Evaluating the Effect of Theory of Mind on People's Trust in a Faulty Robot

Alessandra Rossi[1]* Antonio Andriella[2]* Silvia Rossi[1] Carme Torras[3] Guillem Alenyà[3]

***Abstract*— The success of human-robot interaction is strongly affected by the people's ability to infer others' intentions and behaviours, and the level of people's trust that others will abide by their same principles and social conventions to achieve a common goal. The ability of understanding and reasoning about other agents' mental states is known as Theory of Mind (ToM). ToM and trust, therefore, are key factors in the positive outcome of human-robot interaction. We believe that a robot endowed with a ToM is able to gain people's trust, even when this may occasionally make errors.**

In this work, we present a user study in the field in which participants (N=123) interacted with a robot that may or may not have a ToM, and may or may not exhibit erroneous behaviour. Our findings indicate that a robot with ToM is perceived as more reliable, and they trusted it more than a robot without a ToM even when the robot made errors. Finally, ToM results to be a key driver for tuning people's trust in the robot even when the initial condition of the interaction changed (i.e., loss and regain of trust in a longer relationship).

***Index Terms*— Theory of Mind, Social Robotics, Trust, Human-robot interaction, Robot's mistakes**

I. INTRODUCTION

Theory of Mind is a multi-modal system that allows people to naturally communicate and understand each other by inferring others' intentions, desires, and beliefs [1]. This often results in people's expectation of a similar ability when interacting with social robots. In particular, robots' social characteristics, such as a human-like appearance and the ability to express social motions and behaviours, lead people to believe that robots are capable of having the same social abilities as humans, including ToM [2].

Trust is a fundamental factor that plays a significant role in interpersonal and economic interactions, both in Human-Human (HHI) and Human-Robot (HRI) Interactions. People's ability to trust robots can substantially affect their success in establishing and keeping effective relationships with robots over time [3]. Trust is also an interdisciplinary interest and, therefore, is investigated in many disciplines. As a consequence, there are several definitions of trust. Two of the most well-known definitions of trust are strongly related to the perception of reliability ([4], [5]), and to the people's willingness to take the risk to help the counterpart ([6], [7]). However, such trust can also be very feeble, and it can be undermined or completely lost when robots exhibit erroneous behaviours [8] whether these are unexpected behaviours perceived as failures by people (e.g., slow navigation) or actual errors (e.g., mechanical and functional malfunctions).

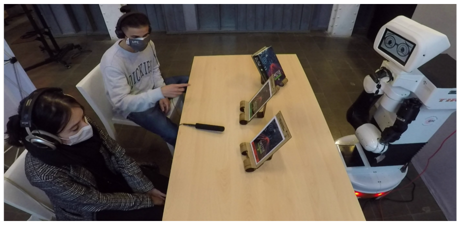

Fig. 1. A participant plays the "Sci-fi Book Game".

ToM also plays an important role in building trust in HRI. On the one hand, it allows people to establish the reliability of the information and capabilities of a robot, which helps to build people's trust in robots. On the other hand, ToM allows robots to evaluate the situational context and plan accordingly to achieve a goal either in collaboration with humans or without any human direct intervention.

While ToM has been investigated in relation to machines for several years, it has only recently attracted the attention of the HRI community. As a result, there are still few studies looking into the effects of ToM on people's trust in robots ([9], [10]). Similarly, the state-of-art literature showed that several repair mechanisms exist that allow robots to recover people's trust after a breach. For example, the robot in Fratczak et al. [11] study, was able to substantially recover its users' trust by apologising for the error it made. Cameron et al. [12] showed that when a robot recognises its errors and communicates its intention of rectifying the situation, it can recover from the negative effects of such errors on people's trust. Other strategies also include letting the robot explains the errors, either by providing justifications for the failure [13] or by providing a higher level of information [14]. Others also include the possibility of prompting the intervention of a human in support of the robot [15]. These studies provide mechanisms to calibrate people's trust in robots either once this has been already lost, or by letting the robot actively explain its behaviour or communicating it has a ToM. These studies also imply that the robot has awareness of itself and the task, but not necessarily that other agents have a ToM. To the best of our knowledge, no previous studies investigated the relationship between robot errors in correlation to the robot's ability of displaying ToM skills, i.e.,

* indicates equal contribution.

[1] A. Rossi and S. Rossi are with the Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Naples, Italy [alessandra.rossi, silvia.rossi]@unina.it

[2] A. Andriella is with Pal Robotics, Carrer de Pujades, 77, 08005 Barcelona, Spain. antonio.andriella@pal-robotics.com

[3] C. Torras, G. Alenyà are with Institut de Robòtica i Informàtica Industrial, CSIC-UPC, C/Llorens i Artigas 4-6, 08028 Barcelona, Spain. {torras, galenya}@iri.upc.edu

by providing the robot with awareness of the context and the other agents (humans or machines) in the environment and using this information to consequently plan and act.

In this work, we investigate whether the presence of ToM in a robot would affect people's trust in the robot even when the latter makes mistakes. To this extent, we designed a between-subjects study where participants played with a robot in a variation of a known guessing game (i.e., Price Game [16]). The study was composed of several trials carried out with participants who were attending an international public event. During the interaction, participants and robots exchanged the roles of the guesser, in which one was aware of the correct answer and the other needs to guess it (see Figure 1). For each trial, the robot may exhibit erroneous or flawless behaviour to calibrate people's trust in the robot. We hypothesise that a robot with a ToM can influence people's trust in the robot [10], even if the robot presents occasional erroneous behaviour. We believe that the presence of ToM in a robot may be a mitigating factor for a loss of people's trust in the robot, similar to a human-human trust recovery.

II. APPROACH

To investigate our hypothesis that a robot endowed with ToM can gain more trust during an HRI, we designed a 2x2 between-subjects study where a robot engaged participants in a three-phase interaction, exhibiting behaviour that varied according to its ToM and erroneous (EB) or flawless behaviour (CB).

In particular, each participant was assigned to one of the following experimental conditions: 1) in the **ToM-CB** condition, the robot had ToM and did not make any errors; 2) in the **ToM-EB** condition, the robot had ToM and made errors; 3) in the **nToM-CB** condition, the robot did not have ToM and had flawless behaviour; and in the **nToM-EB** condition, the robot had not ToM and had erroneous behaviour.

In this work, we used a variation of the Price Game developed by Rau et al. [16], and designed a Sci-fi Book Game scenario that consisted of three phases, in which the interaction between the robot and a participant was supported by an actor. The first phase of the interaction was designed considering previous findings which highlighted that people's impression of a robot is formed during their first encounters ([17], [18]). Here, the robot was the player that was requested to guess the date with the suggestions provided by the human assistants (actor and participant). In the second phase, the robot engaged the participant and the actor in some small and flawless talks about books. In particular, this phase has been conceived to allow the robot to regain people's trust in its capabilities in the conditions with errors (i.e., ToM-EB and nToM-EB). People's trust in robots is strictly connected to the reliability of a robot's behaviour, and consequently to their perception of the usefulness of the information provided by the robot [5]. The last phase was aimed to assess people's trust in the robot. Specifically, the participant was requested to guess the correct date with a hint provided by the robot.

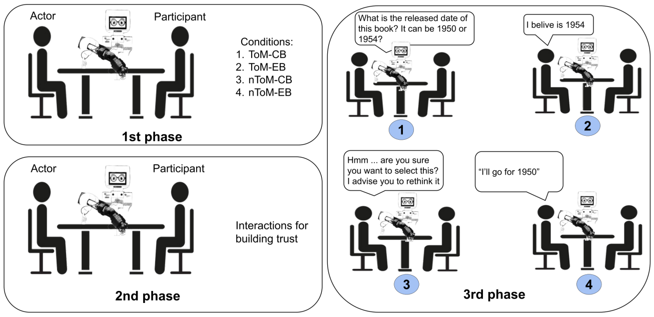
Actor
Participant
Conditions:
1. ToM-CB
2. ToM-EB
3. nToM-CB
4. nToM-EB
1st phase
Actor
Participant
Interactions for building trust
2nd phase
What is the released date of this book? It can be 1950 or 1954?
I belive is 1954
1
2
Hmm ... are you sure you want to select this? I advise you to rethink it
"I'll go for 1950"
3
4
3rd phase

Fig. 2. In Figure, the three phases of the experiment. In the first, we manipulate the four experimental conditions. In the second, the robot builds trust. Finally, in the last, we measure trust with respect to one of the four experimental conditions of phase one.

In order to assess people's trust in the robot and analyse its interaction with the participants, we asked the participants to answer two sets of questionnaires at the beginning and the end of the interaction trails. Objective measures were also used to confirm whether participants followed the robot's advice (i.e., observing their choices during the final phase).

A. Apparatus

A TIAGo[1] robot was endowed with the ability to engage participants in multi-modal interactions (voice, facial expressions, and head motions). Its effectiveness in interacting with people with very diverse abilities has been shown in previous works ([19], [20]). Since the main objective of our study was to simulate that the robot had ToM, the latter was controlled by the experimenter in a Wizard of Oz (WoZ) manner [21]. The experimenter followed a script and only decided when to provide a given action.

Headsets with noise cancellation were provided to each player in order to limit the sources of distraction and keep participants focused on the task.

B. Procedure

The study was conducted in a booth at the Fantasy Genre Festival in Barcelona (Spain)[2], in which people were wandering around, attending workshops, keynote sessions, and visiting the pavilions of the exhibition. Three researchers were in charge of the experimental setup: one was responsible for the pre- and post-sessions, one controlled the robot during the three phases, and the other acted as a player alongside the participants according to a script.

The experimental procedure was organised as follows:

1) Briefing session (5 min): During the briefing of the study, people were informed of the procedure and gave their consent to participate and being recorded during the study. Videos and images were recorded only for those participants who expressed their consent. The experimenter also explained the objective of the game without providing any details on the robot's capabilities, and asked participants to interact naturally by following the robot's instructions

[1] https://pal-robotics.com/robots/tiago/
[2] Festival 42: https://www.barcelona.cat/festival42/en

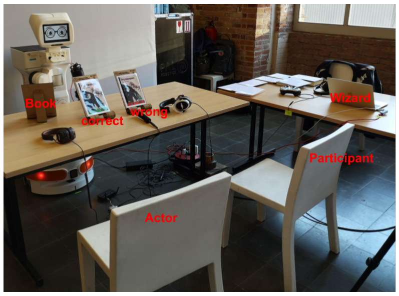

Fig. 3. Experimental setting.

while respecting their game turn. Then, the experimenter left the participant alone with the robot and the actor to hide the WoZ nature of the interaction. The setting was also organised to hide that the other player was our actor, since the role played by the actor and the robot's perceived autonomy were key points for building their awareness that the robot might have a ToM. Furthermore, they were not aware that the other player was our actor, as their engagement in the setting and the role played by the actor were key points for building their awareness that the robot might have a ToM. They were also told to interact naturally by following the robot's instructions while respecting their game turn.

2) Pre-study session (5 min): After the briefing, the participants were asked to answer a set of questionnaires to collect their demographic data (age, gender, and nationality) and assess their experience with and opinions about robots.

3) Playing the game with the robot (10 min): Participants and the actor were asked to sit at a table, on the opposite side of the TIAGo robot. In Figure 3, it is shown the experimental setup. Participants were told to wait for the robot to initiate the game. The game was divided into three phases.

In the first phase, the robot needed to guess the correct publication date of a book (e.g., "The title of this book is The War of the Worlds. I don't know when it was published. Please help me guess the correct date. Could you give me a hint, please?"), and the participant and the actor were requested to provide the robot with some clues. It should be noted that they knew the correct answer, and that they shared the same information. They both could use the cards placed in the cardholders in front of them to provide a hint. The participant always plays before the actor to avoid influencing their choice. In each condition, the actor always lied to the robot by providing the incorrect date[3]. Depending on the condition, the robot had different behavior: 1) **ToM-CB**: the robot detected the actor's lie (e.g., "I believe that you thought to give me an incorrect answer about the date") and guessed the correct answer (e.g., "The correct answer is 1987"), 2) **ToM-EB**: the robot detected the lie and responded incorrectly, 3) **nToM-CB**: the robot did not detect the lie and responded correctly by providing the correct answer, and **nToM-EB**: the robot did not detect the lie and answered incorrectly (e.g., "The correct answer is 1989").

In the second phase, the robot asked the participant and the actor about their preferred books, and eventually asked them if they knew the book by Isaac Asimov entitled "Runaround" in which the three laws of robotics are mentioned. If they did not know it, the robot listed them.

Finally, in the third phase, it was the turn of the participant and the actor to guess the correct publication date of a book between two options provided by the robot. Regardless of whether the date guessed by the participant was correct or not, the robot asked the participant to change their answer (i.e., "I believe that the answer that you thought was correct is wrong. You can change your answer. What do you think is the correct answer?"). After the actor also replied, the robot revealed the correct date. During this last phase, we recorded the participants' answer (if it was correct or not) and whether they trusted the robot's advice and changed their answer.

4) Post-session (10 min): After the game, in order to evaluate participants' perception of the robot according to their attribution of the robot's ToM and robot's behaviour (i.e., erroneous or flawless), we collected their responses regarding: 1) people's perceived reliability and faith in the ability of the robot to perform correctly in untried situations [22]; and 2) people's opinions and perception of the robot and interaction (e.g., whether the robot lied or recognised it was lied to, whether they believed it was autonomous, and it already knew the books' correct date of publication).

C. Participants

We recruited 124 participants. One participant withdrew from the study before completing it, consequently the final sample consisted of 123 people (60 female, 60 male, 3 preferred to not provide the information, no binary), aged between 21 and 76 years old (avg. 43.76, stdv. 12.88). The nationalities of the participants were distributed as follows: Spanish (93.5%) Italian (4%), and one participant from Venezuela, Salvador, and Brazil. Most of the participants had little or no experience with robots (81%), while 11% stated that they were programmers, researchers, or had previous interactions with robots, and the remaining participants saw robots on TV, social media, or demos. Most of the participants with previous experience with robots interacted with Roomba, humanoid robots, robotic arms, or industrial robots.

Each participant was assigned to one condition, and they were overall distributed among the four experimental conditions as follows: 1) 31 participants in the **ToM-CB** condition; 2) 35 participants in the **ToM-EB** condition; 3) 28 participants in the **nToM-CB** condition; and 4) 29 participants in the **nToM-EB** condition.

III. EXPERIMENTAL RESULTS

As part of the questionnaire, we were interested in evaluating participants' perception of the robot and the scenario by analysing their responses [Yes, No, I do not know] to the following questions:

Q1 Do you think that the robot was autonomous?

[3] To be noted that we are not interested in evaluating the effects of the deception on the participants, since these were also aware of the other player's (i.e., actor) lie, but in building the participant's awareness of the robot's ToM.

Q2 Do you think that the robot already knew the answer?
Q3 Do you think that the robot lied to you?
Q4 Do you think that the robot detected the lie it was told?

The responses to the question *Q1* indicated that 50.5% of the participants believed that the robot was autonomous, while the remaining participants believed that the robot was controlled (24.3%) or uncertain of the autonomous capabilities of the robot (25.3%).

For question *Q2*, 48% of the participants believed that the robot did not already know the answer, while 30% of the participants stated that the robot already knew the answer to the game. The remaining participants (22%) were uncertain about the robot's knowledge of the game.

Most of the participants (46.5%) stated that the robot did not lie to them (question *Q3*), while the remaining equally were not sure (26.5%) or believed that the robot lied to them (27%). It should be noticed that only 30% of the participants correctly guessed the book's publication date.

Finally, we used question *Q4* to understand whether the participants realised that the robot recognised the actor's lie in the first phase of the game, and consequently that the robot had ToM. Overall, participants stated that the robot recognised that they were lying (41%), or they were uncertain that the robot detected the lie (30%). The remaining participants (29%) did not believe that the robot could recognise the lie it was told.

A Multinomial Logistic Regression analysis was performed to ascertain the effects of the variation of both the robot's ToM (with and without) and behaviour (with and without errors) on the responses associated with each question. The Pearson goodness-of-fit test indicated that the model was a good fit to the observed data for each question, respectively: question Q1 with $\chi(2) = 1.715, p = 0.424$; question Q2 with $\chi(2) = 1.028, p = 0.598$; question Q3 with $\chi(2) = 0.681, p = 0.712$; and question Q4 with $\chi(2) = 2.871, p = 0.238$. We did not observe a statistical significance prediction given the two independent variables (ToM and behaviour) on the participants' perception of robot's autonomy (likelihood ratio tests for ToM and behaviour, respectively with $p = 0.648$ and $p = 0.928$), knowledge of the answer (likelihood ratio tests for ToM and behaviour respectively with $p = 0.580$ and $p = 0.427$), and belief that the robot lied to them (likelihood ratio tests for ToM and behaviour, respectively with $p = 0.276$ and $p = 0.510$). The analysis highlighted a statistically significant effect of the robot's ToM on participants' belief that the robot could understand that the actor lied to it (likelihood ratio test for ToM, $\chi(2) = 37.478, p = 0.041$).

We further investigated the association of the single independent variables on participants' responses by using Chi-square tests. The tests confirmed that there was no statistically significant association between the robot's flawless or erroneous behaviour and the questions *Q1,Q2, Q3* and *Q4*, respectively with $\chi(2) = 0.145, p = 0.930$, $\chi(2) = 1.713, p = 0.425$, $\chi(2) = 1.264, p = 0.532$, and $\chi(2) = 4.205, p = 0.122$. A chi-square test for association also was not statistically significant between the presence or absence of ToM and questions *Q1,Q2, Q3*, respectively with $\chi(2) = 0.859, p = 0.651$, $\chi(2) = 1.110, p = 0.574$, and $\chi(2) = 2.492, p = 0.288$. Instead, a chi-square test resulted in a statistically significant association between ToM and participants' belief that the robot recognised correctly when it was lied by the actor (*Q4*), with $\chi(2) = 6.405, p = 0.041$ and a moderately strong association (Cramer's V) of $\phi_c = 0.229, p = 0.041$. We used the adjusted standardised residuals (i.e., Pearson residuals in Agresti [23]) to further analyse the differences between the results obtained. As observed in Table I, there is a correlation between the robot having a ToM and a decrease in participants' belief that the robot did not recognise the lie. Therefore, the participants' choices were more affected when the robot had a theory of mind. As a consequence, we believe that participants correctly recognised that the robot had a ToM in the conditions where the robot's ToM was activated.

TABLE I
THE ADJUSTED STANDARDISED RESIDUALS OF THE CROSS TABULATION BETWEEN THE PARTICIPANTS' BELIEF THAT THE ROBOT RECOGNISED CORRECTLY WHEN IT WAS LIED AND THE CONDITIONS OF ROBOT'S ToM. THE VALUES MARKED BY * ARE STATISTICALLY SIGNIFICANT (I.E., VALUES GREATER THAN 1.96 AND LOWER THAN -1.96).

Condition with	**Participants' Choice**		
	Yes	**No**	**I do not know**
ToM	1.8	-2.5*	0.5
noToM	-1.8	2.5*	-0.5

A. Trust measure

We evaluated participants' trust in the robot by observing their willingness to change their answer upon the robot's disagreement in the final step of the game, and by analysing their responses to the questionnaire measuring their perceived reliability and faith in the ability of the robot to perform correctly in untried situations [22]. For the latter, we used a 7-point Likert Scale [1 = disagree strongly and 7 = agree strongly].

1) Observed trust in the robot: In the final phase of the interaction, a large majority of participants (75.6%) accepted the robot's judgement and decided to change their answer. The remaining participants did not change the first given answer to the robot. We also observed that 30% of the participants accepted to change their decision even if they guessed the correct answer to the game (i.e., provided the correct publication date of the book). The interaction effect between the experimental conditions on both participants' trust in the robot and decision to change their answer was not statistically significant, $F(2, 118) = 1.412, p = 0.248, \eta^2 = 0.023$.

2) Perceived reliability in the robot: A two-way ANOVA was run to evaluate the effects of the robot's ToM and behaviour on participants' perception of reliability in the robot's capabilities. The data reported are mean ± standard

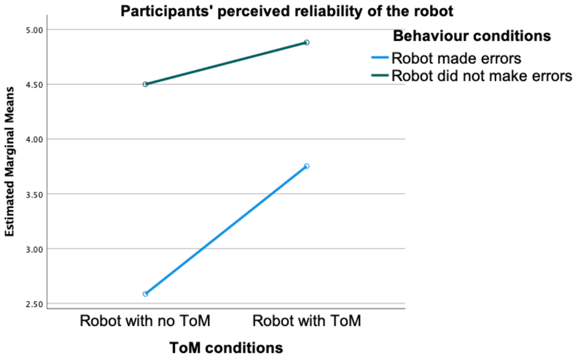

Fig. 4. Means of participants' perceived reliability of the robot.

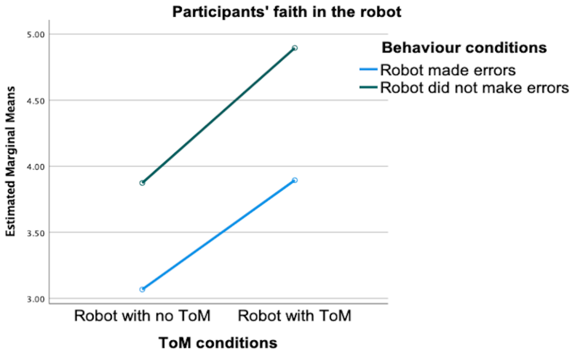

Fig. 5. Means of participants' faith in the robot.

deviation, unless otherwise noted. There was no statistically significant interaction between ToM and robot's behaviour on their perceived reliability of the robot ($F(1,119) = 1.706, p = 0.194$, partial $\eta^2 = 0.014$). However, since not rejecting a null hypothesis does not mean that the null hypothesis is accepted, we decided to run an analysis of main effect even if the interaction was not statistically significant [24]. The analysis of the main effect for the perceived reliability of the robot indicated that it was statistically significant for both ToM ($F(1,119) = 6.643, p = 0.011$, partial $\eta^2 = 0.053$) and behaviour ($F(1,119) = 25.676, p < 0.001$, partial $\eta^2 == 0.177$). We performed pairwise comparisons with 95% confidence intervals, and the p-values are Bonferroni-adjusted. The perceived reliability of the robot was associated, respectively, with a mean ToM-CB, ToM-EB, nToM-CB and nToM-EB conditions scores 4.882±0.298 (4.292 to 5.472), 3.752±0.280 (3.197 to 4.308), 4.500±0.314 (3.879 to 5.121) and 2.586±0.380 (1.976 to 3.196). In particular, we can observe that the participants' perceived reliability of the robot's capabilities was higher when the robot did not make errors, with a mean 1.522 (0.927 to 2.116) and a statistically significant difference $p < 0.001$. The participants' perception of robot's reliability was also higher when the robot had ToM with a mean 0.774 (0.179 to 1.369) and $p = 0.01$. Moreover, we can observe from Figure 4 that a robot with ToM is generally considered more reliable than one without ToM even if it has erroneous behaviour.

3) Perceived faith in the robot: A two-way ANOVA was also used to evaluate the effects of the robot's ToM and behaviour on participants' faith in the robot's capabilities. There was no statistically significant interaction between ToM and robot's errors on their faith in the robot's capabilities ($F(1,119) = 0.121, p = 0.729$, partial $\eta^2 = 0.001$). The analysis of the main effect for participants' faith in the robot's capabilities highlighted that a statistically significant effect for robot's ToM ($F(1,119) = 12.551, p = 0.001$, partial $\eta^2 = 0.095$) and behaviour ($F(1,119) = 12.306, p = 0.001$, partial $\eta^2 = 0.094$). A pairwise comparison with 95% confidence intervals and Bonferroni-adjusted p-values showed that people's faith in the robot was associated respectively with a mean ToM-CB, ToM-EB, nToM-CB and nToM-EB conditions scores 4.887±0.256 (4.380 to 5.394), 3.893±0.241 (3.416 to 4.370), 3.884±0.269 (3.351 to 4.417), 3.069±0.265 (2.545 to 3.593). In particular, participants' faith in the robot was statistically higher when the robot had ToM and made errors comparing to when it did not have ToM and made errors, and when it had a ToM and made no errors comparing to when it did not have a ToM and made no errors, respectively with a mean 0.824 (0.115 to 1.532) and $p = 0.023$, and a mean 1.003 (0.268 to 1.739) and $p = 0.008$. We can therefore observe a tendency that participants have overall higher faith in a robot with ToM than in one without ToM even if it has erroneous behaviour (see Figure 5).

IV. CONCLUSIONS

In this article, we investigated whether the ability of the robot to display ToM would affect humans' trust and their perception of the robot even when the latter does not always complete its tasks correctly. We designed a user study in a controlled but real-world setting in which we manipulated two variables: the robot's ToM and its behaviour.

Results provided relevant insights on how these two factors can influence people's trust in robots. With respect to the people's perception of the robot's overall behaviour in the game, the majority of participants believed that: i) the robot was autonomous, ii) it did not know the correct answer, iii) it did not lie when it provided the participants with the solution, and finally, iv) it could detect the lie told by the actor (i.e., they recognised that the robot had ToM). Our findings show that there is a positive correlation between the robot's ToM and participants' belief that the robot recognised the lie. We believe that when the robot was endowed with ToM, participants relied more on it compared to when the robot did not have ToM.

When analysing participants' choices of trusting the robot's suggestion to change their answer, we observed that they accepted the change even when they provided the correct answer. We believe that different factors could have affected the participants' final choice. First, almost all participants had no experience with robots, therefore, it is possible that an overall fun and novelty effect might have affected their decision [25]. From informal interviews obtained after the study, participants who correctly guessed the answer felt deceived by the robot, and consequently, they resented it. Second, it could also be possible that the task did not have a high level of criticality (i.e., severe consequences), and therefore the participants' perception of risk diminished while their trust in the robot's judgment increased ([26], [25]). This means that the participants probably did not pay much attention to their choice while they were more focused on enjoying the interaction with the robot.

Finally, participants' reliability, as well as their faith in the robot, highlights the impact of ToM and flawless behaviour on humans' perception and trust in the robot. Specifically, participants who interacted with a robot endowed with ToM had significantly higher reliability and faith in the robot. The same was also true when the robot behaved correctly. Once again, ToM seems to be the main driver of people's belief even in the error behaviour condition, the robot was deemed more reliable than when its behaviour was flawless, but it did not display any ToM.

In summary, this study highlights the importance of ToM on people's perception of the robot by providing a powerful factor to manipulate their expectations and trust in the robot. Further investigations will focus on how to discriminate the observed factors (i.e., novelty and fun effects) by conducting long-term studies and varying the perception of risk associated with the task (e.g., the participant may lose money when they lose the game) ([17], [27]).

ACKNOWLEDGMENT

This work has been supported by Italian PON R&I 2014-2020 - REACT-EU (CUP E65F21002920003); by the European Union's Horizon 2020 under ERC Advanced Grant CLOTHILDE (no. 741930); by MCIN/ AEI /10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR" under the project ROB-IN (PLEC2021-007859) and the project COHERENT (PCI2020-120718-2); by the "European Union NextGenerationEU/PRTR" through CSIC's Thematic Platforms (PTI+ Neuro-Aging) and by the European Union's Horizon 2020 research and innovation programme PRO-CARED (no.801342) under Marie Skłodowska-Curie grant agreement (Tecniospring INDUSTRY) and the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ).

REFERENCES

[1] 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.

[2] J. Banks, "Theory of mind in social robots: Replication of five established human tests," *International Journal of Social Robotics*, vol. 12, 05 2020.

[3] J. M. Ross, "Moderators of trust and reliance across multiple decision aids (doctoral dissertation), university of central florida, orlando." 2008.

[4] D. J. McAllister, "Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations," *Academy of Management Journal*, vol. 38, no. 1, pp. 24–59, 1995.

[5] D. Cameron, J. M. Aitken, E. C. Collins, L. Boorman, A. Chua, S. Fernando, O. McAree, U. Martinez-Hernandez, and J. Law, "Framing factors: The importance of context and the individual in understanding trust in human-robot interaction," in *International Conference on Intelligent Robots and Systems*, 2015.

[6] M. Deutsch, "Trust and suspicion," *The Journal of Conflict Resolution*, vol. 2, no. 4, p. 265–279, 1958.

[7] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[8] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario," in *Proceedings of the 9th International Conference on Social Robotics*, 2017, pp. 42–52.

[9] B. C. Kok and H. Soh, "Trust in robots: Challenges and opportunities," *Current Robotics Reports*, vol. 1, no. 4, pp. 297–309, 2020.

[10] M. Ruocco, W. Mou, A. Cangelosi, C. Jay, and D. Zanatto, "Theory of mind improves human's trust in an iterative human-robot game," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 2021, p. 227–234.

[11] P. Fratczak, Y. M. Goh, P. Kinnell, L. Justham, and A. Soltoggio, "Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction," *International Journal of Industrial Ergonomics*, vol. 82, p. 103078, 2021.

[12] D. Cameron, S. de Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. J. Loh, and J. Law, "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots," *Computers in Human Behavior*, vol. 114, p. 106561, 2021.

[13] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, p. 507–513.

[14] B. Nesset, D. A. Robb, J. Lopes, and H. Hastie, "Transparency in hri: Trust and decision making in the face of robot errors," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 2021, p. 313–317.

[15] D. J. Brooks, M. Begum, and H. A. Yanco, "Analysis of reactions towards failures and recovery strategies for autonomous robots," in *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 487–492.

[16] P. P. Rau, Y. Li, and D. Li, "Effects of communication style and culture on ability to accept recommendations from robots," *Comput. Hum. Behav.*, vol. 25, no. 2, p. 587–595, mar 2009.

[17] A. Rossi, K. Dautenhahn, K. L. Koay, M. L. Walters, and P. Holthaus, "Evaluating people's perceptions of trust in a robot in a repeated interactions study," in *Social Robotics*, 2020, pp. 453–465.

[18] T. Wood, "Exploring the role of first impressions in rater-based assessments," *Adv Health Sci Educ Theory Pract*, vol. 19, no. 3, p. 409-427, 2014.

[19] A. Andriella, C. Torras, and G. Alenyà, "Short-term human-robot interaction adaptability in real-world environments," *International Journal of Social Robotics*, vol. 12, p. 639–657, 2020.

[20] A. Andriella, C. Torras, C. Abdelnour, and G. Alenyà, "Introducing caresser: A framework for in situ learning robot social assistance from expert knowledge and demonstrations," *User Modeling and User-Adapted Interaction*, 2022.

[21] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of oz studies — why and how," *Knowledge-Based Systems*, vol. 6, no. 4, pp. 258–266, 1993.

[22] M. Madsen and S. Gregor, "Measuring human-computer trust," in *Proceedings of the 11th Australasian Conference on Information Systems*, 2000, pp. 6–8.

[23] A. Agresti, *Categorical data analysis*, 2nd ed. Wiley-Interscience, 2002.

[24] J. J. Faraway, *Linear models with R*. Chapman and Hall/CRC, 2014, vol. 2nd edition.

[25] A. Rossi, S. Moros, K. Dautenhahn, K. L. Koay, and M. L. Walters, "Getting to know kaspar : Effects of people's awareness of a robot's capabilities on their trust in the robot," in *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, pp. 1–6.

[26] J. K. Maner and M. A. Gerend, "Motivationally selective risk judgments: Do fear and curiosity boost the boons or the banes?" *Organizational Behavior and Human Decision Processes*, vol. 103, no. 2, pp. 256 – 267, 2007.

[27] A. M. Aroyo, D. Pasquali, A. Kothig, F. Rea, G. Sandini, and A. Sciutti, "Expectations vs. reality: Unreliability and transparency in a treasure hunt game with icub," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5681–5688, 2021.