The Effect of Robot Disagreeableness on Trust

Neha Harpale
University of Nevada, Reno
nharpale@nevada.unr.edu

Gautham Yerroju
University of Nevada, Reno
gyerroju@nevada.unr.edu

Tung Dang
University of Nevada, Reno
tung.dang@nevada.unr.edu

ABSTRACT

In this paper, the effect of robot disagreeableness on trust of human is examined through a set of experiments with a humanoid robot as well as a survey based on Godspeed questionnaires. The paper focuses on two main conditions where the robot disagrees with the human in warranted and unwarranted manners. The agree condition is used as a baseline. Following intuition and evidence from human-human interaction research, two hypotheses related to those conditions are proposed. The first one is that a robot with warranted disagreements tends to gain more trust than a robot that always agrees. Similarly, the second is that a robot with warranted disagreements should gain more trust than one with unwarranted disagreements. A short quiz game is utilized to build trust between a human and a robot. Then an one-shot investment game is performed to measure it quantitatively. Collected data suggest that both hypotheses seem valid, but statistical results reveal that there is no strong evidence to support the proposed hypotheses.

1 INTRODUCTION

Robots have been increasingly used in many areas like assistive care, search and rescue, bomb disposal, navigation, etc. [4], where safety of individuals is at stake. With highly sophisticated robots, there will also soon be situations where robots can become capable of strategies which are not obvious to a human operator or collaborator. Trust is an important factor to consider when introducing robots into such domains, especially as it has been shown that trust can often determine the overall acceptance and usage of a system [5, 8]. In do or die scenarios robots might have to disagree with humans for optimal outcomes. Disagreements are usually beneficial to solving problems, as they bring diversity to ideas and generally lead to better performance [4]. It is therefore beneficial for robots in these situations to be able to disagree with a human operator or collaborator. Our goal in this study is to investigate the effect of disagreeableness of a robot on the trust a human ascribes to it. We also study how trust varies between warranted disagreeableness and unwarranted disagreeableness.

2 BACKGROUND

Hancock et al. have performed a meta-analysis of factors affecting trust in human-robot interaction [6], where they measured the effects of human, robot, and environmental factors on perceived trust. They concluded that a robot's performance and attributes were the largest contributors to the development of trust in HRI, with environmental factors playing only a moderate role. Though

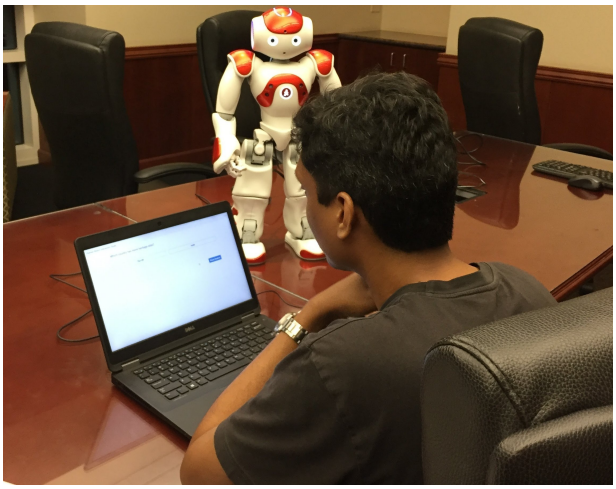

Figure 1: Nao robot and a participant during the experiment. The participant answers the quiz, plays investment game and fills the questionnaire on a laptop computer, while the NAO stands to the side in autonomous life mode. The Nao speaks its opinion about each of the participant's choices as soon as they select an option on the computer.

they studied the effects of a human's propensity to trust on the development of trust, they did not study disagreeableness as a factor. Salem et al. have investigated how erroneous robot behavior may influence trust in a robot [10]. This paper focused on the effects of errors performed by robots on trust. Participants were asked by the robot to perform unusual tasks to assess their trust in the robot. They found that the robot's performance did not substantially influence participants' decisions to not comply with its requests, but whether the requested task and its effects were revocable or irrevocable. The work done by Mathur and Reichling in [7] uses a game-theory based methodology (one-shot investment game) to assess social trustworthiness of robots. However, their work focused on the effects of the Uncanny Valley [11] and subtle facial expressions on trust. They found that robots in the Uncanny Valley were indeed perceived as less trustworthy. The effect of disagreeableness on robot likeability and perceived similarity has been studied by Takayama et al. [12], but their focus was on studying the effect of separation of voice from an embodied robot on agreement. By simulating a desert survival task, they found that people felt more positive toward a disagreeing robot whose voice came from a separate control box, where as people preferred the voice coming from the robot for agreement. Their study was focused on mitigating the negative perception of disagreeing robots. As far as we are aware, disagreeableness as a factor affecting trust was not the focus of

Conference, 2018, USA

ACM ISBN 123456789/18/02…$0.00
https://doi.org/10.114

these studies. We are interested to know if warrantedness can be used as a factor to affect perceived trust.

3 METHODOLOGY

Our aim is to study the effects of warranted and unwarranted disagreeableness on human trust in a robot. In our two-step experiment, the robot establishes different types of trust with a human subject in the first step (trust-building) and that trust is then measured in the second step (evaluation). This section will describe the first and second steps, introduce the conditions (the manipulated variable), then discuss the hypotheses and how they relate to the conditions. Then, the experimental setup is detailed.

3.1 Step 1: Trust Building Exercise

The first step is used to build human trust in the robot for different types of disagreeableness or agreeableness. The participant engages with the robot in completing a quiz. Each question in the quiz has two possible responses. When the participant chooses one response, the robot speaks its opinion by either agreeing or disagreeing with the participant, i.e. by either speaking an opinion in support of the participant's response, or in support of the other response. Then, the participant can proceed to view the correct response, accompanied by a justification statement. Then he or she can proceed to the next question.

The questions we picked are based on obscure statistical facts and we assume that participants in general are not aware of the correct responses. For each question, we created two sets of statements for each response. One set is the robot opinion supporting each response, and the other set is a justification for each response shown when the correct answer is revealed to the participant. A list of questions used is shown in Table 2.

The correct responses to these questions is irrelevant to us because we control how often the participant is wrong (which is 40% of the time). Depending on whether we want the participant to be right or wrong, either of the two responses can be declared as the correct answer, supported by their respective justification statement. For example, one of the questions is "Which country consumes more rice? Bangladesh or Indonesia?". If the participant picks Bangladesh and we want him/her to be wrong, we declare that the participant is wrong, and use the justification statement for Indonesia: "The Indonesian government takes a lot of effort to keep rice production high to meet the country's high rice demand.".

Similarly, we control whether the robot agrees or disagrees with the participant by using different opinion statements. For the above example, if the robot wants to agree with the participant, it says: "It should be Bangladesh, because its cuisine is mainly rice based.". The robot also adds a prefix to convey its agreement, for example "I agree...", or "I don't think so..." before stating the opinion statements. There are three prefix phrases each for agreement and disagreement, and each prefix is attached to the robot opinion randomly, to avoid monotony in its responses. Also note that the robot uses qualifiers like "I think..." or "The right answer might be..." to ensure that participants do not confuse the robot's opinion for knowledgeability.

3.2 Step 2: One-shot Investment Game

Even though the investment game [3] has been studied thoroughly in other research fields such as economics, it is still questionable approach in the case of HRI research. Furthermore, to the best of our knowledge, there is no well-established method to measure the trust in the HRI literature, so we were very wary when using this game in our project. Thus, to ensure the reliability of the results collected from this game, we introduced a direct question related to trust in our post-experiment survey to further evaluate the statistical results shown in the section 4. The game is explained more clearly as following: The participant needs to invest some (imaginary) money in the robot, and the robot would return anywhere from half to twice the participant's investment amount. After the quiz, the participants are given these instructions and told to invest anywhere from $1 to $100 in the robot based on how much they trust it. Based on the amount they choose to invest, we get a measure of trust they have in the robot. After the participants make an investment, they are told that the robot will take some time to make decision about how much to return, and are asked to complete a questionnaire in the meanwhile.

We do not immediately show the participants how much the NAO returns, because what matters to us is the amount the participants chose to invest in the robot, and we do not want the robot's return value to affect the post-survey questionnaire. However, to give the participants a sense of completion, we have the robot return a random value between 100% to 130% of the investment value, which the robot says after participants complete the questionnaires. We also use this randomly generated number to determine optional cash prizes for the participants with the top three return values.

3.3 Conditions

For the experiment the manipulated variables, that is the type of agreement or disagreement of the robot are:

- **Agree (A):** Robot always agrees
- **Disagree Warranted (DW):** Robot disagrees (with good reason)
- **Disagree Unwarranted (DU):** Robot disagrees (without good reason)

For the condition A (Agree), the robot always agrees with whatever response the participant picks. For condition DW (Disagree Warranted), the robot only disagrees with the participant when the participant's answer is wrong. For condition DU (Disagree Unwarranted), the robot disagrees when the participant's answer is correct. Note that the robot always disagrees 40% of the time for all conditions.

3.4 Hypotheses

Based on the conditions, we hypothesize the following:

- **H1** Agreeable robots are trusted less often than robots with warranted disagreeableness
- **H2** Robots with warranted disagreeableness are trusted more often than robots with unwarranted disagreeableness

Both H1 and H2 are based on evidence in human-human interactions. H1 for instance alludes to people giving less credibility to the words of a sycophant. This is also supported by the findings in [12],

where robots which disagreed with participants were found to be more persuasive than robots which always agreed with the them. H2 is based on human-human interaction, where a person with warranted disagreement might be considered more trustworthy than a person who disagrees without a good reason. This can also apply to to human-robot interactions as evidenced in *The Media Equation* by Reeves and Nass [9], where plenty of research is discussed on how people treat computers and new media as real people.

3.5 Questionnaires

After the participants make the investment, they are administered a subset of the Godspeed Questionnaire [2] to obtain subjective measures of their perceptions of the robot. The Godspeed questionnaire is widely used to gauge perceptions of anthropomorphism, animacy, likeability, perceived intelligence and perceived safety. In this paper, we used a shorten version of the original Godspeed questionnaires (e.g. eliminated irrelevant questions such as perceived safety). We also asked the participants general comments about explaining their choice of investment amount, whether they trust the robot and why, and finally other comments they might have about the robot.

3.6 Experimental Setup

The experiment took place in a meeting room in the Knowledge Center building in the University of Nevada, Reno campus, over roughly 12 hours from 10am to 10pm. The participants' age ranged from 18 to 33, with over 60% between the age of 18 and 23. We invited the participants to take part in the experiment by word of mouth. Conducting the experiment for each participant took approximately 12 minutes on average.

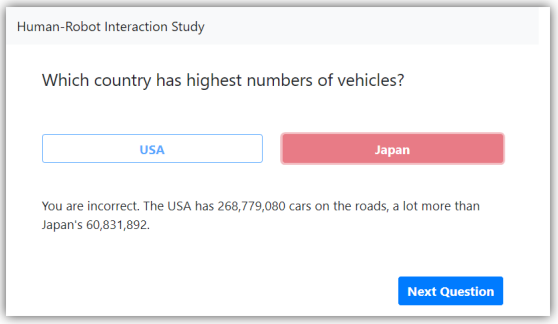

Figure 2: Designed user interface for the quiz. The participant clicks on a response, then the robot speaks its opinion. A "Show Answer" button appears at this point, which the participant can click to reveal the correct answer. The state after clicking the "Show Answer" button is shown in in this figure.

Upon entering the room, the participants first sign the consent form, after which they are seated in front of a laptop computer and the NAO robot set in autonomous life mode (shown in Figure 1). On the computer screen, they are shown instructions for the quiz step, after which a sample question is shown so that they get a sense of how the quiz will operate. We also stood by and explained to them verbally if they seemed like they were confused. The participants then went through all the ten quiz questions. An example of this is shown if Figure 2. For each question after the participant makes a choice, all interaction on the screen is disabled, so the participant can focus on the robot. After the ten questions, a summary screen is displayed, which shows a list of questions each accompanied by either a green tick to indicate correct response or a red X to indicate an incorrect response. For each question, there is also an icon of NAO with either the word "Agreed" or "Disagreed" beside it to indicate robot's agreement with the participant.

When the participants click "Proceed", instructions to the Investment Game are displayed. After they read the instructions (and/or we explain to them verbally), they can proceed to a screen which has a slider with a range of $1 and $100, and the participants can drag the slider to choose an investment value within that range. The slider's default position is halfway, at $50. This is shown in Figure 3. After they make a decision on the value they would like to invest, they click the "Invest" button. At this point a spinner is shown, indicating that there is a wait here. We tell them that the NAO will take some time to make a decision on how much to return, and direct them to a different window to fill the post-experiment questionnaire. The spinner is designed to be displayed for 30 seconds, after which the button changes to "View Result". After the participants complete the questionnaire, they are asked to check back on the result from the robot and we direct them to the original window, where the "View Result" button appears. On clicking that button, The NAO speaks out the amount that is returned to them: "Here's ... dollars. Thank you for investing!". Their original investment and NAO's return are also shown on the screen.

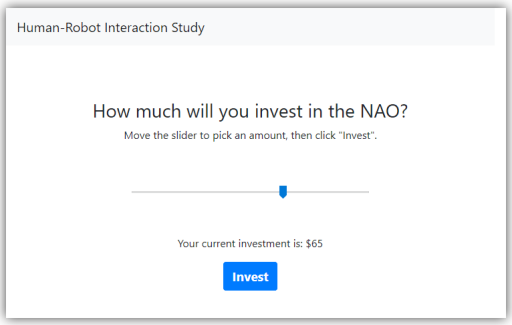

Figure 3: User interface for the investment game. The participant needs to drag the slider to make a choice between $1 (at the far left) and $100 (at the far right), then click the "Invest" button. At this point, a loading spinner will be displayed and the participant is told that the NAO is thinking about a how much money to return. After 30 seconds, the button changes to "View Result", and clicking on it will make the NAO speak how much it returns, and displays that amount on the screen.

Finally, the participants are debriefed about the experiment. They are told about the $5 cash prize for the top three participants with the highest return amount by the robot from the investment game, and are given instructions on how to enroll if they choose to do

so. Then, we explained to them that we controlled how often they were wrong, and that their responses to the quiz questions are not necessarily right or wrong as was displayed. They are asked to take some snacks and thanked for their participation.

3.7 Implementation

The quiz application and the investment game were designed as a web application. The server was built using the Python framework Flask and the frontend was designed using Bootstrap and jQuery. The Python version of the NAO's official SDK, PyNaoQi was used to control the NAO. Each time the the initial web page is loaded, a drop down box allows the experimenters to select a condition (A/DU/DW). Then, an increasing ID starting from 1001 is assigned to the current run. As the participants make choices for each quiz question, a message is sent to the Flask server which makes the NAO speak its opinion, then shows a button on the UI to allow the participant to view the correct response. At the end of the investment game when the participants click "View Result", the NAO is instructed by the server to speak the returned value. The return value is generated using a random number generator. Instead of using the RNG provided by Python, we used the RNG exposed by the underlying operating system for more reliable randomness. At the end of the investment game, the web application saves the participant's ID, condition, their investment value, the return value, and their responses to questions in a CSV file. The implementation of this system is available on Github [1].

A configuration file is used to control exactly for which of the ten questions the participant would be correct (the "correct answer" mask), and for which of the questions the robot would agree with the participant ("robot agreement" mask). These "correct answer" and "robot agreement" masks are explicitly defined for each of the three conditions, and are the same for every participant, to ensure consistency.

4 EXPERIMENTAL RESULTS

We were able to run the experiment for 27 participants, but one of them had to be discarded because our software malfunctioned and results for that participant were not saved, leaving 26 participants. We randomly assigned conditions to participants initially, and towards the end, tried to balance the numbers between the three conditions, and to have equal number male and female participants. We had 8 participants for the Agree condition, 8 participants for the Disagree Warranted condition, and 10 participants for the Disagree Unwarranted condition.

4.1 Statistical Analysis of Investment Game

To evaluate the effect of robot disagreeableness on trust of human, three independent datasets on each condition (A: Agree, DU: Disagree Unwarranted, DW: Disagree Warranted) were collected to assess quantitatively the two proposed hypotheses. More specifically, we assumed that the amount of monetary investment in the second game correlates positively to the trust of human on the robot. Therefore, in this section, the invested money from the investment game is considered as a primary variable to assess the trust. However, due to the known complexity of measuring the trust, the feedback provided from participants in the survey is carefully examined as a secondary check to verify the validity of the acquired measurements.

The testing hypotheses are two following comparisons: H1) the NAO robot programmed under condition DW should gain more trust than one in condition A, and similarly H2) the NAO in condition DW should also gain more trust than NAO in DU. There are many statistical methods to perform such analysis on the acquired datasets. A simple, yet naive, approach could be comparing the mean or median values to quickly understand the trend of the data. As shown in Figure 4, the medians of the data support our hypotheses which supposed that the amount of investment $Invesment(A) < Invesment(DW)$ and $Invesment(DU) < Invesment(DW)$.

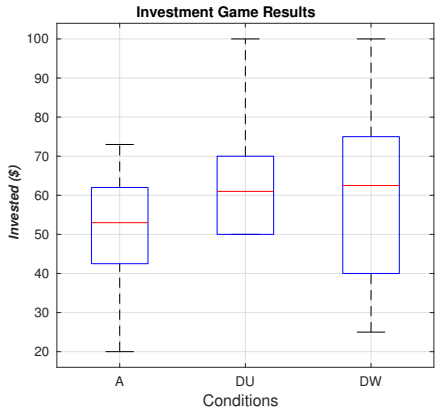

Figure 4: Box plot of investment amount from participants for the three conditions.

To further evaluate the results statistically, few statistical techniques were applied to compute the $p-$value, then compare it to the significance level 0.05 as a common practice in statistics. In general, most of such methods assume that the measurements have normal distribution which usually holds with a large sample size. An alternative for small dataset is the t-test, but this method is very sensitive to outliers. In our experiment, the sample size is small, less than 30, and possibly contains outliers as pointed out from feedback of participants in the post-experiment survey; hence, non-parametric methods that do not assume the normality are more suitable in this case. Essentially, such approaches perform the test on the rank of each data point rather than its raw value. For instance, the Wilcoxon method is often applied to one- and two-group cases. Moreover, for tests with more than 2 groups, another technique is the Kruskal-Wallis analysis. Since there are three independent sets on three conditions in the experiment, either Wilcoxon or Kruskal-Wallis test could be utilized to reveal the differences between testing conditions.

4.1.1 Wilcoxon Test. We want to test conditions A vs. DW (this corresponds to H1), as well as conditions DU vs. DW (corresponds to H2). The null hypothesis is considered as no difference between two sets. The alternative hypothesis here is one-sided, which means that the median of condition 1 is less than the median of condition 2. To support this alternative hypothesis, the $p-$value must be smaller

than 0.05. Shown below are results using Wilcoxon test for two hypotheses using the MATLAB statistics toolbox:

- A vs. DW: $p = 0.17$
- DU vs. DW: $p = 0.6087$

The large p–value shown above means that there is no statistical difference between conditions; or equivalently, we failed to reject the null hypothesis.

4.1.2 Kruskal-Wallis Test. In this method, we can evaluate 3 sets at the same time. The null hypothesis here is that the collected dataset came from the same distribution. In contrast, the alternative hypothesis is that they did not come from the same distribution. The p–value provided from this test is 0.3440, which implies that there is no strong evidence to reject the null hypothesis.

Overall, statistical results computed as above do not show significant evidence to support our hypotheses. Nevertheless, these results could be explained due to excessive outliers in our datasets which could be drawn from comments of participants on reasons why they chose that amount of money (e.g. '*[I invested] 100, since I am a risk seeker person*'; '*[I invested] 65 [since] 65 is my lucky number*'). Hence, this violates our main assumption in this test which assumes positive correlation of the investment and the trust. Another possible reason is the small sample size of our datasets which is not large enough to filter out noisy data points. For these reasons, to further verify the proposed hypotheses, we chose an alternative to evaluate our hypotheses using the secondary variable extracted from the post-experiment survey in which we asked the participants directly '*Do you trust the robot that you played with, and would you please explain briefly main reasons for your choice?*'. We manually encoded the answers into one of three categories: 'Y: yes', 'N: no', and 'NS: not sure'. The results are shown in Figure 5. Overall, the trust in conditions DW and DU are higher than one in the condition A. Also, in the DW case, not a single participant claimed to distrust the robot. Therefore, we would conclude here the designed experiments including the Q&A and investment games present reasonable outcomes that align with the proposed hypotheses, but they fail to fully support those ideas. In future work, one should consider getting more data and also refine the experiments to create a stronger link between two games.

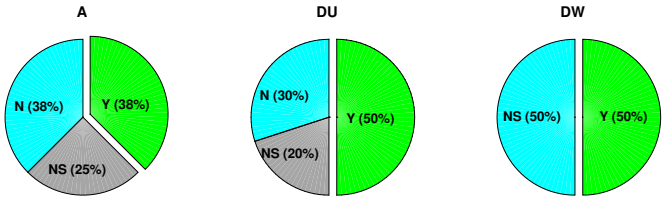

Figure 5: Pie plot of trust from participants for the three conditions.

4.2 Analysis of Godspeed Questionnaire

We collected results for four scales in the Godspeed Questionnaire: Anthromorphism, Animacy, Likeability and Perceived Intelligence. The box plots for different conditions of the GQS scales are shown in Figure 6. All the scales for all the conditions passed consistency tests, with Cronbach's Alpha greater than 0.8, except condition DU in Anthropomorphism, and conditions DU and DW in perceived intelligence. As we checked again in the experiment log, in the first 3 experiments, we forgot to turn on the Autonomous Life mode of the NAO robot which might lead to very low rates in the Anthropomorphism criterion. The Cronbach's Alpha values are shown in Table 1.

Table 1: Cronbach Alpha values for the Godspeed questionnaire.

Measurements	Cond	Mean	STD	Cronbach's Alpha
Anthropomorphism	A	3.2812	0.7	0.8140
	DU	2.8750	0.5922	0.5703
	DW	3.1562	1.2602	0.9370
Animacy	A	3.5938	1.1721	0.9488
	DU	3.4250	0.7076	0.8100
	DW	3.5625	0.9039	0.8270
Likeability	A	4.3750	0.5285	0.9079
	DU	3.8800	0.6339	0.9098
	DW	3.8800	0.6339	0.9098
Perceived Intelligence	A	3.8750	0.6944	0.8889
	DU	3.2000	0.8563	0.5606
	DW	4.0625	0.6232	2.8e-16

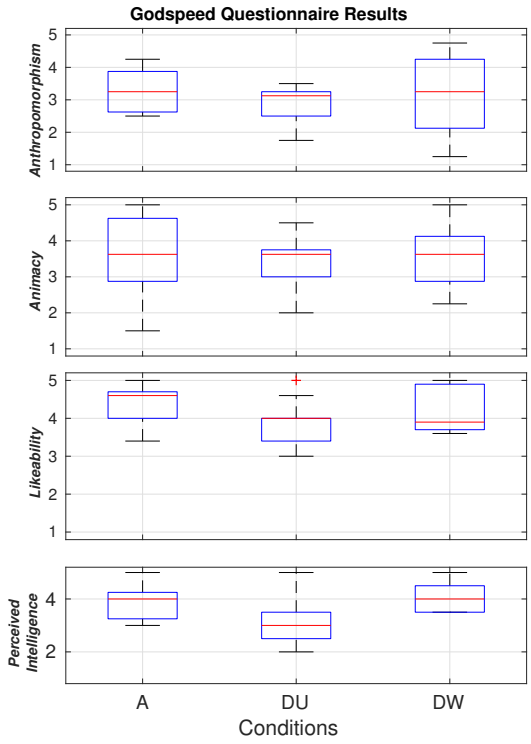

Figure 6: Box plots of the Godspeed Questionnaire for the three conditions, across the four scales of animacy, anthropomorphism, likeability and perceived intelligence.

The main idea of using animacy, likeability, anthropomorphism and perceived intelligence is as indicators of whether the participants believed they were interacting with an intelligent agent; and more importantly, the participants should enjoy the interaction with the robot. Box plots of these scales (6) seem to show that the participants did indeed rate them higher, with medians ranging between 3 to 5 on the scales.

We also wanted to see whether our manipulated variable had any effect on the GQS scales. For likeability, the box plot indicates that it would be highest for condition A and lowest for the condition DU. This is because we assumed that participants liked the robot when it agreed with them (condition A), disliked the robot when it disagreed with them (condition DW), and disliked it even more when it disagreed with them and was also wrong (condition DU), which seems to be the case. We also notice that perceived intelligence is lowest for Disagree Unwarranted, and is on the higher side for Agree and Disagree Warranted. This is expected, because in the DU condition, every time the robot disagreed with the participant, it was also wrong, so it would give the impression of lower intelligence.

4.3 Other Comments from the Survey

Beyond observations described in the above parts, we got overall very positive and constructive feedback from the participants. Many of them expressed explicitly their interest in the robot and our project, e.g *'I really like this project and think this robot could be a useful tool'*. We also found that that there were a few participants who did not fully understand the experiment procedure; for example '*I am confused whether it was correct or not because the computer sometimes gave me different answers*'. This is either because we did not explain clearly or they did not read the instruction carefully. So after first 10 tests, we decided spend more time to explain the experiment to the participants.

5 CONCLUSION

In conclusion, we designed an experiment to assess the effect on different types of agreeableness on trust in a robot. The first step of the experiment is a trust-building exercise which establishes trust with the Agree, Disagree Unwarranted, and Disagree Warranted conditions. The second step is the investment game which is meant to measure the level of trust the participants have on the robot. This is followed by a questionnaire, which includes four scales from the Godspeed questionnaire and some general questions about why participants chose to trust/distrust the robot. From statistical analysis of the investment made by participants, we found that there was no significant effect of the three condition on the investment value. This was attributed to too many outliers in the data, where the investment amount did not truly reflect participant trust in the robot (as evidenced by the comments) and our small sample size. However, inspection of the box plots indicates support for our hypotheses. Results from the Godspeed questionnaire indicated that most participants considered the robot as intelligent (with medians ranging from 3-5 on the scales), and Likeability was lowest for the Disagree Unwarranted condition. This is because from participants' point of view, in addition to disagreeing with them, the robot also turned out to be wrong multiple times. In addition, we also open-sourced our source code including a web-based framework to communicate with the NAO robot, and a simple interface to carry out quickly quiz-like experiments with the NAO.

While our analysis did not yield significant results statistically, the raw numbers do seem to indicate support for our hypotheses. In future work, we would address two items. First, we would try to make the trust evaluation step more representative of participants' trust in the robot, because based on comments we failed to do that effectively. One way we can do this is to make the trust evaluation step similar to the trust building step. For example, the evaluation step could also be a quiz, but this time, after the robot speaks its opinion, the participants can change their answers. By counting how often they changed their answers after a robot suggested a different answer, we can measure their level of trust. Second, we would try to increase the sample size significantly so that we can achieve statistically significant results to either prove or disprove our hypotheses more definitively.

ACKNOWLEDGMENTS

The authors would like to thank Dr. David Feil-Seifer and students in his lab for providing inputs, discussion, as well as support through preparation of this paper.

REFERENCES

[1] 2018. Web App for HRI Project. (2018). https://github.com/GauthamYerroju/hri_spring_2018

[2] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[3] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10, 1 (1995), 122 – 142. https://doi.org/10.1006/game.1995.1027

[4] J. Casper and R. R. Murphy. 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 33, 3 (June 2003), 367–385. https://doi.org/10.1109/TSMCB.2003.811794

[5] Victoria Groom and Clifford Nass. 2007. Can robots be teammates? Benchmarks in human-robot teams. 8 (10 2007), 483–500.

[6] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* 53, 5 (2011), 517–527. https://doi.org/10.1177/0018720811417254 arXiv:https://doi.org/10.1177/0018720811417254 PMID: 22046724.

[7] M. B. Mathur and D. B. Reichling. 2009. An uncanny game of trust: Social trustworthiness of robots inferred from subtle anthropomorphic facial cues. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 313–314. https://doi.org/10.1145/1514095.1514192

[8] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253. https://doi.org/10.1518/001872097778543886 arXiv:https://doi.org/10.1518/001872097778543886

[9] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press, New York, NY, USA.

[10] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. 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 (HRI '15)*. ACM, New York, NY, USA, 141–148. https://doi.org/10.1145/2696454.2696497

[11] Jun'ichiro Seyama and Ruth S. Nagayama. 2007. The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. *Presence: Teleoper. Virtual Environ.* 16, 4 (Aug. 2007), 337–351. https://doi.org/10.1162/pres.16.4.337

[12] Leila Takayama, Victoria Groom, and Clifford Nass. 2009. I'm Sorry, Dave: I'm Afraid I Won'T Do That: Social Aspects of Human-agent Conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2099–2108. https://doi.org/10.1145/1518701.1519021

Table 2: List of questions used in the quiz. Question 0 is a sample question used to demonstrate the UI to the participant and is not counted in the final score. A prefix phrase for agreement/disagreement (e.g. "I agree with you", "I don't think so") is attached before each robot response, chosen randomly from 3 possible prefixes each for agreement and disagreement to avoid monotony.

No.	Question	Choices	Opinion statements	Justification statements
0	Which country has more heritage sites?	The UK	I think it's The UK, because the British Empire ruled around half of the world at one point.	The UK has over 30 heritage sites, including the Stone Henge, the English Lake District and Durham Castle.
		India	I think it's India, because it is a melting pot of cultures since a long time.	India has over 35 heritage sites, including the Red Fort, Ellora Caves, and the Taj Mahal.
1	Which country has highest road network size?	USA	I think it's USA, as it generally has very good infrastructure.	The total road length of USA is 6,722,347 and that of China is 4,696,300.
		China	I think it's China, as it as the largest population.	USA's highway roads are 77,017 km long, which is less than China's 131,000 km highways.
2	Which country consumes more rice?	Bangladesh	It should be Bangladesh, because its cuisine is mainly rice based.	Bangladesh consumes more rice than Indonesia. It consumed 25% more rice than Indonesia in 2014 alone.
		Indonesia	It should be Indonesia, as its government spends a lot to maintain high rice production to meet demands.	The Indonesian government goes to a lot of effort to keep rice production high to meet the country's high rice demand.
3	Who has the most Oscar nominations?	Katharine Hepburn	Katharine Hepburn should be correct, because she has acted longer.	Katharine Hepburn had 66 years of acting career and Meryl Streep only has 46 years.
		Meryl Streep	Meryl Streep should be correct, because she was nominated more times.	Meryl Streep has the highest Oscar nominations, 21, compared to Katharine Hepburn, who has 12.
4	Which country has the longest coastline?	Russia	It's Russia, because it looks larger on a map.	Russia shares a coastline with 3 oceans and 4 seas.
		Philippines	It's Philippines, because it's an island country.	Philippines is an island country with a very long coastline of 38,287 km, longer than Russia's coastline of 37,653.
5	Which country has highest numbers of vehicles?	USA	It should be USA, because they are more affordable with financing being common.	The USA has 268,779,080 cars on the roads, a lot more than Japan's 60,831,892.
		Japan	It should be Japan, because its automotive industries are one of the largest industries in the world.	Japan's automotive industry is one of the largest of all types of industries in the world, producing an average of more than 89,554,219 cars per year.
6	Which country has largest army?	India	I think it's India, because it has the larger population.	Despite North Korea having the largest military institution in the world, India's population is much larger, and its army is larger than North Korea's at 1,395,100.
		North Korea	I think it's North Korea, because it has the largest military institution in the world.	Despite India's much larger population, North Korea has the largest military institution in the world and a significant population is enrolled in its army, which is 1,190,000 strong.
7	Which country has more languages?	Mexico	It should be Mexico, because it is a larger country and is diverse.	Mexico is one of the most culturally diverse countries in world, where 68 languages are spoken.
		Papua New Guinea	It should be Papua New Guinea, because it was ruled by external powers for more than 60 years.	Having been ruled by many countries, Papua New Guinea is home to 830 languages.

Table 2 continued from previous page

8	Which country has more gold?	India	India should be correct, as it's the largest consumer of the precious metal in the world.	In addition to having one of the largest populations, India is also the largest consumer of precious metals in woe world, gold included.
		Japan	Japan should be correct, as it has the world's third largest economy.	One of the world's strongest economies, Japan is also the eighth largest hoarder of gold in the world.
9	Which country has the largest mall?	Thailand	I think it's Thailand, as it has a more bustling economy.	Thailand has the largest mall, which covers an area of 132,606 sq. ft.
		Philippines	I think it's Philippines, as it has a more bustling economy.	Philippines has the largest mall, which covers an area of 5,970,000 sq. ft.
10	Which is more popular chocolate brand?	KitKat	KitKat is correct, because it became more popular after it launched.	KitKat is sold over 80% Of the world, a brand more easily recognized by Cadbury.
		Cadbury	Cadbury is correct, because it was established a hundred years before KitKat.	Cadbury had a 100 year head start over KitKat and a much more diverse product portfolio, known all across the world.