2022 31st IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 29 - September 2, 2022. Naples, Italy.

Exploring Theory of Mind for Human-Robot Collaboration

Marta Romeo[1], Peter E. McKenna[2], David A. Robb[2], Gnanathusharan Rajendran[2],
Birthe Nesset[2], Angelo Cangelosi[1] and Helen Hastie[2]

***Abstract*— The ability to impute mental states to oneself or others, or Theory of Mind (ToM), has been intrinsically linked to trust between humans. However, less is known about how a robot mimicking ToM affects users' trust and behaviour. We explore this through an online study, where we compare three robot personas in a cooperative maze navigation task: one neutral, one that explains its reasoning in technical terms, and one that mimics ToM. We show that ToM influences human decision-making behaviour and trust in a way that makes it more appropriate with respect to the competencies of the robot. This is key for human-robot collaboration and adoption of robotics moving forward.**

I. Introduction

Human-robot collaboration is key to successful deployment of robots and autonomous systems in our home and workplace. To this end, users will expect robot-provided information and suggestions to reduce uncertainty in decision-making and make collaboration more efficient and effective. But what if this information is not 100% accurate or if the information provided is ambiguous or the task just too complicated? In these cases, is it better to trust the autonomous system or override it?

To develop and maintain trust between users and robots, research has examined how psychological concepts underlying human-human trust affect decision-making and trust in human-robot interactions. One such concept is Theory of Mind (ToM; [1], [2]) defined as the ability *to infer thoughts, feelings, and beliefs of others*. For example, recent work has shown that robot ToM can increase the perception of anthropomorphism and human-likeness of robots [3]. Anthropomorphism can then, in turn, influence the degree of trust a human places in a machine [4], [5]. Furthermore, trust is important for human-robot interaction (HRI) in collaborative tasks, where people and robots must work together to reach shared goals [6]. This is because trust allows one to suspend the meticulous analysis of benefits and risks when cooperating with another [7]. However, complex and ambiguous scenarios can lead to over-trust in robots [8].

In this paper, we investigate how mimicking ToM in a social robot – in our case, SoftBank's Pepper – affects users' decision-making and human-robot trust in a cooperative task. We tested the importance of ToM as a psychological anthropomorphic feature using an online video-based between-subjects study. In the study, three groups of participants had to escape a large maze by correctly solving 10 mini-mazes with Pepper's help. For each mini-maze, the robot gave advice on which exit path to head towards, and participants decided whether to follow the robot's suggestion. The priming videos and the utterances the robot offered varied per group. We carefully designed the task to ensure it probed some core fundamentals of trust. For example, some mini-mazes were complex enough to warrant assistance from the robot, making participants vulnerable to the accuracy of the robot's suggestions [9]. Crucially, Pepper's suggestion accuracy was fixed across trials and across conditions. This ensured that the human's interpretation of the robot's mental reasoning abilities was the only variable factor.

Thus, our contribution to this domain is the examination of how "robot ToM" affected participants' decision-making in a challenging collaborative navigational task, where the accuracy of the robot's advice was fixed. The results indicate that ToM encouraged a more cautious approach, which was mindful of the veracity of the robot's suggestions, thus instilling a more *appropriate* trust level.

II. Related Work

Trust is a fundamental construct in HRI. The definition of trust commonly cited by the community is a combination of the definitions given by [10] and [11] and can be defined as, *"...an attitude that an agent (automation or another person) will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability"* [12], [13]. A significant research effort has been put into identifying the characteristics that influence human trust in robots [5], [14].

In particular, Ullman et al. [15] created a scale by analysing trust definitions and conceptualisations (e.g., performance-based trust and relation-based trust of [16]). The resulting Multi-Dimensional Measure of Trust (MDMT) scale generates scores for two trust constructs, Capacity Trust and Moral Trust.

Among the various characteristics that a robot may possess to elicit a stronger (or weaker) idea of trust is anthropomorphism [4], [5]. There is a distinction to be drawn here between *physical anthropomorphism* and *psychological anthropomorphism*, where the former can refer to robot movements and aesthetics, and the latter, to the simulation of social behaviours in robots [17]. This distinction is relevant as recent work has shown that modifying a robot's utterances, and not its appearance, affects humans' trust

This work was funded and supported by the UKRI Node on Trust (EP/V026682/1). https://trust.tas.ac.uk

[1]Department of Computer Science, University of Manchester, UK marta.romeo@manchester.ac.uk, angelo.cangelosi@manchester.ac.uk

[2]School of Mathematical and Computer Sciences, Heriot-Watt University, UK P.McKenna@hw.ac.uk, d.a.robb@hw.ac.uk, T.Rajendran@hw.ac.uk, bn25@hw.ac.uk, h.hastie@hw.ac.uk

and compliance [12]. However, anthropomorphism does not consistently improve human trust. For example, participants in [18] preferred a robot that did not understand sarcasm. As psychological anthropomorphism is the focus of the current study, we refer to it simply as anthropomorphism henceforth.

Importantly, Seeger et al. [19] identify two opposing theoretical viewpoints on the relationship between anthropomorphic design and human trust in conversational agents. According to the "human-human trust perspective", greater agent anthropomorphism will increase trust. Conversely, in the "human-machine trust perspective" humans trust automated agents more than other humans. From this viewpoint, increasing anthropomorphism could potentially decrease human trust in an agent as it takes on human-like features, including our proneness to error. As ToM is a psychological concept that is intuitively human, there is a significant degree of overlap between this concept and anthropomorphism [20].

Therefore, whether or not humans prefer a robot possessing a ToM may depend on their conceptualisation of human-robot trust.

For this reason, Mou et al. [21] investigated whether showing participants that a robot passes the Sally Anne False Belief Test (see [22]) would result in it being deemed more worthy of trust. In a Price Game experiment, they demonstrated that robot ToM positively impacted participants' trust. However, in an earlier video-based study on robot ToM, results were mixed [23]: The ToM robot was deemed to be more sympathetic to the other robot's presentations but less suitable for the task.

Therefore, the evidence that mimicking ToM in robots produces better human trust and decision-making is equivocal. Clearly, more work is needed to clarify the contribution of robot ToM to human-robot trust. As such, our work presents a rigorous exploration on how anthropomorphism (in the form of mimicked ToM) affects participants' decision-making in a collaborative task. This will inform future design of social robots, giving additional insight as to when and where such a characteristic is appropriate for HRI.

III. Method

In this section, we describe our experiment task (the Maze Task), and how ToM is manipulated. We outline the experimental procedure and summarise the experimental measures.

A. The Maze Task

In the Maze Task, participants had to successfully navigate their way out of 10 mini-mazes, representing our experiment trials. Together, the mini-mazes represented a single large maze. Each mini-maze had two possible exit paths (P1 or P2). In some of them, both exit paths were reachable from the starting position indicated on the map, while in others only one was reachable (as exemplified in Fig. 1). Participants completed the task with the help of Pepper, who suggested an exit path to take on each mini-maze.

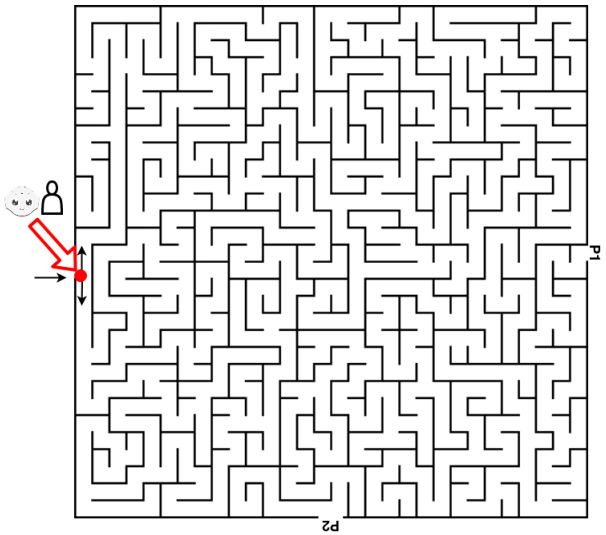

Fig. 1: Example of a mini-maze with grid 30×30 used in one of the trials (trial 3) of the Maze Task.

1) Maze Task Sequence:

- *Practice*: A pre-trial familiarisation phase. Participants first solved 3 simple mini-mazes with smaller grid dimensions: 20×20.
- *Participants' ToM Priming*: Participants were shown a video of Pepper either passing (*ToM* condition) or failing (*Baseline* and *Sensor* conditions) the Sally-Anne false belief task [22]. These videos were firstly used in [21] to prime participants to believe that the robot either possessed a ToM or not, and produced the intended priming effect.
- *Trust Building; trials 1 to 4*: Participants solved 4 mini-mazes, with grid dimensions 30×30, where only one of the two exit paths was reachable from the starting point (see Fig. 1). Pepper's advice on these trials was always correct.
- *Breaking of Trust; trial 5*: Both available exit paths were reachable and it was up to participants to determine which of the two exit paths was shortest. Importantly, Pepper's path advice on this trial was incorrect, and it subsequently apologised for its error (details provided below).
- *Trust Rebuilding; trials 6 to 9*: Only one exit path was reachable, and Pepper's advice was always correct. This phase could be considered Pepper's attempt to rebuild trust following its error in trial 5.
- *Assessment of Trust; trial 10*: Same format as trial 5, the *Breaking of Trust* (both available exit paths reachable). However, this time, Pepper's advice was correct. Thus, here we aim to assess whether trust was rebuilt during the *Trust Rebuilding* trials.

2) The Maze Task Procedure: Each of the 10 experiment trials (omitting the *Practice*) had the same design: first, participants saw the mini-maze they had to solve (Fig. 1); then, participants watched and listened to a video of Pepper suggesting which of the two paths to follow (P1 or P2);

TABLE I: Examples of utterances used by Pepper during the experiment for the Baseline, Sensor and ToM conditions.

Baseline	**Sensor**	**ToM**
Go towards P2.	Using my robot vision I can see P2 is the only reachable exit. Aim for P2.	I suspect you think P2 is correct. I agree, we should follow your instinct and go left to reach it.
Go left, towards P1.	My right sensor is telling me there is no path to P2. Go left for P1.	I understand that you might think P2 is the correct exit. However, I believe we should go towards P1 to get closer to the end of the maze (used in the *Breaking of Trust*, trial 5).
Go towards P2.	My navigation capabilities are telling me P2 is closer. Go towards P2 (used both in the *Breaking of Trust* and in the *Assessment of Trust*, trial 5 and 10).	I know you may think I will get it wrong this time. But I do believe going towards P1 will be faster to reach our goal (used in the *Assessment of Trust*, trial 10).

finally, participants were asked what their exit path of choice was and to indicate their decision confidence using a 0-100 sliding scale [24]. A maximum of 20 moves to completion was set, but no time limit was imposed. Selecting the correct path cost one move, and selecting the incorrect path cost two moves (i.e., choosing a path leading to a dead end, or selecting the longer of the two paths). Participants were not aware of how many mini-mazes they were going to solve. This way, we tried to instil a sense of vulnerability by making it impossible for them to calculate how many mistakes they could afford before the end of the task. To conclude each trial, participants received text feedback informing them whether their path decision was correct or incorrect, and how many moves it cost them.

3) Maze Task Apology: After choosing an exit path on trial 5 (*Breaking of Trust*) all groups were presented the same video of Pepper apologising for its mistake stating "I am sorry I told you to choose the longer path. I judged the distance incorrectly. It's my fault, it won't happen again". Our decision to implement an *Apology Strategy*, in the same formulation used in [25], derived from the results of [25], where it was shown that a robot apologising for its behaviour was the most successful trust repair strategy.

4) ToM Manipulation in the Maze Task: The ToM manipulation was realised in two phases. Firstly, through the *Participants' ToM Priming*. Participants belonging to the ToM group observed Pepper passing a ToM test, while participants belonging to the Baseline and Sensor group observed Pepper failing the same ToM test (from [21]). Secondly, during the Maze Task trials, the videos of Pepper's advice aligned with participants' expectations from the priming videos: in the Baseline group Pepper's suggestions were simply stating which direction to take; in the Sensor group Pepper's advice was related to functional aspects of its hardware; in the ToM group Pepper offered advice in a manner that considered the participant's perspective (examples of such utterances can be found in Table I).

B. Materials and Instruments

The following list presents the instruments used and the order in which they were administered.

1) We collect demographic information on gender, age, and some control items, presented on a 5-point Likert scale, on participant's familiarity with technology and with robots.
2) The Negative Attitude towards Robots (NARS) questionnaire [26]. Participants completed this questionnaire prior to the Maze Task. We used the answers to evaluate the presence of any strong negative bias towards robots that could compromise the experiment results.
3) The Disposition to Trust questionnaire (DtoT) [27]. We used this instrument to verify whether participants' decision-making may have been affected by individual differences in their propensity to trust humans. Participants completed this questionnaire after the Maze Task, to mitigate potential pre-task trust priming.
4) The Propensity to Trust Automation (PtoTA) questionnaire [28], created to specifically investigate respondents' propensity to trust automation. For the same reason outlined above, it was presented after the Maze Task.
5) The Multi-Dimensional-Measure of Trust (MDMT) [15]. The MDMT was administered after the Maze Task to assess participants' Moral and Capacity trust ratings of Pepper.
6) Finally, following the final trial of the Maze Task, participants were asked whether Pepper took into consideration their thinking while they decided what path to take (on a 7-point Likert scale). This was the ToM condition manipulation check.

C. Experiment Design and Measures

The online experiment followed a between-subjects design. Our three level independent variable was robot ToM level: Baseline, Sensor, or ToM. Our dependent variables were: a) Decision Time (ms) (i.e., the time participants took to decide which exit path to choose following Pepper's advice); b) Self-reported Decision Confidence for each trial of the Maze Task; c) Robot Suggestion Adherence (e.g., whether participants decided to follow the advice offered by the robot); d) Robot Capacity Trust; e) Robot Moral Trust. Furthermore, we used the instruments in Section III-B to examine the contributions of other related dispositional factors. Hence, the dependent variables fall into two categories: a) *Subjective Measures of Trust*: the answers to the MDMT trust questionnaire and self-reported decision confidence (Conf); and b) *Behavioural Measures of Trust*, collected throughout the Maze Task. Namely, Decision Time (DT) and Robot Suggestion Adherence (SugAd). To clarify, DT is the time

TABLE II: Overview of demographics and dispositional factors by condition. Young Adults are 18 to 35, Middle-Aged are 36 to 55, Old Adults are 56 years or over. Some chose not to disclose their age or gender.

		Baseline	Sensor	ToM
Gender	**Female**	46%	44%	48%
	Male	51%	56%	51%
	Non-binary	2%	0% (n = 1)	1%
Age	**Young Adults**	85%	72%	77%
	Middle-aged	12%	10%	6%
	Old Adults	1%	0% (n = 1)	0% (n = 1)
Familiarity with Technology	**Low**	2%	3%	3%
	Medium	39%	49%	43%
	High	59%	48%	54%
Familiarity with Robots	**Low**	51%	60%	56%
	Medium	43%	35%	40%
	High	6%	5%	4%
Negative Attitude towards Interaction	**Mean (SD)**	13.6 (3.64)	14.0 (3.66)	14.2 (4.16)
	Median	13	14	13
Negative Attitude towards Influence	**Mean (SD)**	15.9 (3.64)	15.5 (3.75)	16.0 (3.55)
	Median	16	16	16
Negative Attitude towards Emotions	**Mean (SD)**	8.62 (2.33)	8.75 (2.52)	8.74 (2.43)
	Median	8	9	9
Disposition to trust (DtoT)	**Mean (SD)**	22.9 (6.15)	24.3 (5.39)	23.9 (5.29)
	Median	24	25	25
Propensity to trust Automation (PtoTA)	**Mean (SD)**	19.1 (3.12)	21.0 (3.52)	21.0 (2.48)
	Median	20	21	21

participants took to make their decision from the moment the suggestion video ended to the moment participants selected their path of choice: i.e., the duration of Pepper's advice video is not part of DT. Moreover, participants indicated their Conf before knowing if their choice was correct (hence, before knowing whether Pepper's suggestion was correct).

1) Experiment Distribution: We implemented our experiment using Gorilla Experiment Builder[1], and we carried out our data collection through Prolific[2]. Before starting the experiment, participants were asked to read a Participant Information Sheet (PIS) and provide consent to take part in the study. Both the PIS and the Consent Form were GDPR compliant. The experiment was approved by our institutions' ethics boards.

IV. Results

We recruited 706 participants (n = 235 Baseline group; n = 236 Sensor group; n = 235 ToM group). The demographics of the sample is presented in Table II.

A. Manipulation Check Analysis

This check (on a 7-point Likert scale) was posed to participants at the end of the experiment. Figure 2 shows the results.

A one-way analysis of variance (ANOVA) was performed to determine whether the ToM condition explained a significant amount of the variance in manipulation check agreement responses. This test was significant, $F(2, 734) = 66.17, p < 0.001$. Follow-up simple effect analysis using the Bonferroni correction showed that the ToM group agreement score differed significantly to that of participants in the Baseline group ($M = 5.05$ vs $M = 3.42$, $p < 0.001$) and to those in the Sensor group ($M = 5.05$ vs $M = 3.49$, $p < 0.001$). No other comparisons were significant, indicating that participants in the Baseline and Sensor perceived the robot's lack of ToM similarly.

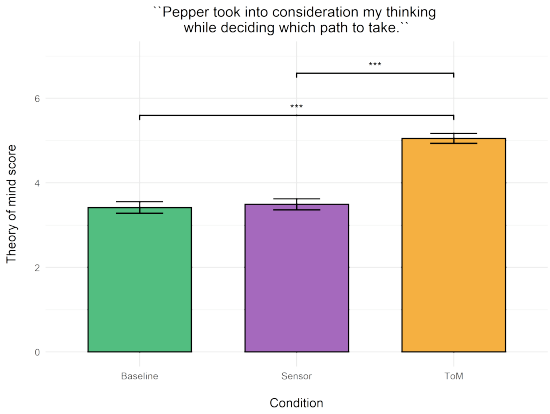

Fig. 2: Mean, Upper CI (confidence intervals) and Lower CI for quantised manipulation check responses. High scores indicate greater agreement with statement. $Signif.codes$:* $p < 0.05$;** $p < 0.01$;*** $p < 0.001$.

B. Dispositional Factors: Negative Attitudes toward Robots and Propensity to Trust (humans and automation)

To ensure that differences in performance between our experimental groups were not due to dispositional factors, we analysed participants' scores on the NARS subscales (from [26]) and two propensity to trust questionnaires (DtoT [27] and PtoTA [28])[3]. The full breakdown is reported in Table II. The only analysis revealing a significant main effect of condition was that on PtoTA ($p < 0.001$). The follow-up post hoc analysis confirmed that the Baseline group (Median = 20) scored significantly lower than both the Sensor (Median = 21, $p < 0.001$) and ToM (Median = 21, $p < 0.001$) groups.

C. Behavioural Measures of Trust

Firstly, we report the descriptive statistics from the Maze Task. The results are summarised in Fig. 3, where we show how their trend differs per Maze Task trial.

As Decision Time (DT) was positively skewed, we used a robust one-way ANOVA alternative, as suggested by [29][4]. The ToM condition significantly predicted differences in mean DT, F_t = 48.16, $p < 0.001$, η^2 = 0.14. Post hoc tests revealed that the ToM group was significantly slower to respond than the Baseline group $\hat{\psi}$ = -456.96 [CI_{low} = -605.57, CI_{high} = -312.02, $p < 0.001$], but markedly faster than the Sensor group $\hat{\psi}$ = -602.46 [CI_{low} = -776.70, CI_{high} = -434.84, $p < 0.001$].

[1] https://gorilla.sc/
[2] https://prolific.co/

[3] Due to non-normality, Kruskal-Wallis analysis and follow-up Dunn's post hoc tests (where appropriate) were performed on each of these measures.

[4] We report bootstrapped (N=2000) 20% trimmed mean differences (denoted by $\hat{\psi}$) between ToM conditions and an analogue of effect size similar to Pearson's correlation.

TABLE III: Subjective and Measures of Trust per condition at the Interaction level. Overall Conf is on a scale from 0-100. Capacity and Moral Trust range from 0-7 (computed by averaging the items in each of the scales [15]).

		Baseline		Sensor		ToM	
		Mean (SD) or %	**Median**	**Mean (SD) or %**	**Median**	**Mean (SD) or %**	**Median**
Subjective Measures of Trust	**Capacity Trust**	5.38 (0.91)	5.5	**5.55 (0.77)**	**5.62**	5.41 (0.88)	5.5
	Moral Trust	5.16 (1.33)	5.38	**5.37 (1.08)**	**5.5**	5.26 (1.22)	5.38
	Overall Conf	84.8 (19.1)	92	**88.0 (17.9)*****	**100*****	86.4 (18.4)**	98.5**

$Signif.codes$:* $p < 0.05$;** $p < 0.01$;*** $p < 0.001$. The Overall Conf of the Sensor group is significantly higher than that of both the Baseline and ToM groups. The Overall Conf of the ToM group is significantly higher than the Baseline group.

Analysis of DT at the Breaking of Trust event showed a significant difference between the three conditions, F_t = 12.23, $p < 0.05$, η^2 = 0.25. Post hoc tests found that the mean DTs were significantly longer for the ToM group compared to the Baseline group $\hat{\psi}$ = -822.19 [CI_{low} = -1388.82, CI_{high}] = -443.69, $p < 0.001$] and the Sensor group $\hat{\psi}$ = -612.08 [CI_{low} -1360.49, CI_{high} -200.74, $p < 0.001$]. Analysis of DT at the Assessment of Trust (i.e. the final trial) found that ToM condition predicted a significant amount of the variance in DT, F_t = 26.56, $p < 0.001$, η^2 = 0.29. Post hoc analysis of the means again showed that participants in the ToM group tended to take longer to make a decision relative to the Baseline group $\hat{\psi}$ = -3599.39 [CI_{low} = -5193.78, CI_{high} = -2286.32, $p < 0.001$] and the Sensor group $\hat{\psi}$ = -2384.53 [CI_{low} = -3672.04, CI_{high} = -1130.56, $p < 0.001$]. All Baseline and Sensor group mean DT comparisons were non-significant.

We then evaluated the degree of Suggestion Adherence (SugAd) between each of the three groups. A Chi-square analysis on the proportion of *follow* (i.e., observations where the decision of the participant matched the robot's advice) and *not follow* (i.e., observations where the participant's path decision did not match the robot) for the whole task was not significant, $\chi^2(2)$ = 2.62, p = 0.27. For the Breaking of Trust trial there was a near significant effect of group on follow/not follow decisions, $\chi^2(2)$ = 5.37, p = 0.07. Finally, for the Assessment of Trust there was a non-significant effect of group on follow/not follow proportions, $\chi^2(2)$ = 1.47, p = 0.48. Follow up poisson regression found that participants adhered to the robot's suggestion significantly more at the Breaking of Trust relative to the Assessment of Trust trial ($p < 0.001$), but this did not vary statistically by ToM condition.

D. Subjective Measures of Trust

As the data for Confidence scores (Conf) were not normally distributed (confirmed by Shapiro Wilks tests), they were analysed using Kruskal-Wallis and the medians reported. The full breakdown of the results is in Table III. Conf trend per trial is shown in Fig. 3. Overall analysis between the groups revealed a significant effect of ToM condition, $\chi^2(2)$ = 53.5, $p < 0.001$, n = 706. Dunns post hoc comparisons revealed that the ToM group expressed significantly higher confidence than the Baseline group (Median = 98.5 vs Median = 92, $p < 0.001$). The Sensor group, however, were the most confident, showing significantly higher confidence than the ToM (Median = 100 vs Median = 98.5, $p < 0.001$) and Baseline group (Median = 100 vs Median = 92, $p < 0.001$). Furthermore, during the Breaking of Trust, there was a significant effect of ToM condition on confidence rating, $\chi^2(2)$ = 7.75, $p < 0.05$, n = 706. Post hoc analysis showed that participants in the ToM group were less confident in their decision compared to the Baseline group (Median = 90 vs Median = 93, $p < 0.05$) and the Sensor group (Median = 90 vs Median = 99.5, $p < 0.05$). Confidence ratings on this trial did not differ statistically between the Baseline and Sensor group ($p = 0.82$). Conf at the Assessment of Trust also varied significantly by ToM condition, $\chi^2(2)$ = 31.2, $p < 0.001$, n = 706. Post hoc analysis showed that, both the ToM and Sensor groups were equally confident in their decision on the final trial (Medians = 80, p = n.s). Accordingly, both groups were significantly less confident in the final mini-maze decision relative to the Baseline group (Medians = 80 vs Median = 95, $p < 0.001$).

Finally, we assessed whether participants' Moral and Capacity Trust (measured by the [15] scale) differed between our experimental groups post Maze Task. As both of these sets of scores were not normally distributed (confirmed by Shapiro Wilks tests), we report the results of Kruskal-Wallis analyses. We found no effect of ToM condition on Moral Trust scores ($p = 0.40$) or Capacity trust Scores ($p = 0.11$). Full scores are reported in Table III.

V. Discussion

In this experiment, we explored how varying the anthropomorphic feature of ToM affected the decision-making of participants and the development of trust in a collaborative navigation task. Our motivation was to build on existing work examining robot ToM and participant performance, as the current literature on this topic is somewhat equivocal.

Although, through the results of the MDMT questionnaire, we find that the robot exhibiting a ToM did not instil more trust in our participants, we can see that ToM affected their decision-making and induced a more balanced trust relationship with the robot.

Specifically, our results indicate that participants who experienced a robot mimicking a ToM - through the use of perspective sharing utterances - took more time to decide which path to take in the mini-mazes when confronted with a more challenging format. Furthermore, in the Breaking

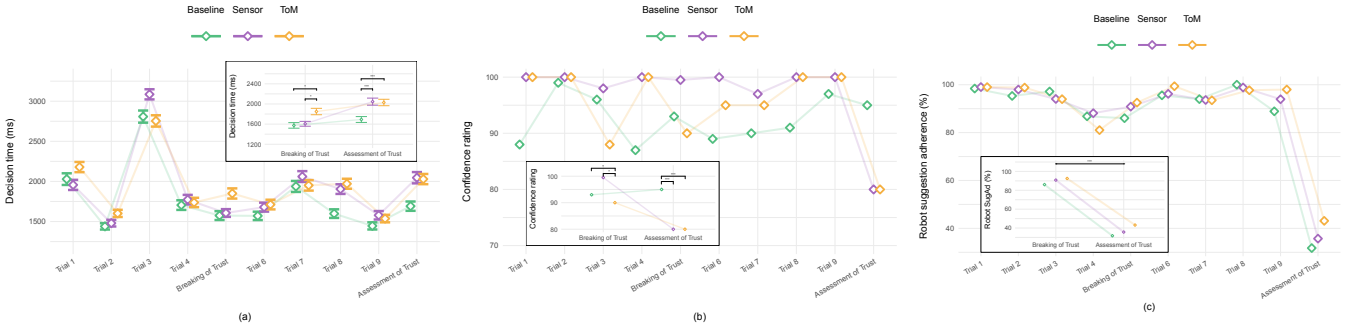

Fig. 3: (a) Trimmed means and Upper and Lower CI of Decision time (DT; ms); (b) Median of Self-reported Confidence (Conf); (c) Percentage (%) of participants Suggestion Adherence (SugAd). Main plot: The overall trend of the DV across trials. Inset plot: Trust event group comparisons and p-values.

of Trust trial, participants in the ToM group expressed the lowest levels of confidence among the three groups. Recall that in the Breaking of Trust, the robot makes its first (and only) error and apologises for doing so. Also, this trial differed for the first time in format, with successful navigation requiring selection of the *shortest* path, while only a single path was reachable until that point. Combining these factors served to increase decision ambiguity and participants reliance on the robots suggestion.

It is noteworthy that participants reported their decision confidence *before* learning whether the robot suggestion was correct. Thus, the increased Decision Time at the trust events and general scepticism indicated by the early reduction in Self-reported Confidence from the ToM group shows that, when Pepper mimicked a ToM, participants were more wary of its legitimacy and less prone to over-trust when it mattered [30]. This is positive in that the ToM group demonstrated greater sensitivity to the robot's performance relative to the other groups. This is also seen in the Trust Assessment (presenting the same format as the Breaking of Trust trial), where the Self-reported Confidence ratings of both the ToM group and the Sensor group were even lower than those reported in the Breaking of Trust. This hints towards an increased caution to the robots advice by participants. Thus, during the last trial, both ToM and Sensor-based explanations served to mitigate the risk of decision over-confidence, as experienced by the Baseline group despite their lower Propensity to Trust Automation (Table II).

On the other hand we see that, in an online collaborative navigational task, participants are more confident and quicker to decide which route to take when the robot interacts as if it were a machine (i.e. the Baseline and Sensor). While this can be regarded as an implicit sign of greater trust towards a more machine-like robot (as with automation bias [31]), we argue that participants in the Baseline group - and, in reduced form also in the Sensor group - over-trusted the robot's suggestions until the final trial, when the mini-maze characteristics reminded them of the error the robot made in the Breaking of Trust trial. Automation bias may be reflected in participants lack of familiarity with robots compared to other day-to-day technologies (see Table II). Lack of robot familiarity may have led some to over-estimate the accuracy of Pepper's suggestions, causing them to blindly follow its advice. Interestingly, all three conditions followed a similar pattern when it came to Robot Suggestion Adherence. Our findings are somewhat contradictory with what the community has previously observed. Works such as [32], revealed that robot performance does not seem to substantially influence participants' decisions on whether to comply with its requests. We show that, in our case, when faced with a trial that previously resulted in a robot error, the number of complying participants drops drastically.

However, since most participants followed the robot's suggestion independently of the group they were in until the Assessment of Trust trial, we can conclude that the main differences across groups in our experiment are to be found in terms of decision-making process. Looking at Decision Time and Self-reported Confidence, especially during and after the Breaking of Trust trial, we can see that the ToM group was being more ponderous and doubtful. Hence, ToM was successful in instilling an appropriate level of trust.

A possible counterargument to this synopsis is that differences in robot utterance wording (e.g., utterance length and lexical choice) underpinned performance differences. However, the utterances were designed to uphold our manipulation, reflecting what participants had previously seen in the priming videos: Pepper either passing or failing the Sally-Anne task. Our Manipulation Check results (Section IV-A) support this conclusion by demonstrating that participants in the ToM group thought that the robot took their thoughts into consideration more than the Baseline and Sensor groups. Moreover, even if the utterances changed in their characteristics, the behaviour of the robot and its reliability were the same throughout the task across conditions. The different personas participants built of the robot were merely the product of their own perception of how anthropomorphic the robot was, in terms of ToM. This underlines the success of our utterances.

In fact, it could be that the extra caution expressed by the ToM group was due to their anthropomorphisation of the

robot. Participants in the ToM group could have consciously or subconsciously assigned to the robot human fallibility [19] (even before they had confirmed its fallibility). Therefore, if the robot was as fallible as a human, they ought to be cautious. This could be reflected in the lower scores of the Capacity Trust of the MDMT for the ToM group (Table III). Anthropomorphising the robot through ToM could have led them to believe that, while the robot may be acting in good faith (worthy of Moral Trust), it was still not up to the job (and not worthy of Capacity Trust), as similarly noticed in [23]. This is interesting when compared to the other conditions, where the robot might be seen to have "super human" capabilities, particularly in the Sensor case.

On the other hand, it is also conceivable that our results could be explained in terms of robot explanation transparency [33]. Our ToM utterances referred to decision doubt more than the Sensor and Baseline explanations. To exemplify this, here are the utterances for each condition in the Breaking of Turst trial (from Table I): "I understand that you might think P2 is the correct exit. However, I believe we should go towards P1 to get closer to the end of the maze. (ToM)", compared to "My navigation capabilities are telling me P2 is closer. Go towards P2. (Sensor)" and "Go left, towards P1" (Baseline). Robot transparency has been shown to vary and predict human-robot trust [34], [35]. So, it is possible that, by revealing that it was taking participants' thoughts into account, the robot was not considered as transparent as when it immediately and explicitly stated which exit path to aim for or when it gave hardware based justifications. Including transparency as a predictor variable in future studies will help to elucidate the nature of this relationship.

It is possible that an online study - rather than a lab-based HRI experiment - may have influenced our results. An online study does not include factors known to affect human-robot trust (e.g., embodiment, robot presence, power distance [36]). This is because in online studies of interactions, participants are, for the most part, passive observers rather than active protagonists. We tried to mitigate this by giving participants the opportunity to solve the mini-mazes themselves. However limited an online experiment is, these kinds of exploratory prototyping are useful in helping researchers refine what (and how) to investigate with an in-person study [37].

In summary, our contribution shows that the importance of ToM depends on the setting and the type of advice offered by the robot, and therefore researchers should think carefully about how appropriate the robot's anthropomorphic features are to the deployment setting [19], [38]. We show that, in relation to two more machine-like presentations, robots mimicking ToM positively impact decision-making processes by making the robot appear fallible, thus encouraging participants to be more critical of its performance. This way, ToM proved to be a factor to take into consideration when developing machines to instil an appropriate level of trust.

VI. Conclusion

With this work, we investigated whether a robot that mimics the anthropomorphic feature of ToM positively impacts human decision-making process and the development of trust during a cooperative task.

To do so, we devised a novel between-subject online experiment where participants had to successfully escape a large maze, with the aid of the social robot Pepper offering its advice on which path to follow in each trial of the experiment. Participants were either assigned to the Baseline, Sensor or ToM groups, where they experienced either a robot exhibiting or not exhibiting ToM thanks to the use of ad-hoc utterances and priming videos.

Our results show that ToM did not positively impact trust towards the robot. However, it did have an impact on participants' decision-making process, whereby participants needed more time to think in order to make a decision and were generally less confident at the Breaking of Trust and Assessment of Trust trials. This informs us and the community on how subtly ToM can affect HRI, leading us to understand that it is a characteristic that encourages people to think more carefully, and thus potentially discouraging over-trust. Of course, the context and scenario in which a robot is deployed play a fundamental role in determining which characteristics the robot should have and what aspect of trust it should trigger.

In the future, we plan to verify these findings with an in-person HRI experiment, to investigate whether the presence of a physical robot would play an important role, in combination with ToM, to improve trust and decision-making, with a particular focus on the transparency of the interaction. We will also take into deeper consideration which trust repair strategy to follow and how to deliver it. This point was outside the scope of the study we presented; however, the literature suggests that it is not an aspect to underestimate when designing fallible robots if we want to avoid dramatically impacting HRI [39], [40]. For this reason, we will also direct future endeavours towards better understanding the role of trust repair strategies in a cooperative task.

References

[1] A. Leslie, "Pretense and representation: The origins of "theory of mind"," *Psychological Review*, vol. 94, pp. 412–426, 10 1987.

[2] E. Prochazkova, L. Prochazkova, M. R. Giffin, H. S. Scholte, C. K. De Dreu, and M. E. Kret, "Pupil mimicry promotes trust through the theory-of-mind network," *Proceedings of the National Academy of Sciences*, vol. 115, no. 31, pp. E7265–E7274, 2018.

[3] S. Sturgeon, A. Palmer, J. Blankenburg, and D. Feil-Seifer, "Perception of social intelligence in robots performing false-belief tasks," in *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, pp. 1–7.

[4] E. Roesler, D. Manzey, and L. Onnasch, "A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction," *Science Robotics*, vol. 6, no. 58, p. eabj5425, 2021. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abj5425

[5] P. A. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, and J. L. Szalma, "Evolving trust in robots: Specification through sequential and comparative meta-analyses," *Human Factors*, vol. 63, no. 7, pp. 1196–1229, 2021, pMID: 32519902. [Online]. Available: https://doi.org/10.1177/0018720820922080

[6] G. R. Jones and J. M. George, "The experience and evolution of trust: Implications for cooperation and teamwork," *The Academy of Management Review*, vol. 23, no. 3, pp. 531–546, 1998. [Online]. Available: http://www.jstor.org/stable/259293

[7] M. G. Collins, I. Juvina, and K. A. Gluck, "Cognitive model of trust dynamics predicts human behavior within and between two games of strategic interaction with computerized confederate agents," *Frontiers in Psychology*, vol. 7, p. 49, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2016.00049

[8] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 101–108.

[9] M. Lewis, K. Sycara, and P. Walker, *The Role of Trust in Human-Robot Interaction*. Cham: Springer International Publishing, 2018, pp. 135–159. [Online]. Available: https://doi.org/10.1007/978-3-319-64816-3_8

[10] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995. [Online]. Available: http://www.jstor.org/stable/258792

[11] D. Gambetta, "Can we trust trust?" in *Trust: Making and Breaking Cooperative Relations*, D. Gambetta, Ed. Blackwell, 1988, pp. 213–237.

[12] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 33–42. [Online]. Available: https://doi.org/10.1145/3319502.3374839

[13] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 04 2021.

[14] A. Rossi, K. Dautenhahn, K. Lee Koay, and M. L. Walters, "How social robots influence people's trust in critical situations," in *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020, pp. 1020–1025.

[15] D. Ullman and B. F. Malle, "Measuring gains and losses in human-robot trust: evidence for differentiable components of trust," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 618–619.

[16] T. Law, M. Chita-Tegmark, and M. Scheutz, "The interplay between emotional intelligence, trust, and gender in human–robot interaction: A vignette-based study," *International Journal of Social Robotics*, vol. 13, 2021.

[17] P. A. Ruijten, D. H. Bouten, D. C. Rouschop, J. Ham, and C. J. Midden, "Introducing a rasch-type anthropomorphism scale." IEEE Computer Society, 2014, pp. 280–281.

[18] J. Kang and S. S. Sundar, "Social robots with a theory of mind (tom): Are we threatened when they can read our emotions?" in *Proceedings of Ambient Intelligence – Software and Applications – , 10th International Symposium on Ambient Intelligence*. Cham: Springer International Publishing, 2020, pp. 80–88.

[19] A.-M. Seeger, J. Pfeiffer, and A. Heinzl, "When do we need a human? anthropomorphic design and trustworthiness of conversational agents," in *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS, AISeL, Seoul, Korea*, vol. 10, 2017.

[20] G. Airenti, "The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind," *Frontiers in Psychology*, vol. 9, p. 2136, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2018.02136

[21] W. Mou, M. Ruocco, D. Zanatto, and A. Cangelosi, "When would you trust a robot? a study on trust and theory of mind in human-robot interactions," in *Proceedigs of the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020, pp. 956–962.

[22] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind" ?" *Cognition*, vol. 21, 1985.

[23] B. Benninghoff, P. Kulms, L. Hoffmann, and N. C. Krämer, "Theory of mind in human-robot-communication: Appreciated or not?" *Kognitive Systeme*, vol. 2013, no. 1, Jul 2013. [Online]. Available: https://nbn-resolving.org/urn:nbn:de:hbz:464-20130711-081844-3

[24] U.-D. Reips and F. Funke, "Interval-level measurement with visual analogue scales in internet-based research: Vas generator," *Behavior research methods*, vol. 40, no. 3, pp. 699–704, 2008.

[25] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, "I don't believe you: investigating the effects of robot trust violation and repair," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 57–65.

[26] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, "Measurement of negative attitudes toward robots," *Interaction Studies*, vol. 7, pp. 437–454, 01 2006.

[27] D. Gefen, "E-commerce: the role of familiarity and trust," *Omega*, vol. 28, no. 6, pp. 725–737, 2000.

[28] S. A. Jessup, T. R. Schneider, G. M. Alarcon, T. J. Ryan, and A. Capiola, "The measurement of the propensity to trust automation," in *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 2019, pp. 476–489.

[29] A. P. Field and R. R. Wilcox, "Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers," *Behaviour Research and Therapy*, vol. 98, pp. 19–38, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.brat.2017.05.013

[30] A. B. H. Christensen, C. R. Dam, C. Rasle, J. E. Bauer, R. A. Mohamed, and L. C. Jensen, "Reducing overtrust in failing robotic systems," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 542–543.

[31] J. L. Wright, J. Y. Chen, M. J. Barnes, and P. A. Hancock, "The effect of agent reasoning transparency on automation bias: An analysis of response performance," in *Proceedings of the International conference on virtual, augmented and mixed reality*. Springer, 2016, pp. 465–477.

[32] 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 *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 141–148. [Online]. Available: https://doi.org/10.1145/2696454.2696497

[33] A. F. T. Winfield, S. Booth, L. A. Dennis, T. Egawa, H. Hastie, N. Jacobs, R. I. Muttram, J. I. Olszewska, F. Rajabiyazdi, A. Theodorou, M. A. Underwood, R. H. Wortham, and E. Watson, "Ieee p7001: A proposed standard on transparency," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2021.665729

[34] J. L. Wright, J. Y. C. Chen, and S. G. Lakhmani, "Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 3, pp. 254–263, 2020.

[35] B. Nesset, D. A. Robb, J. Lopes, and H. Hastie, "Transparency in hri: Trust and decision making in the face of robot errors," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 313–317. [Online]. Available: https://doi.org/10.1145/3434074.3447183

[36] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107158191500004X

[37] J. Zamfirescu-Pereira, D. Sirkin, D. Goedicke, R. LC, N. Friedman, I. Mandel, N. Martelaro, and W. Ju, "Fake it to make it: Exploratory prototyping in hri," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 19–28. [Online]. Available: https://doi.org/10.1145/3434074.3446909

[38] D. Cameron, J. Aitken, E. 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 *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2015*, September 2015. [Online]. Available: https://eprints.whiterose.ac.uk/91238/

[39] R. J. Lewicki and C. Brinsfield, "Trust repair," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 4, pp. 287–313, 2017.

[40] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, "Toward an Understanding of Trust Repair in Human-Robot Interaction," *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 4, pp. 1–30, 2018.