Human Trust After Robot Mistakes: Study of the Effects of Different Forms of Robot Communication

Sean Ye[†1], Glen Neville[†1], Mariah Schrum[†1],
Matthew Gombolay[1], Sonia Chernova[1], and Ayanna Howard[1]

***Abstract*— Collaborative robots that work alongside humans will experience service breakdowns and make mistakes. These robotic failures can cause a degradation of trust between the robot and the community being served. A loss of trust may impact whether a user continues to rely on the robot for assistance. In order to improve the teaming capabilities between humans and robots, forms of communication that aid in developing and maintaining trust need to be investigated. In our study, we identify four forms of communication which dictate the timing of information given and type of initiation used by a robot. We investigate the effect that these forms of communication have on trust with and without robot mistakes during a cooperative task. Participants played a memory task game with the help of a humanoid robot that was designed to make mistakes after a certain amount of time passed. The results showed that participants' trust in the robot was better preserved when that robot offered advice only upon request as opposed to when the robot took initiative to give advice.**

I. INTRODUCTION

As robots continue to become more sophisticated, an increasing number of applications will involve collaboration between humans and robots. Mixed human-robot teams will become more prominent in air travel, hospitals, factories and even consumer homes. As these mixed human-robot teams become more prevalent, implementing effective communication strategies between a robot and its human teammate will be vital to achieving team goals. The interactions within these teams can take many forms, e.g. a pilot asking a robotic co-pilot for advice on how to approach a difficult maneuver, a robot factory worker offering help to a human co-worker, or a person asking a robot butler for assistance in finding a missing item.

Due to the increasing number of human-robot teams, it is important to consider the factors that affect a user's trust of a robot. One of the largest factors affecting continued use of robotic systems and user trust is robotic failure [6]. However, due to the complex nature of the environments and tasks in which robots will have to operate, service breakdowns, failures, mistakes and/or errors are inevitable. These failed or unexpected behaviors can cause a degradation of trust between the robot and the teammates/users. Whether a user continues to look to the robot for assistance may be impacted by this loss of trust. How then should robots communicate effectively to create trust?

The first three authors contributed equally to this work. † indicates corresponding authors.

[1]Georgia Institute of Technology, Atlanta, GA {seancye, gneville, mschrum3, mgombolay3, chernova, ah260@gatech.edu}@gatech.edu

A few studies in the human-robot interaction (HRI) literature have explored speech interactions [19], [20] but the role of who initiates the interaction and when a robot speaks has not been studied in relation to user trust. We seek to understand how different forms of communication (FOC) can impact the gain and degradation of trust in human-robot teams. In our study we identify four common FOCs: unsolicited, solicited, pre-corrective and post-corrective and investigate the effect each has on trust when a robot makes mistakes. In order to test the effects of the FOCs on trust, we created a cooperative Simon Says game that a human participant plays alongside a robot partner. This game simulates a task with high mental strain. During this task, the robot provides assistance in the form of advice via one of the identified FOCs. At predefined times, the robot fails and gives incorrect advice to the participant. Trust measurements are taken periodically to measure changes in user trust over time. We seek to determine how the FOC affects both rise and fall in trust when mistakes occur.

This work provides the following three contributions. First, we identify four forms of communications and distinguish these from other aspects of human-robot interactions. Second, we evaluate the effects of these FOCs on user trust through a collaborative game. We compare the differences on trust when the robot does not make mistakes and after it begins to make mistakes. Third, we compare the effect of the FOC used on how often participants agree with the robot. We found that a humanoid robot giving solicited advice had a significant effect on both trust and agreement as compared to the other FOCs.

II. RELATED WORK

A. Trust and Collaboration

As robot capabilities continue to improve, robots will move into households and workplaces and interact collaboratively with people. Trust will play an important role in the relationship between humans and robots [13], [2]. For instance, Hancock et al. has shown that an individual's trust affects how soon they will intervene as the robot progresses towards task completion [6]. Freedy et al. found that a person will intervene sooner if they trust the robot less [2]. Xu et al. found that a person's behavior changes based on whether their first encounter was with a robot providing incorrect versus correct advice [22]. Furthermore, Steinfeld et al. demonstrated that a user's level of trust also affects the willingness of people to accept information produced by the robot and to follow a robot's advice [18].

Understanding how trust in robotic systems is maintained and lost is important when developing robots that people will continue to interact with. Within the human-robot interaction community, the two largest factors that have been found to impact trust are robot appearance and robot behavior. Findings show that anthropomorphic robots tend to be favored compared to robots that appear more mechanical [7], [1], [5], [4]. Mori's concept of the "uncanny valley" describes the effect of technologies resembling closely but not exactly resembling a human [12]. Researchers have proposed that the degree of human-likeness affects trust and that people very quickly form a 'mental model' of a robot which influences the expectations they have for its behavior [11], [3], [9]. In this way, the appearance and behavior of a robot are crucial in building trust [21].

Another relevant aspect of robot behavior is the nature of speech interactions between humans and robots. Related work has examined how a robot's use of hedge or discourse markers can influence perceptions of the robot [20], [19]. Hedge markers are defined as words people use in sentences to soften the tone. For example, "I'm not an expert, but" could be used to reduce the impact of the sentence. Discourse markers include words such as "so", "well", "right", and "anyways" which connect speech phrases. Researchers have found that markers that affect interaction directness impacted the effectiveness of human-robot communication [19].

B. Trust and Robotic Failures

While it is clear that distrust of robots can be detrimental in human-robot teams, over-trust can have similarly negative consequences. Over-reliance on robots can lead to dangerous situations in safety critical environments where a person trusts a robot when it fails. For example, Robinette et al. conducted a study in which a contrived fire emergency situation was created and participants were encouraged to follow a robot to safety [16]. Many participants followed the robot despite obvious faults in the robot's guidance. Further work by Robinette et al. defined two situations where people can over-trust robots: misjudging the abilities or intent of the robot and misjudging the risk involved [14]. In these works, failures or service breakdowns are an important part of influencing how much a person continues to trust a robot. Lee et al. found that strategies mitigating the effects of these types of failures include apologies, compensation, or providing more options to the user [8]. Other researchers have also found that the specific timing of when robots apologize or promise to do better impact the repair of trust [15]. While specific strategies have been shown to maintain trust after robotic mistakes, how users perceive different types of robot communication with respect to trust is still an open area of research.

III. Forms of Communication

In this work, we examine how forms of robot communication affect user trust. Our work differs from prior studies of communication and trust which looked at robot features such as hedge and discourse markers or other aforementioned trust repair strategies. In our study, the FOC is defined as a combination of who begins the interaction and when the robot provides information. The timeline of the progression of each FOC is visualized in Fig. 1. Here, we define the four FOCs.

1) **Unsolicited:** The robot aids the user *preemptively*, whether or not the user has prompted the robot.
2) **Solicited:** The robot aids the user only when the user has *directly requested* help from the robot.
3) **Pre-Corrective:** The robot aids the user only when the *user makes a mistake*, at the moment of the mistake and *before completion of the task*.
4) **Post-Corrective:** The robot aids the user only when the *user makes a mistake*, *after completion of the task*.

FOCs are unique from other previously studied trust repair strategies. Our selected FOCs specifically capture the differences in initiation strategy, i.e. whether the user initiates the interaction or the robot initiates the interaction. They also investigate the timing of interaction, i.e. whether the robot corrects the user at the time of failure or after completion of the task. FOCs are different from robot features such as hedge markers, appearance, or directness because FOCs change whether the user receives information or not. For example, if a person does not ask for information, this is a different form of communication than when the robot gives you information unprompted. However in both these cases, the robot could use the same hedge markers, have the same appearance, and have the same level of directness.

IV. Methodology

A. Materials

We conducted our study using a Pepper robot from SoftBank Robotics. Pepper is an animated humanoid robot with text to speech, speech recognition, and computer vision capabilities.

We developed a Simon Says game, pictured in Fig 2. In this version of the Simon Says game, a sequence of colored lights (white, yellow, blue and green buttons) are flashed in predefined sequences. After the sequence is shown, the participant must copy the sequence by pressing the buttons in the exact same order as flashed. Participants can change their answer by pressing a "reset" button (located on the top left of the box). Thus, users could change their answer if Pepper provided advice.

One sequence of the game had a length of eight color flashes. This was selected based on prior psychological experiments that have shown individuals can recall, on average, about seven items from short-term memory [10]. The game involved ten interaction rounds. In each round, there were three sequences followed by a 14-point questionnaire. Participants therefore completed 30 sequences of the game.

The Simon Says game was chosen because it allowed us to measure trust in discrete time blocks. It is also a game that is

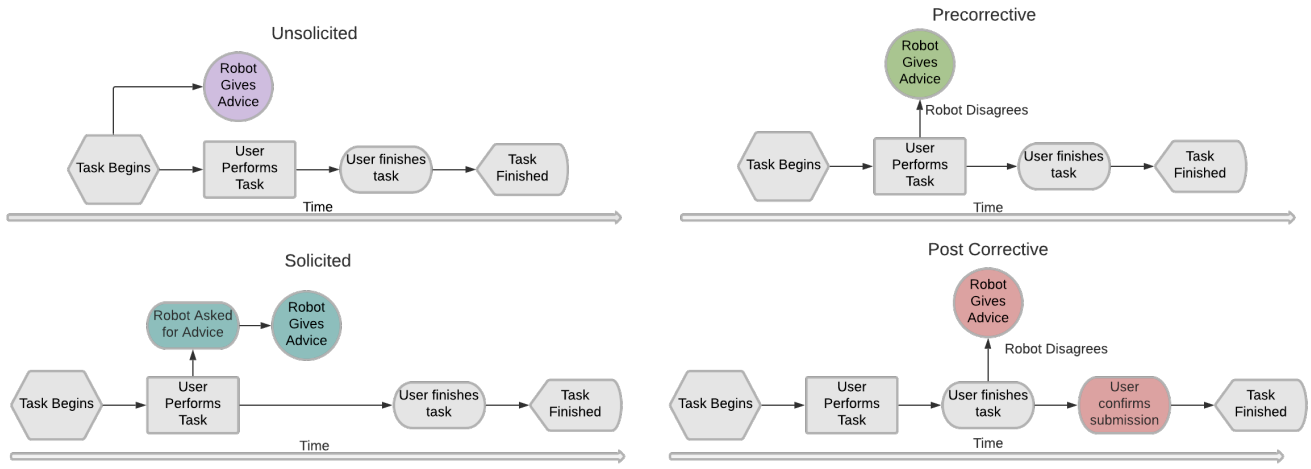

Fig. 1. Diagrammatic Representation of Forms of Communication: This diagram shows when a robot gives advice with respect to a task that a user is completing. The colored/darkened nodes are locations where the robot provides a different type of advice, either by initiating differently or at a different timing.

Fig. 2. Simon Says Game

easy to understand so participants can be quickly briefed and do the experiment in a short time. The feedback for correct or incorrect sequences is immediate so participants are able to quickly judge the advice received. Finally, each sequence length takes a relatively short time to complete which allows many interactions with Pepper over a short period of time.

B. Participants

We recruited 44 participants (Male=28, Female=16) who were students and staff between the ages of 18 and 64 to participate in this experiment. The mean age was 22.6 with a 7.5 standard deviation. Participants were randomly assigned to each of the four FOC conditions. This experiment was conducted with the approval of the university's Institutional Review Board (IRB) and all subjects signed approved consent forms.

C. Experimental Design

This study was a 1 x 4 between subject design to test the effects of each FOC on trust. Each participant only experienced one FOC when interacting with the robot.

In order to obtain an understanding of how different types of communication affect trust we developed a collaborative trust exercise. In this exercise, participants played a Simon-Says game (described in Section IV-A). While the participants played the game, Pepper assisted the participants by giving advice. However, this advice was not always correct. Participants were told that Pepper shared the same goal of completing the task. In our experimental design, Pepper provided the correct sequences to the user during the first four rounds of the game. Four rounds without failure was the minimal amount needed for maximum trust to be reached as participants' trust scores plateaud and reached a steady state. This was found in a pilot study. In the next six rounds, Pepper provided an incorrect sequence once in each set of three, for an effective failure rate of one-third. Our expectation was that trust would rise from the baseline for the first four correct rounds then fall at a certain rate for the remaining incorrect rounds. This hypothesis was tested with the measures described in the Metrics section.

The type of advice Pepper gave took the form of one of the FOC's. In the unsolicited condition, Pepper provided the user with the sequence unprompted every time. In the solicited condition, the participant must ask Pepper for advice directly. Only when asked would Pepper provide advice. In the pre-corrective condition, Pepper would only provide advice when the user made a "mistake" when entering the sequence based on Pepper's knowledge of the sequence. This means that Pepper would not only correct the participant when the participant was wrong, but also when Pepper had incorrect knowledge of the sequence and thus thought the participant was wrong (even if the participant was right). Pepper provided this advice immediately when the user inputted a color that differed from Pepper's belief. In the post-corrective condition, Pepper again provided advice only when the participant disagreed with Pepper's knowledge of the sequence. However, Pepper did not interrupt the participant and only provided the advice once the participant had finished entering the entire sequence. Again, Pepper would provide this advice when Pepper's knowledge of the sequence disagreed with the input of the participant.

The experiment proceeded as follows: The participant was introduced to Pepper, and the game was explained. The experimental setup is shown in Fig. 3. The participant was told that the study was being conducting to test Pepper's new computer vision algorithm. Before the experiment began, participants played the game for one round without Pepper to familiarize themselves with the game mechanics. In the solicited case, they were also shown how to ask Pepper for help.

The experiment then began and the participant took the 40-question baseline trust survey. The participant then played the game for one round, i.e. three back-to-back sequences of flashing lights in which the participant entered his or her own input after each sequence. At this point, the participant took the 14-question trust survey. This was repeated for ten rounds, resulting in a total of 30 sequences and ten 14-question surveys. The sequences were the same for each participants. Additionally, the sequences when Pepper gives incorrect advice were also the same across all conditions. After the 10 rounds, the participant took the final 40 question survey. This timeline of events is pictured in Fig 4. Each round took approximately two to three minutes totalling an experiment time of around 30 minutes.

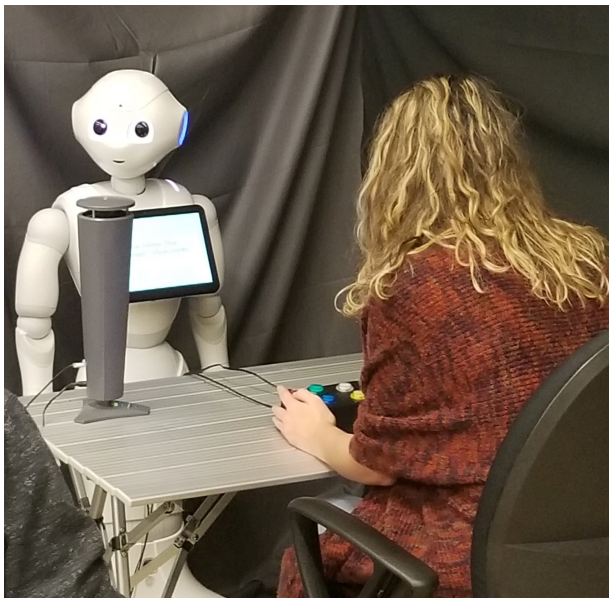

Fig. 3. Experimental Setup

V. Hypotheses

We hypothesize about the relationship between the FOCs and their impact on trust. As a subjective metric of trust we utilize a standard 14-point survey. As an objective metric, we measure the amount of agreement between the participant's response and the advice of the robot. We predict overall trust differences in Hypothesis 1 and rate of trust rise and fall differences in Hypothesis 2.

Hypothesis 1-A: *Users will have greater trust in the robot when the robot gives solicited advice as compared to the other FOCs* because solicited advice is less intrusive than the other forms of communication. Furthermore, a person is likely to be more forgiving when the advice was prompted for by the user.

Hypothesis 1-B: *Users will have greater trust in the robot when the robot gives post-corrective advice as compared to pre-corrective advice (Post-Corrective > Pre-Corrective).* The main difference between pre- and post- corrective advice is that pre-corrective interrupts the participant which may be viewed as rude or annoying by the participant. Therefore, we believed that post-corrective would result in higher trust scores than pre-corrective.

Hypothesis 1-C: *Users will have greater trust in the robot when the robot gives unsolicited advice as compared to pre-corrective advice (Unsolicited > Pre-Corrective).* Similar to the previous hypothesis, pre-corrective is more interruptive than unsolicited so we hypothesized that users would trust a robot using unsolicited advice more.

Hypothesis 2-A: *Users' trust will decrease at a faster rate when the robot gives pre-corrective advice compared to the other FOCs during robot failure.* We hypothesized that pre-corrective advice would affect trust more negatively than the other conditions because the robot is interrupting the user with false information. Based on our pilot study, users claimed that this distraction greatly impacted how well they could complete the task.

Hypothesis 2-B: *Users' trust will increase at a faster rate when the robot gives pre-corrective advice compared to the other FOCs when the robot is correct.* We hypothesized that users would prefer the interruptive nature of pre-corrective when the robot is correct because it notifies the user immediately when a mistake occurs and allows them to correct the mistake before completing the task.

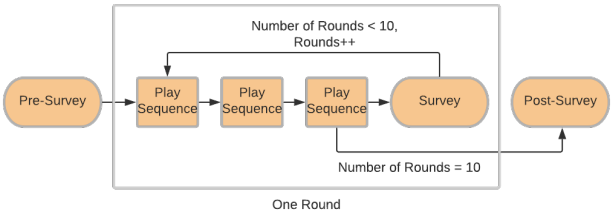

Fig. 4. Summary of experiment flow

VI. Metrics

A. Explicit Metrics

Trust metrics consisted of a 14-point and 40-point survey described in Schaefer [17]. We used these metrics as they have been thoroughly tested and are explicitly designed for measuring trust over time. The 40-point survey was given to the participant prior to the experiment to obtain a baseline

trust score and the 14-point survey was used after each round of the game giving 11 aggregate trust measurements. The internal consistency between the two surveys can be found in Schaefer [17]. The 14-point survey consisted of the following questions rated at 10% intervals from 0% to 100%: What % of the time will Pepper function correctly, be dependable, be reliable, be unresponsive, be predictable, act consistently, malfunction, provide feedback, meet the needs of the mission/task, provide appropriate information, have errors, communicate with people, perform exactly as instructed, and follow directions?

B. Implicit Metrics

In order to further measure participant trust, two implicit measures were recorded. These were *participant agreement with Pepper* and *number of sequences correct*. Agreement with Pepper was recorded as a percentage and was used as a measure of trust alongside the explicit measures mentioned above. The number of sequences correct was collected to verify whether or not the FOC also had no impact on performance.

VII. Results

In this section, we report the results of the study and the statistical findings. We establish statistical significance at the $\alpha = 0.05$ level. Trust scores were split into two groups: before Pepper begins failing (rounds 1-4) and after Pepper begins failing (rounds 5-10). The first trust score taken before the game began was used as a baseline. A summary of the adjusted trust score over all game rounds is shown in Fig. 5. Adjusted trust scores were created by subtracting the baseline trust score from every participant's subsequent trust score.

We conducted a linear mixed effects analysis of the relationship between trust and form of communication. As fixed effects, we used the FOC's interaction with failures before and after along with their baseline trust, age, race, gender, and education. The baseline trust was determined by the pre-survey participants took before interacting with Pepper. As random effects, we had intercepts for each participant. Mistake rounds were encoded three different ways within the model. In the first encoding, rounds 1-4 were grouped together as no mistake rounds and rounds 5-10 were grouped together as mistake rounds. In the second encoding, the specific slopes of trust rise in rounds 1-4 were modeled. In the third encoding, the trust fall in rounds 5-10 were modeled. The residuals of the model were checked to be normally distributed by a visual inspection of the Q-Q plot.

Main effects were found for the baseline trust score, trust scores before and after failure, rate of trust increased, and rate of trust decrease. No main effects for the form of communication were found. However, interaction effects were found between the form of communication and the difference in trust scores between no mistake and mistake rounds. Interaction effects were statistically different between solicited and pre-corrective $F(1,484) = 7.39, p = 0.009$, solicited and post-corrective $F(1,484) = 6.086, p = 0.002$, and unsolicited and post-corrective, $F(1,484) = 5.567, p = 0.033$. Fig. 6 shows these interactions.

One significant interaction effect between the slope of trust rise and form of communication was found. Unsolicited and post-corrective had statistically different slopes ($F(1,484) = 3.895, p = 0.016$) between rounds one through four. No other interaction effects were found in both the trust rise and trust fall slopes.

A one-way ANOVA ($F(3,38), p \leq 0.001$) determined that the form of communication significantly impacted the percentage of times a participant agreed with Pepper. Levene's test, the Shapiro-Wilk test, and the Bartlett test were used for checking outliers, normality of residuals, and homogeneity of variances respectively. A Tukey post-hoc test revealed statistically significantly higher ratio of agreements between the solicited case ($M = 0.92$, $SD = 0.14$) than the pre-corrective ($M = 0.68$, $SD = 0.12$, $p < 0.001$) and post-corrective ($M = 0.66$, $SD = 0.10$, $p < 0.001$). There was no statistically significant results between unsolicited ($M = 0.79$, $SD = 0.10$) and the other groups. Fig. 7 shows these results.

Finally, a one-way ANOVA determined no significant differences were found between the FOC groups on the percent of sequences a participant got correct.

VIII. Discussion

The results indicate that the form of communication (FOC) used by Pepper affects the difference in trust loss once mistakes occur. In other words, the change in trust before and after Pepper makes mistakes is impacted by the form of communication. Here we will analyze each significant interaction effect. We found evidence to support Hypothesis 1-A but we did not find any significant differences to support Hypothesis 2.

A. Positive Effect of Solicited Advice on Trust

Trust decreased significantly less after Pepper began making mistakes in the solicited case compared to both the pre-corrective and post-corrective case. While the interaction effect with unsolicited did not show statistical significance, the difference in trends can be seen in Fig. 6. Participants thus maintain trust much better in the solicited case. This is not the same as saying solicited advice loses trust slower than the other cases because we found no interaction effects for the slope during the mistake rounds. Instead, we compare the average trust scores during the correct rounds versus the average trust scores during the mistake rounds.

This significant interaction effect corresponds to our first hypothesis that users would trust Pepper when offered solicited advice over other forms of advice. We believe this to be the case because when Pepper begins making mistakes, she doesn't initiate the interaction and interrupt or distract the participant who is completing the sequence. Pepper takes around 2 or 3 seconds to speak the sequence which many participants commented caused them to forget the sequence due to the interruption. The solicited case allows users to prompt for advice only when they want to. In addition, if Pepper is wrong when a user asks for advice, perhaps they

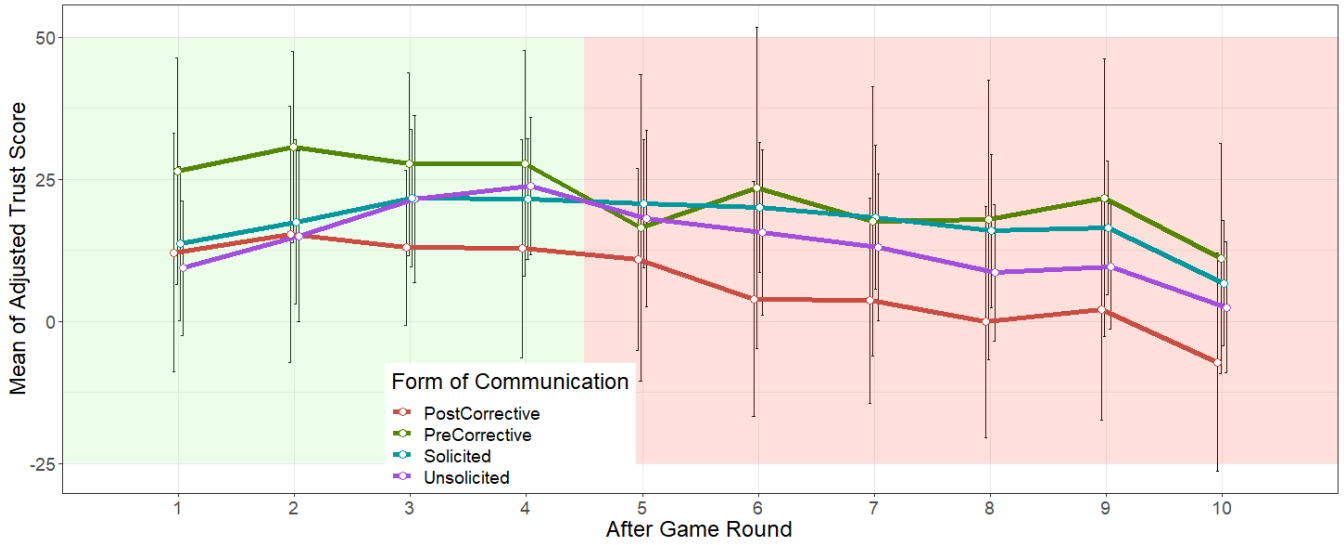

Fig. 5. Mean of adjusted trust score over all game rounds. Error bars represent standard deviation. Green region (rounds 1-4) represents Pepper giving completely accurate advice, red region (rounds 5-10) represents Pepper giving mistakes in her advice

Fig. 6. Interaction effects of forms of communication before and after Pepper fails. Statistical levels are indicated here: $p^* \leq 0.05$, $p^{**} \leq 0.01$

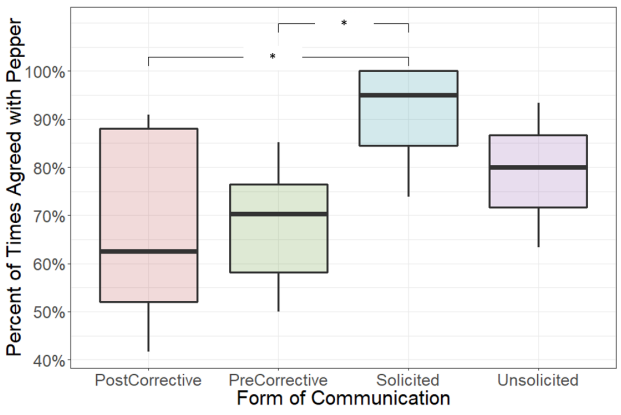

Fig. 7. Comparison of percent of times participants agreed with Pepper for each condition. The significant levels are indicated here: $p^* \leq 0.001$

are more forgiving because they didn't know the sequence themselves. However, our study seeks only to demonstrate which FOC affects trust the most, not the reasons behind it.

B. Agreement with Pepper

The results show that the frequency with which participants agreed with the answer provided by Pepper was affected by the FOC. We found statistically significant differences between the solicited condition and pre-corrective and post-corrective conditions. Since solicited communication was initiated by the participant it was more likely that they were unsure and in need of advice. Perhaps participants were more willing to accept Peppers answer due to this factor. On the other hand post-corrective required the users to have already formulated an answer to the task and therefore participants are less inclined to change their answer.

IX. Conclusion

Our results indicate that the form of communication can positively or negatively impact a user's perception of a robot. We found that using solicited advice decreased the degradation in trust users had when a robot started making mistakes in our task. Further work can be done on identifying which form of communication is most suitable for different tasks. Our experiment had relatively low risks for when the robot made a mistake which may have affected which form of communication participants preferred. For example, in a high risk scenario, participants may prefer a robot that interrupts through unsolicited, pre-corrective, or post-corrective advice if the information is crucial. Another future study, could analyze the relationship between the form of communication and how often mistakes are revealed to the user. In this study, the solicited case could have the potential to obfuscate the true error rate of the robot.

Social robots that are designed to serve the public must be viewed as trustworthy sources of information and assistance

if they are going to be actively used. Unfortunately, as with all systems, mistakes are inevitable. These robot mistakes can potentially lessen the public's trust of these robotic systems. Therefore it is important to create robots that build and maintain trust with the people they serve. This study demonstrated that the ways in which robots provide information and how they initiate an interaction can affect how trustworthy they are perceived by users.

ACKNOWLEDGMENT

This work was partially supported by funding from the National Science Foundation under Award #1849101, the Air Force Office of Sponsored Research under Award #FA9550-17-1-0017 and the NSF Accessibility, Rehabilitation and Movement Science Fellowship under Grant #1545287.

REFERENCES

[1] Julia Fink. Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. 2012.

[2] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. Measurement of trust in human-robot collaboration. *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems, CTS*, (August 2018):106–114, 2007.

[3] J. Goetz, S. Kiesler, and A. Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pages 55–60, Nov 2003.

[4] Matthew Gombolay, Anna Bair, Cindy Huang, and Julie Shah. Computational design of mixed-initiative human-robot teaming that considers human factors Situational awareness, workload, and workflow preferences. *International Journal of Robotics Research (IJRR)*, 36(5-7):597–617, 2017.

[5] Matthew Gombolay, Xi Jessie Yang, Brad Hayes, Nicole Seo, Zixi Liu, Samir Wadhwania, Tania Yu, Neel Shah, Toni Golen, and Julie Shah. Robotic assistance in coordination of patient care. In *Proceedings of Robotics: Science and Systems (RSS)*, Ann Arbor, MI, U.S.A., June 20-22 2016.

[6] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527, 2011.

[7] Pamela J. Hinds, Teresa L. Roberts, and Hank Jones. Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction*, 19(1-2):151–181, 2004.

[8] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (May 2016):203–210, 2010.

[9] Sau-lai Lee, Ivy Yee-man Lau, S. Kiesler, and Chi-Yue Chiu. Human mental models of humanoid robots. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 2767–2772, April 2005.

[10] George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

[11] Takashi Minato, Michihiro Shimada, Hiroshi Ishiguro, and Shoji Itakura. Development of an android robot for studying human-robot interaction. In Bob Orchard, Chunsheng Yang, and Moonis Ali, editors, *Innovations in Applied Artificial Intelligence*, pages 424–434, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[12] M Mori. Bukimi no tani [the uncanny valley]. *Energy*, 7:33–35, 01 1970.

[13] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian Jentsch. Building Appropriate Trust in Human-Robot Teams Mental Models : Building Blocks of Trust. *AAAI Spring Symposium*, 2013.

[14] Paul Robinette, Ayanna Howard, and Alan R Wagner. *Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?* Springer US, 2017.

[15] Paul Robinette, Ayanna M Howard, and Alan R Wagner. Timing is Key for Robot Trust Repair. *7th International Conference on Social Robotics (ICSR 2015)*, Oct. 2015.

[16] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. Overtrust of robots in emergency evacuation scenarios. *ACM/IEEE International Conference on Human-Robot Interaction*, 2016-April:101–108, 2016.

[17] Kristin E. Schaefer. *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"*, pages 191–218. Springer US, Boston, MA, 2016.

[18] Aaron Steinfeld, Terrence Fong, Moffett Field, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common Metrics for Human-Robot Interaction. (1), 2006.

[19] Megan Strait, Cody Canning, and Matthias Scheutz. Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. *Conference on Human-robot interaction*, pages 479–486, 2014.

[20] Cristen Torrey, Susan R. Fussell, and Sara Kiesler. How a robot should give advice. *ACM/IEEE International Conference on Human-Robot Interaction*, pages 275–282, 2013.

[21] Michael L Walters, Dag S Syrdal, and Kerstin Dautenhahn. Avoiding the uncanny valley : robot appearance , personality and consistency of behavior in an attention-seeking home scenario for a robot companion. pages 159–178, 2008.

[22] Jin Xu and Ayanna Howard. The Impact of First Impressions on Human- Robot Trust during Problem-Solving Scenarios. *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, pages 435–441, 2018.