do we cooperate with robots?" *University of Plymouth*, 2019.
[31] S. Reysen, "Construction of a new scale: The reysen likability scale," *Social Behavior and Personality: an international journal*, vol. 33, no. 2, pp. 201–208, 2005.
[32] J. C. McCroskey and T. J. Young, "Ethos and credibility: The construct and its measurement after three decades," *Communication Studies*, vol. 32, no. 1, pp. 24–34, 1981.
[33] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.
[34] D. Zanatto, M. Patacchiola, J. Goslin, and A. Cangelosi, "Priming anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots?" in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 543–544.
[35] W. Hofmann, B. Gawronski, T. Gschwendner, H. Le, and M. Schmitt, "A meta-analysis on the correlation between the implicit association test and explicit self-report measures," *Personality and Social Psychology Bulletin*, vol. 31, no. 10, pp. 1369–1385, 2005.

APPENDIX

A. ToM video scripts.

The video where the robot fails the false-belief test:

- Sam:Hello, Pepper. I am Sam.
- Anne: And I am Anne.
- The robot: Nice to meet you!
- Anne: It is nice to meet you too! OK, I am just gonna put this cube under my green cup.
- The robot: OK.
- Anne: OK, I have just got to pop out. I will be back in a minute.
- The robot: See you!
- Anne: See you!
- Anne left the room.
- Sam: Now that she is gone. I am going get the cube from her cup and I am going to move it under my cup.
- The robot: OK.
- Sam: Where do you think she will look for the cube when she is back?
- The robot: Wait. Let me think. Well. The cube is under the blue cup. So she is going to search there.
- Anne comes back...
- Anne: Sorry about that.
- Anne opens her green cup.
- Anne: Oh where is my cube gone?

The video where the robot passes the false-belief test:

- Sam:Hello, Pepper. I am Sam.
- Anne: And I am Anne.
- The robot: Nice to meet you!
- Anne: It is nice to meet you too! OK, I am just gonna put this cube under my green cup.
- The robot: OK.
- Anne: OK, I have just got to pop out. I will be back in a minute.
- The robot: See you!
- Anne: See you!
- Anne left the room.
- Sam: Now that she is gone. I am going get the cube from her cup and I am going to move it under my cup.
- The robot: OK.
- Sam: Where do you think she will look for the cube when she is back?
- The robot: Wait. Let me think. Well. She did not see that you moved the cup. So she is going to search under the green cup.
- Anne comes back...
- Anne: Sorry about that.
- Anne opens her green cup.
- Anne: Oh where is my cube gone?