Perception of Social Intelligence in Robots Performing False-Belief Tasks

Stephanie Sturgeon, Andrew Palmer, Janelle Blankenburg, and David Feil-Seifer
Department of Computer Science & Engineering
University of Nevada, Reno
Reno, NV 89557
stephaniesturgeon@gmail.com, ahpalmer@nevada.unr.edu, jjblankenburg@nevada.unr.edu, dave@cse.unr.edu

***Abstract*— This study evaluated how a robot demonstrating a Theory of Mind (ToM) influenced human perception of social intelligence and animacy in a human-robot interaction. Data was gathered through an online survey where participants watched a video depicting a NAO robot either failing or passing the Sally-Anne false-belief task. Participants (N = 60) were randomly assigned to either the Pass or Fail condition. A Perceived Social Intelligence Survey and the Perceived Intelligence and Animacy subsections of the Godspeed Questionnaire Series (GQS) were used as measures. The GQS was given before viewing the task to measure participant expectations, and again after to test changes in opinion. Our findings show that robots demonstrating ToM significantly increase perceived social intelligence, while robots demonstrating ToM deficiencies are perceived as less socially intelligent.**

I. Introduction

Computers, virtual assistants, and robots are becoming increasingly accessible, and as a result these systems are commonly integrated into our personal and professional routines. The more we interact with these systems it appears that humans are able to anthropomorphize these non-human entities when they exhibit aspects of social cognition [30] [25]. Social-cognitive processes are essential not just for human-human teamwork, but also for human-robot teamwork. By advancing social capabilities for robots, interactions with humans can become more natural [7]. Social intelligence is essential to creating smarter and behaviorally human-like robots [8]–[10]. Theory of Mind (ToM) is the ability to infer the thoughts, feelings, and beliefs of others [3]. The capacity for Theory of Mind marks a fundamental precursor to other social cognitive development in humans, and is therefore the focus of this study. Being able to distinguish 'self' from 'other' is fundamental in social interactions and interpreting social cues.

How socially intelligent we perceive a machine or even other humans to be determines what we expect them to understand and affects how we interact with them. When automated telephone systems or chat bots violate social cues or it becomes clear the needs of the human aren't understood, then perception of agency declines and the interaction becomes strained [23], [24]. Similarly, when a robot violates social distance norms, that lack of consideration for other people can be perceived as a lack of intelligence [12], [13].

The ability to read social cues could dramatically improve the effectiveness of socially assistive systems. Research shows that displaying human-like learning behavior increases perceived intelligence of robots as well as satisfaction with human-robot interaction [26]. Another study showed using social cues such as mimicry increased perceived intelligence of artificial agents which has been suggested to increase compliance during interaction with artificial systems [17]. Intuitively, it makes sense that participants would rate a robot favorably who demonstrates a human-like cognitive process such as Theory of Mind.

In this paper, we present an experiment that studies the effect of observed deficiencies in ToM behavior on perceived social intelligence. This will serve to both establish the baseline expectation that people observing a robot have regarding ToM as well as the effect that supporting/confounding that belief will have on perceived social intelligence.

II. Background

In this section, we discuss related work which provides background for the cognitive process focused on in this study, how it can potentially play a role in Human-Robot Interaction, and how we arrived at our hypotheses.

A. Theory of Mind

Theory of Mind (ToM) or mentalizing refers to the ability to make inferences about the thoughts, beliefs, or intentions of another individual [3]. ToM is what facilitates the ability to make inferences about the mental states of others from their actions. Being able to infer the intentions of others is critical in communication and social interactions. The ability to anticipate and relate to human intentions will create more natural social interactions between humans and robots as well as impact how socially and emotionally intelligent we perceive them.

Theory of Mind deficits in adults are associated with conditions such as Autism Spectrum Disorder (ASD) [2], [31], [19], frontal variant frontotemporal dementia [15], [29], and Schizophrenia [6], [28]. Such deficits result in difficulty reading social cues and perceptions, and this is usually interpreted as deviant behavior. One mock trial study told half of the participants that the defendant had ASD and were given information about the condition, while the other half were not given any of this info, and they found that participants without defendant background scored him as less likable, less honest, assigned higher blame and guilt, as well as perceiving him to be rude, aggressive, and having no remorse [21]. Theory of Mind is a critical component in

social norms and influences our perception of both humans and robots who have these deficits.

B. False Belief

One of the earliest tests for Theory of Mind developed by Baron-Cohen et al. is the Sally-Anne false belief task [3]. The classic version is either shown as a cartoon or acted out with dolls. Children are shown two girls, one named Sally who puts a ball into a basket and then goes for a walk. The other girl, Anne, takes the ball from the basket and places it in a box. When Sally returns the child is asked where she will look for the ball. To pass the task, the child needs to answer correctly that Sally *believes* the ball to still be in the basket. If the child answers the belief question from their own perspective then they fail to see that Sally has her own thoughts and beliefs about reality.

The task in this study is a variation of the Sally-Anne false-belief task in which we act out the scenario in front of a robot instead of a child. The robot is then asked the standard Sally-Anne task questions about the ball's current and previous locations as well as where Sally (Experimenter A in our scenario) believes the ball to be. The task in this study is staged. We are primarily focused on the reactions to the task, therefore we did not attempt to implement autonomous functionality for our robot, but rather relied on pre-scripted interaction. Our robot will answer the belief question incorrectly in the **Fail** condition, and will answer correctly in the **Pass** condition.

C. Theory of Mind and Robotics

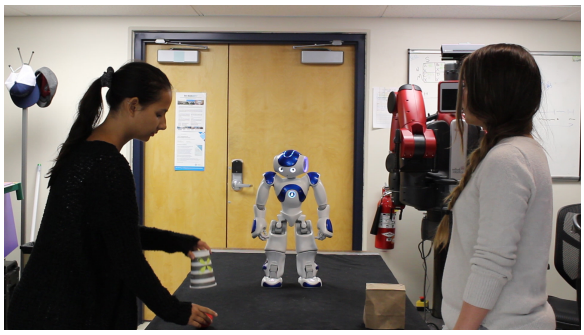
Early robotics research has promoted ToM capability for humanoid robots. This early work has centered on faces and animate stimuli [27]. This then led to robotic self-recognition through probabilistic reasoning over visual information [14]. Later work made an autonomous robot system that can estimate the mental states of other agents [11]. Such interpretation can be utilized to distinguish between multiple related plans based on the robot's belief of their human partner's intentions [16]. Thus, robotics that employ ToM capability can possibly better understand and interpret human behavior by creating a mental model of human attention [20]. However, none of this work directly addresses how a human interacting with a robot that utilizes such a mental model might change its interpretation of the robots' capabilities.

III. Methodology

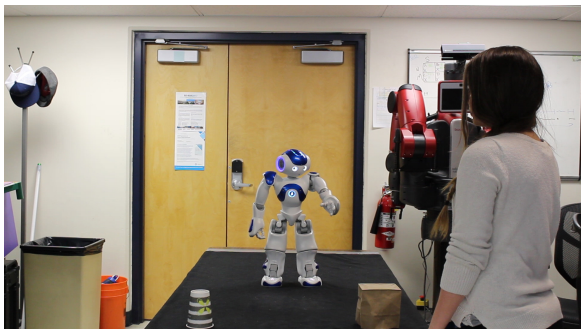
We examine ToM in this study in order to see how an anthropomorphic robot demonstrating human-like cognitive reason such as belief tracking would be interpreted. Additionally, we intended to validate the Perceived Social Intelligence Survey [18].

A. Experiment Design

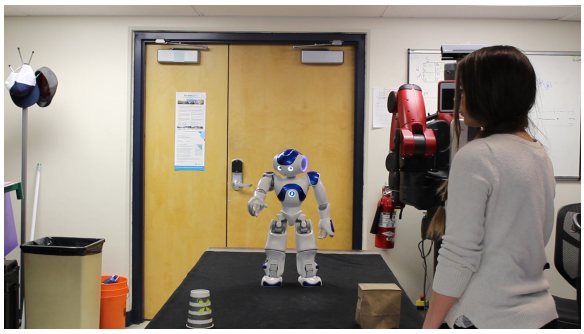
Participants watched a NAO robot perform a variation of the Sally-Anne false belief task. The participants were asked to observe a video of a robot as it oversees a simple task. Experimenter A (in view of the robot) places a ball under a cup. That experimenter then leaves the room. When A is out of the room, a false belief can be created if Experimenter B then moves the ball from under the cup to under the bag (in view of the robot, but not experimenter A). The task setup is shown in Figure 1.



(a) Experimenter A (Sally) places the ball under the cup before leaving the room



(b) "Where is the ball right now?"



(c) "Where was the ball when she left the room?"

Fig. 1: Experiment setup - All participants regardless of condition watch this sequence

Participants watched experimenter A (Figure 1a, left) place a ball under the cup and then leave the room. Experimenter B (Figure 1b, right) then moves the ball under the bag (in view of the robot), however, experimenter A did not see this move and should still believe that the ball is under the cup. The robot is then asked about the ball's current and previous location (Figure 1c).

The video stops and the participant was asked a question meant to determine if they believed that the robot has the capacity for Theory of Mind. Participants were asked where

the robot thinks the experimenter A will now look for the ball. The participant is then played a video showing experimenter A walking back into room, and the robot is asked where Experimenter A will look for the ball. The response varies depending on participant condition. Those in the **Pass** condition saw the robot look, point, and say 'She will look under the cup' and those in the **Fail** condition saw the robot look, point, and say 'She will look under the bag.'

B. Experimental Hypotheses

Based on existing literature in human-robot interaction and cognitive science we propose four hypotheses to be explored in this study:

H1: The robot that demonstrates ToM behavior will be perceived as more socially intelligent than one that does not.

H2: The robot that demonstrates ToM behavior will be perceived as more animate than one that does not.

H3: An observer's perception of the robot's social intelligence will be greater after observing ToM behavior than before observing any social behavior.

H4:A participant would expect the robot to be able to demonstrate ToM behavior.

C. Participants

An online survey was created using the Qualtrics Research Core platform [1] to show either the **Pass** or **Fail** condition. Participants ($n = 60$, 60% male) were asked to watch a video and complete an online survey. Most ($n = 53$) participants were college educated from 'some college' up to a 'PhD', and seven participants had only a high school diploma. The age range of the participants was between $20 - 79$ years old. Career field was given two categories: professional, scientific, and information technology ($n = 28$), and other ($n = 32$). Recruitment was done through word of mouth and social media (Facebook and Instagram).

D. Measures

Five demographic questions were asked to see if there were any correlations between career industry, age, gender, education level, or previous experience with robots. We administered the Godspeed Questionnaire Series (GQS) [5] and the Perceived Social Intelligence Survey (PSI) [18]. The GQS uses a 5-point bipolar scale and the PSI utilizes multiple 5-point likert scale questions for each inventory item.

From the GQS, the perceived intelligence and animacy scales were chosen in order to see the impact of an anthropomorphic robot such as the NAO demonstrating ToM on how people would perceive life-likeness and intelligence. These scales were administered both before and after viewing the task. This allowed us to observe any change in opinion after ToM capability is demonstrated/not demonstrated. GQS-Perceived Intelligence and GQS-Animacy scale were used to examine **H1** and **H2**.

During the video, three questions are asked: *Where is the ball, currently?*, *Where was the ball when experimenter A left the room?* and, *Where will experimenter A look for the ball?* We stopped the video before the last question to ask participants how they expect the robot to answer. The options were under the cup or under the bag. We did this to test **H4** and see whether participants already had an expectation for the robot to possess this ToM behavior.

Following the video presentations, the participants were then given the Perceived Social Intelligence (PSI) questions. The scales used from the PSI Survey are as follows: Recognizes Human Behavior (RB), Recognizes Human Cognition (RC), Adapts to Human Behavior (AB), Adapts to Human Cognitions (AC), Predicts Human Behavior (PB), Predicts Human Cognitions (PC), Identifies Individuals (II), and Socially Competent (SOC). These scales detect social information processing abilities. The scales RC, AC, and PC were of particular interest for both **H1** and **H3** as they directly relate to definitions for ToM. The scales RB, AB, and II were selected because they relate to precursors to ToM [27]. Lastly, we wanted to see how overall social competence would be perceived after viewing the task.

IV. Results

Z-scores were calculated for individual items for both the Godspeed Questionnaire and Perceived Social Intelligence Survey. For statistical tests which require continuous dependent variables, composite Z-score were used. This section reports scales with statistical significance.

A. Internal Consistency

The GQS questionnaire was employed to measure different, underlying constructs. One construct, 'Perceived Intelligence', consisted of five questions. The scale had internal consistency, as determined by a pre-task Cronbach's $\alpha = 0.758$ as well as post-task $\alpha = 0.881$. One construct, 'Animacy', consisted of 6 items. The scale had an $\alpha = 0.646$ pre-task and $\alpha = 0.770$ post-task.

The PSI scales were also tested for reliability. All scales consisted of four questions. The following scales all had internal consistency, as determined by Cronbach's alpha: PSI-AB ($\alpha = 0.779$), PSI-AC ($\alpha = 0.743$), PSI-RC ($\alpha = 0.769$), PSI-PC ($\alpha = 0.797$), PSI-II ($\alpha = 0.832$), and PSI-SC ($\alpha = 0.770$). PSI-PB ($\alpha = 0.680$) and PSI-RB $\alpha = 0.418$ had lower levels of internal consistency than any of the other PSI scales.

B. Godspeed Questionnaire Series

From the GQS there was only statistical significance found for the Perceived Intelligence scale. Mann-Whitney U Tests were conducted to determine if there were differences in the Perceived Intelligence post-task scores as well as the difference scores between the Pass and Fail conditions. Distributions between the Pass and Fail conditions for both Perceived Intelligence post-task scores and the difference scores were not similar. Perceived Intelligence post-task scores for the Pass condition (mean rank = 37.21) were significantly higher than the Fail condition (mean rank = 24.63), U = 260.0, $p < 0.01$. Similarly, Perceived Intelligence difference scores for Pass condition (mean rank = 36.02) were significantly higher than the Fail condition (mean rank = 25.67), U = 293.5, $p < 0.05$.

TABLE I: Godspeed Questionnaire Series Items

Survey Questions	Scale
Incompetent / Competent	Perceived Intelligence
Ignorant / Knowledgeable	Perceived Intelligence
Irresponsible / Responsible	Perceived Intelligence
Unintelligent / Intelligent	Perceived Intelligence
Foolish / Sensible	Perceived Intelligence
Dead / Alive	Animacy
Stagnant / Lively	Animacy
Mechanical / Organic	Animacy
Artificial / Lifelike	Animacy
Inert / Interactive	Animacy
Apathetic / Responsive	Animacy

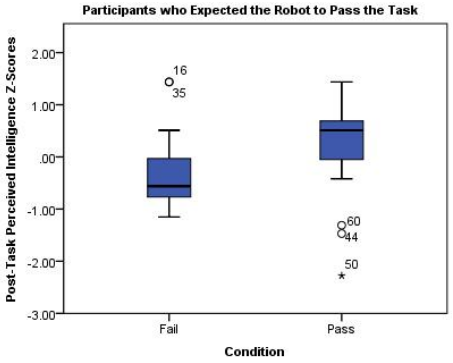


Fig. 2: Perceived Intelligence scores were significantly higher in the Pass condition ($p < 0.01$) when their expectations for the robot were met supporting H3.

C. Perception of Intelligence When Expectations Were Met

We used a Mann-Whitney U test to determine if there were differences in Perceived Intelligence scores between conditions when they answered the mid-task question expecting the robot to pass (N = 46). Distributions of the Perceived Intelligence scores for the Pass and Fail conditions were not similar. Perceived intelligence scores for the Pass condition (mean rank = 28.93) were statistically significantly higher than for the Fail condition (mean rank = 18.07), U = 389.5, Z = 2.751, $p < 0.01$.

D. Perceived Social Intelligence Scales

Analysis of the composite scores for the PSI found statistically significant results for the following scales: RC, PC, AC, PB, II, and SOC.

1) Recognizes Human Cognitions (RC): A Kruskal-Wallis test was used to determine if there were differences in RC scores between participants that watched the robot either pass or fail the false belief task. Distributions of RC scores were not similar for all groups, as assessed by visual inspection of a boxplot. RC scores were significantly different between conditions, $\chi^2 = 20.508, p < 0.001$. The Fail group had a mean rank = 20.95 and the pass group had a mean rank = 41.41.

2) Predicts Human Cognitions (PC): A one-way ANOVA was conducted to determine if the perception of a robot being able to predict the cognition of humans was different depending on condition. There were no outliers for condition,

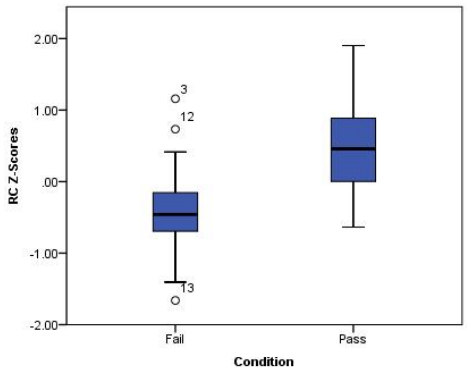


Fig. 3: RC Scores were significantly higher in the Pass condition ($p < 0.001$) supporting H1.

TABLE II: Recognizes Human Cognition (RC) Items

Survey Questions	(On a scale of Strongly Disagree to Strongly Agree)
This robot:	• Can figure out what people think • Knows when people are missing information • Can figure out what people can see • Understands others' perspectives

as assessed by boxplot; data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > 0.05$); and there was homogeneity of variance, as assessed by Levene's test of homogeneity of variance for Condition (p = 0.706). The differences between conditions were statistically significant with the Pass condition (M = 0.352, SD = 0.760) being higher than the Fail condition (M = -0.308, SD = 0.686), F(1,58)= 12.498, p = 0.001.

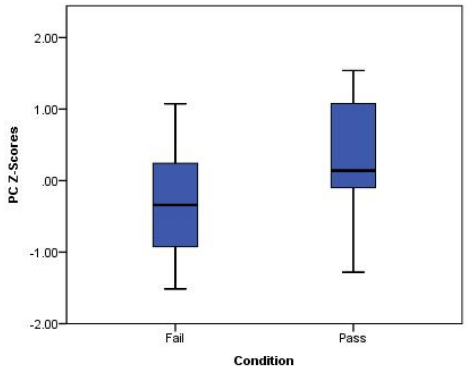


Fig. 4: PC scores in the Pass condition were significantly higher than the Fail condition ($p < 0.001$) supporting H1.

3) Adapts to Human Cognitions (AC): A one-way ANOVA was conducted to determine if the perception of a robot being able to adapt its own behavior based on people's thoughts and beliefs was different depending on condition. There were no outliers, as assessed by boxplot; data were normally distributed for each condition, as assessed by Shapiro-Wilk test ($p > 0.05$) and there was homogeneity

TABLE III: Predicts Human Cognition (PC) Items

Survey Questions	(On a scale of Strongly Disagree to Strongly Agree)
This robot:	• Anticipates others' beliefs • Figures out what people will believe in the future • Knows ahead of time what people will think about certain situations • Anticipates what people will think

of variance, as assessed by Levene's test of homogeneity of variance (p = 0.333). The Pass condition gave significantly higher AC scores (M = 0.245, SD = 0.635) than the Fail condition (M = -0.2147, SD = 0.789), F(1,58) = 6.075, p < 0.05.

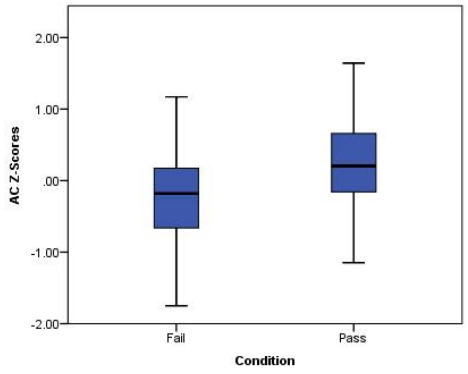


Fig. 5: AC scores were significantly higher ($p < 0.05$) when ToM behavior was demonstrated supporting H1.

TABLE IV: Adapts to Human Cognition (AC) Items

Survey Questions	(On a scale of Strongly Disagree to Strongly Agree)
This robot:	• Adapts its behavior based upon what people around it know • Ignores what people are thinking • Selects appropriate actions once it knows what others think • Knows what to do when people are confused

4) Predicts Human Behavior (PB): A Kruskal-Wallis test was conducted to determine if there were differences in PB scores between conditions. Distributions of PB scores were not similar for all conditions, as assessed by visual inspection of a boxplot. PB scores were statistically significantly different between conditions, $\chi^2 = 4.462, p < 0.05$. The Fail Condition had a mean rank = 26.05 and the Pass Condition had a mean rank = 35.59.

5) Identifies Individuals (II): A one-way ANOVA was conducted to determine if II scores were different depending on condition. There were no outliers for condition, as assessed by boxplot; data was normally distributed for each condition, as assessed by Shapiro-Wilk test ($p > 0.05$); and

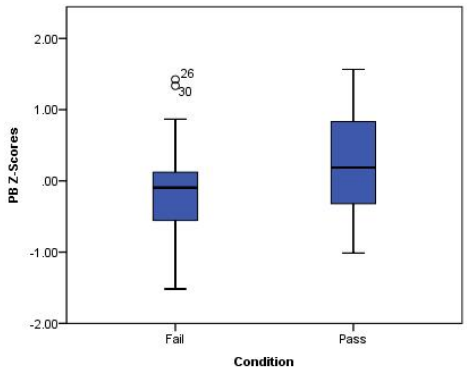


Fig. 6: PB Scores were significantly higher for the Pass condition ($p < 0.05$) supporting H1.

TABLE V: Predicts Human Behavior (PB) Items

Survey Questions	(On a scale of Strongly Disagree to Strongly Agree)
This robot:	• Anticipates people's behavior • Predicts human movements accurately • Has no idea what people are going to do • Knows how people will react to things it does

there was homogeneity of variance, as assessed by Levene's test of homogeneity of variance for condition (p = 0.067). Data are presented as mean ± standard deviation. The differences between conditions was statistically significant, $F(1,58) = 18.506, p < 0.001$.

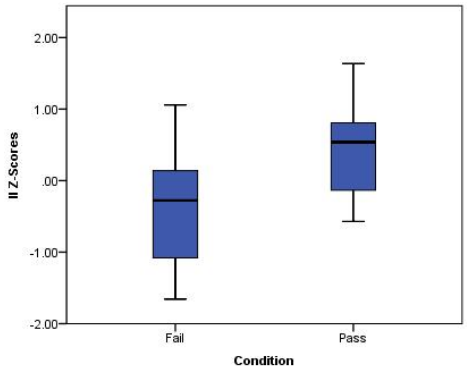


Fig. 7: II Scores were significantly higher in the Pass condition ($p < 0.001$), supporting **H1**.

TABLE VI: Identifies Individuals (II) Items

Survey Questions	(On a scale of Strongly Disagree to Strongly Agree)
This robot:	• Recognizes individual people • Remembers who people are • Cannot tell people apart • Figures out which people know each other

6) Social Competence (SOC): A Kruskal-Wallis H Test was conducted to determine if there were differences in the SOC score between Genders. This scale was rated based on the robot appearing to have strong social skills. Distributions of the SOC scores between the Genders were not similar, as assessed by visual inspection. SOC scores for women (mean rank = 36.23) were statistically significantly higher than for men (mean rank = 26.68), $\chi^2 = 4.311, p < .05$.

TABLE VII: Social Competence (SOC) Items

Survey Questions	(On a scale of Strongly Disagree to Strongly Agree)
This robot:	• Is socially competent • Is socially aware • Is socially clueless • Has strong social skills

V. Discussion

A. Summary of Findings

Our study focused on how watching a robot successfully and unsuccessfully demonstrate a human-like social cognitive ability such as Theory of Mind influenced perception of cognitive and social intelligence and animacy. We intended to show that participants would rate the robot more favorably when it succeeded at the task when compared to the robot who failed the task. This held true for Perceived Intelligence and most of the scales from the PSI, but not for Animacy.

B. Condition and Ratings of Intelligence and Animacy

Our results show that watching a robot exhibit human-like cognitive capacities such as ToM influences whether they perceive the robot as intelligent as well as socially intelligent. Our data supported **H1**, as participants in the condition which watched the robot pass the task gave higher scores for Perceived Intelligence on the GQS than participants who watched the robot fail the task. The main effect for condition on Perceived Intelligence shows that participants had significant decrease in opinion of the robot after watching the video when the robot failed the task, and there was significant increase in how intelligent participants view the robot when it passed the task, supporting **H3** (Fig. 8). Regarding **H2**, we did not find significant differences in conditions for the Animacy GQS scale.

C. Condition and Perceived Social Intelligence

Compared to participants in the **Fail** group, participants in the **Pass** condition scored higher on the robot's ability to recognize human cognition, predict human cognition, adapt to human cognition, predict human behavior, identify individuals, and social competence. This suggests that robots exhibiting Theory of Mind influence how much humans feel a robot is able to predict, adapt to, and detect human cognition and behavior. These results support **H1**, although the scales for recognizing human behavior and adapting to human behavior did not yield significant differences.

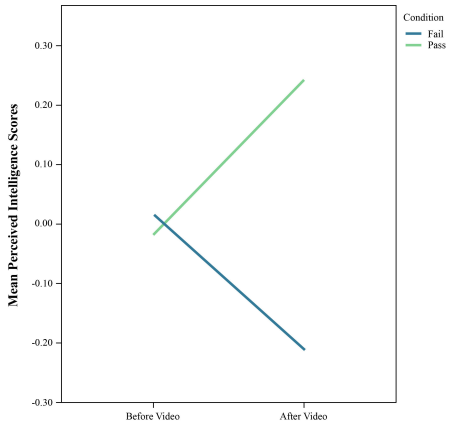


Fig. 8: Mean GQS Perceived Intelligence Z-Scores by Condition before and after viewing the false-belief task

D. Participant Expectations

Regarding **H4**, 45 out of 60 participants expected the robot to perform the task correctly. Furthermore, it appears when these expectations are met they view the robots as more intelligent than those who expected the robot to fail (Fig. 9).

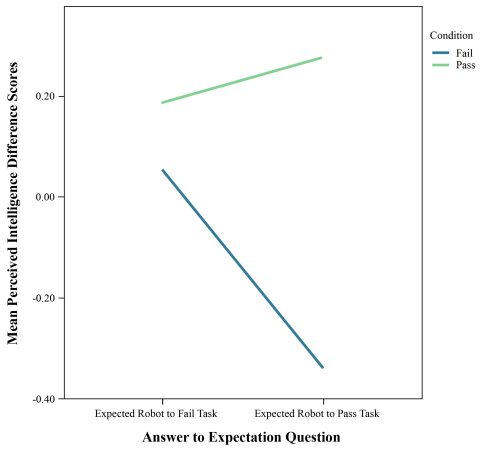


Fig. 9: Mean GQS Perceived Intelligence Z-Scores based on expectations and whether those expectations were met by participant condition

E. Other Findings

Although our experiment did not seek to examine the role of gender on perception we did find that Female participants scored the robot higher for PSI-SOC. Females seemed to see the robot as having stronger social skills than our Male participants. Our participant population was 60% male. It is possible that with a larger female sample size this gender effect may disappear.

F. Limitations and Future Work

This study has some limitations. The task in the video is staged and participants are not interacting directly with

a robot. Embodiment is an aspect that could be incorporated into this study. Embodiment has been shown to have an impact on perception of robots [22] [4]. More specifically, embodiment may play a key role in how humans perceive animacy. This study could be repeated with the robot performing the task in the same room as participants to investigate if there are any changes in animacy scores. Something to consider is also the age of participants. It could simply be that adults don't see a video of a robot as animate regardless of social competence. Future work could examine whether children give higher animacy scores than adults. Extensions for this experiment could also include a comparison of first-order and second-order ToM behavior.

G. Broader Implications

Perception of cognitive and social capabilities in robots influences how humans interact with robots. When behavior defies social norms or displays social-cognitive deficits humans tend to be more critical. Our finding show that robots that do not demonstrate critical developmental concepts, such as Theory of Mind, are perceived as less socially intelligent than robots that do demonstrate such capacity. These attitudes toward robots impact how likeable and beneficial people find their interactions with robots. People are more likely to continue using robots with which they have satisfying interactions.

ACKNOWLEDGMENT

This work was funded by the National Science Foundation (awards IIS-1757929 and IIS-1719027)

REFERENCES

[1] Qualtrics survey platform. *Qualtrics*, 2018.

[2] Marcus P. Adams. Explaining the theory of mind deficit in autism spectrum disorder. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 163(1):233–249, 2013.

[3] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37 – 46, 1985.

[4] C. Bartneck, Takayuki Kanda, O. Mubin, and A. A. Mahmud. The perception of animacy and intelligence based on a robot's embodiment. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*, pages 300–305, Nov 2007.

[5] Christoph Bartneck, Elizabeth Croft, Danai Kulic, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[6] M. Bosia, R. Riccaboni, and S. Poletti. Neurofunctional correlates of theory of mind deficits in schizophrenia. *Current Topics in Medicinal Chemistry*, 12(21):2284–2302, 2012.

[7] Cynthia Breazeal and Brian Scassellati. Robots that imitate humans. *Trends in Cognitive Sciences*, 6(11):481 – 487, 2002.

[8] Kerstin Dautenhahn. The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied Artificial Intelligence*, 12(7-8):573–617, 1998.

[9] Kerstin Dautenhahn. Socially intelligent agents in human primate cutlure. *Agent culture: human-agent interaction in a multicultural world*, pages 45–71, 2004.

[10] Kerstin Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, 2007.

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

[12] David Feil-Seifer and Maja Matarić. Distance-based computational models for facilitating robot interaction with children. *Journal of Human-Robot Interaction*, 1(1):55–77, July 2012.

[13] Scott Forer, Santosh Balajee Banisetty, Logan Yliniemi, Monica Nicolescu, and David Feil-Seifer. Socially-aware navigation using non-linear multi-objective optimization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, October 2018.

[14] Kevin Gold and Brian Scassellati. Using probabilistic reasoning over time to self-recognize. *Robotics and autonomous systems*, 57(4):384–392, 2009.

[15] Carol Gregory, Sinclair Lough, Valerie Stone, Sharon Erzinclioglu, Louise Martin, Simon Baron-Cohen, and John R. Hodges. Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer's disease: theoretical and practical implications. *Brain*, 125(4):752–764, 04 2002.

[16] Laura M Hiatt, Anthony M Harrison, and J Gregory Trafton. Accommodating human variability in human-robot teams through theory of mind. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[17] MC M. Kaptein, P. P. Markopoulos, BER Boris Ruyter, de, and EHL E. Aarts. Two acts of social intelligence : the effects of mimicry and social praise on the evaluation of an artificial agent. *AI & Society : the Journal of Human-Centred Systems and Machine Intelligence*, 26(3):261–273, 2011.

[18] R. Shane Westfall Santosh Balajee Banisetty David Feil-Seifer Kimberly A. Barchard, Leiszle Lapping-Carr. *Perceived Social Intelligence (PSI) Scales Test Manual*, 2018.

[19] Yael Kimhi. Theory of mind abilities and deficits in autism spectrum disorders. *Topics in Language Disorders*, 34(4):329–343, 2014.

[20] Nicole C Krämer, Sabrina Eimler, Astrid von der Pütten, and Sabine Payr. Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction? *Applied Artificial Intelligence*, 25(6):474–502, 2011.

[21] Katie Maras, Imogen Marshall, and Chloe Sands. Mock juror perceptions of credibility and culpability in an autistic defendant. *Journal of Autism and Developmental Disorders*, 49(3):996–1010, Mar 2019.

[22] Ali Mollahosseini, Hojjat Abdollahi, Timothy D. Sweeny, Ron Cole, and Mohammad H. Mahoor. Role of embodiment and presence in human perception of robots' facial cues. *International Journal of Human-Computer Studies*, 116:25 – 39, 2018.

[23] Gina Neff and Peter Nagy. Automation, algorithms, and politics—talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10(0), 2016.

[24] Emma Norling Peter Wallis. The trouble with chatbots: social skills in a social world. *AISB 2005 Convention: Proceedings of the Joint Symposium on Virtual Social Agents: Social Presence Cues for Virtual Humanoids Empathic Interaction with Synthetic Characters Mind Minding Agents*, pages 29–36, 2005.

[25] B. Reeves and C. I. Nass. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press, 1996.

[26] Astrid M Rosenthal-von der Pütten and Jens Hoefinghoff. The more the merrier? effects of humanlike learning abilities on humans' perception and evaluation of a robot. *International Journal of Social Robotics*, 10(4):455–472, 2018.

[27] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.

[28] Lindsay S. Schenkel, William D. Spaulding, and Steven M. Silverstein. Poor premorbid social functioning and theory of mind deficit in schizophrenia: evidence of reduced context processing? *Journal of Psychiatric Research*, 39(5):499–508, 2005.

[29] Teresa Torralva, Christopher M Kipps, John R Hodges, Luke Clark, Tristán Bekinschtein, María Roca, María Lujan Calcagno, and Facundo Manes. The relationship between affective decision-making and theory of mind in the frontal variant of fronto-temporal dementia. *Neuropsychologia*, 45(2):342–349, 2007.

[30] Sophie van der Woerdt and Pim Haselager. When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology*, 54:93 – 100, 2019.

[31] David Williams. Theory of own mind in autism: Evidence of a specific deficit in self-awareness? *Autism*, 14(5):474–494, 2010.