May 3rd, 12:00 AM - 12:00 AM

Investigating Trust and Trust Recovery in Human-Robot Interactions

Abigail L. Thomson
Augustana College - Rock Island

Investigating Trust and Trust Recovery in Human-Robot Interactions

Abby Thomson

Abstract: As artificial intelligence and robotics continue to advance and be used in increasingly different functions and situations, it is important to look at how these new technologies will be implemented. An important factor in how a new resource will be used is how much it is trusted. This experiment was conducted to examine people's trust in a robotic assistant when completing a task, how mistakes affect this trust, and if the levels of trust exhibited with a robot assistant were significantly different than if the assistant were human. Each participant was asked to watch a computer simulation of the three-cup monte or shell game where the assistant would give advice, and the participant could choose to follow, ignore, or go against the advice. The hypothesis was that participants would have higher levels of trust in the robotic assistant than the human, but mistakes made by the robot would have a larger impact on trust levels. The study found that while there was not a significant difference between the overall levels of trust between the robotic assistant and the human one, mistakes did have a significantly larger impact on the short-term trust levels for the robotic assistant versus the human.

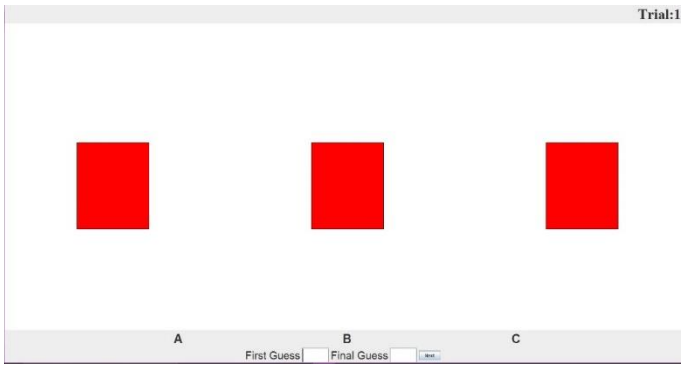
Background: The question of trust in an important and timely one. As AI (Artificial Intelligence) and robotics continue to advance, they are being used in increasingly different ways and situations. With any new technology, it is important to understand how people will interact and use it, an important factor in how people use a technology is how much they trust it. In a paper titled "Network Operations: Developing Trust in Human and Computer Agents" Dzindoley, Beck and Pierce describe two possible types of inappropriate reliance people can exhibit towards AI [4]. The first type of inappropriate reliance is misuse, depending on AI too much and not

noticing or being critical of mistakes. The other type is disuse, not accepting assistance from an AI even when it is clearly more capable. To examine people's use of an AI, and see if people misuse or disuse it, Dzindoley, Beck and Pierce conducted an experiment where participants where shown a series of images, some with a camouflaged solider and some without. Participants were asked to identify whether there was a soldier present in the image, then received feedback from either an AI or previous participants. Participants stated they felt the AI had the highest accuracy, but when given the choice between depending on the AI or on themselves, the majority choose to depend on themselves, even when they and been explicitly told the AI had a significantly higher accuracy rate.

There are a number of factors that can affect people's trust in a robot or AI. Li, Rong and Thatcher found in their study, "Does Technology Trust Substitute Interpersonal Trust? Examining Technology Trust's Impact on Individual Decision-Making," that people were more trusting when the robot was visibly managed by a person [7]. Mistakes made by a robot have been shown to affect people's response to it. Studies have varied on whether it is a positive or negative effect. In Alem, Eyssel, Rohlfing, Koop and Joublin's paper, "To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability," they described the result of their experiment that found that small mistakes made by a robot during a conversation increases people's positive feelings towards the robot [8]. On the other hand, Salem, Lakatos, Amirabdollahian and Dautenhahn describe in their article "Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust" an experiment where they found that mistakes made by a robot in a social setting caused an increase in negative feelings towards the robot, and made people less likely to follow its directions [9].

There are other aspects that affect how people view others' trustworthiness and how quickly they regain trust when lost. When interacting with other people, a study by Haselhuhn, Schweitzer and Wood found that how a person views others' moral character affects how quickly they regain trust [6]. A study by Haselhuhn, Kennedy, Kray, Zant and Schweitzen found that women were more forgiving and quicker to regain trust than men [5]. While both of these studies focused on human interaction, these same aspects could affect how people interact with robots.

Procedure: The participants in this study were 62 undergraduate college students. The experiment was advertised to the student population through emails, posters and by word of mouth; based on this, participants self-selected to be involved in the experiment. No particular group was targeted, but Computer Science majors and minors were excluded from trials with the robotic assistant due to the fact that they would most likely not be fooled by deceptions made about the robot's capabilities. Participants were told that the purpose of the experiment was to compare human, trained human and robotic visual tracking ability. This deception was done to help prevent potential participant bias. Before the trials, the participants answered a short questionnaire asking their age, gender, major and rate on a scale of 1 to 5 how likely they would use assistance from an AI in four different situations. Then the assistant would introduce themselves. In half of the trials the assistant was a robot named TAVI (Tracking Animations Very Intelligently). In the rest of the trials the assistant was a human confederate.



Screenshot of simulation

The basic structure is that the participants play twenty-four rounds of a computer simulation of the Three Cup Monte game, also known as the Shell Game. In the game, there are three identical cups, one of which has a ball hidden underneath it. The cups then rapidly switch positions. The player's goal is to guess where the cup with the ball under it is at the end, and each time their final guess is correct they receive one entry into a drawing for a gift card. For the first twenty rounds, the participant would guess which cup the ball was under, be given advice from an assistant, then the participant would make their final guess. After the twentieth round, the participant was asked to choose whether he or she wanted to answer the final four rounds on his or her own, without any help from the assistant or have the assistant answer for him or her.

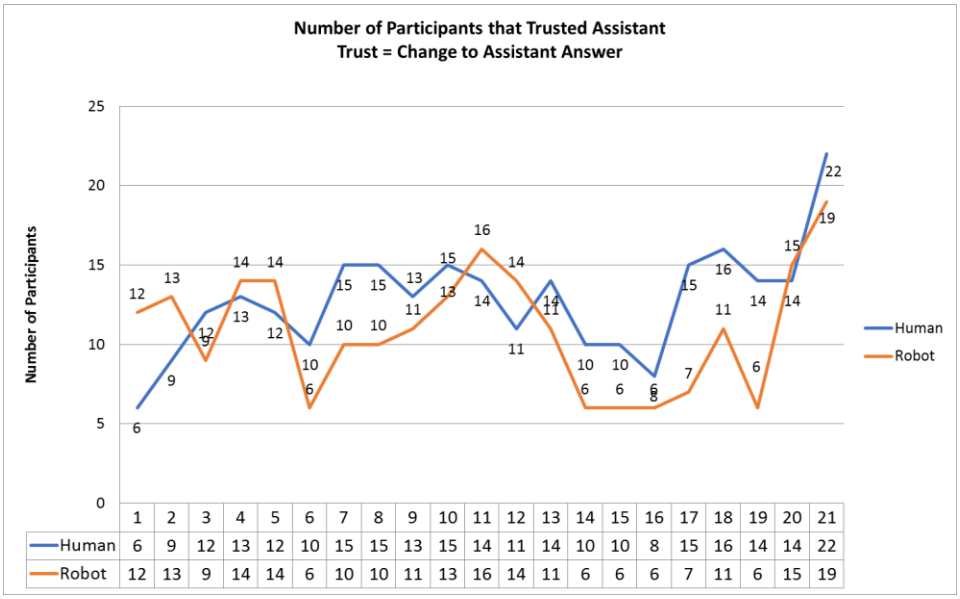
There were multiple layers of deception in this experiment besides concealing the true purpose of the study. Participants were lead to assume that if they watched the simulation closely, they would be able to succeed in determining the correct answer. In reality, the changes in cup positions were just a sequence of random switches that had no bearing on what the "correct" answer for that round would be. This was done so that each participant's independent accuracy would be about 33%, significantly lower than the 70% accuracy of the assistant. The order of correct answers was preset and both the human and robotic assistant were following a set script. The order of correct and incorrect guesses is as follows (where C is a correct answer and I is an incorrect one);

TAVI, built using Lego Mindstorm EV3 kit

C C C C I C C C C C I I I I I C C C C C. With this comes another aspect of deception, TAVI was not in fact able to track anything. Instead TAVI merely moves its "head" back and forth and speaks a letter when the touch sensor, held by the experimenter, is pressed. This gave TAVI the

illusion of watching and responding to what was happening on the computer screen. During each trial, the computer simulation collected the assistant type (human or robot), participants' first and final guesses for each round, and whether the participant decided to depend on the assistant or answer on their own for the last four rounds.

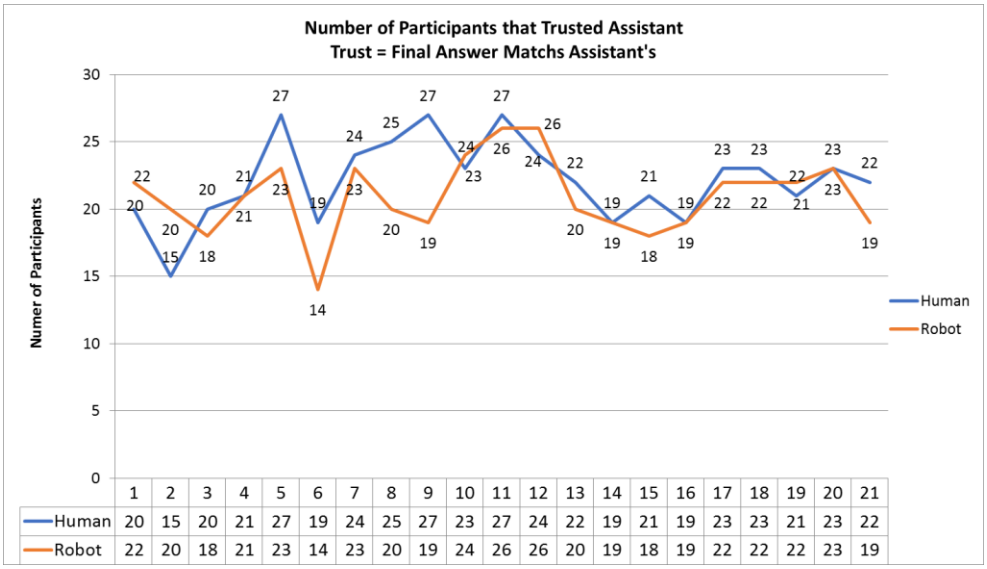
Results: The data collected during the simulation was used to help address the question of whether there was a difference between how participants trusted a human or robotic assistant. In particular, examining how or if participants changed their answer after hearing the assistant's guess. To compare trust, it is important to first define what is meant by trust, and how it would be measured. Two different definitions of trust were used to help create a more complete understanding of people's actions. In the first, a participant was defined as trusting the assistant if they changed their final guess to match the assistant's answer. So if the participant guessed **A** first, then the assistant said **B** and the participant put **B** for their final answer, than the trial would be categorized as the participant trusted the assistant for that round. If in the same example the participant guessed **B** for their first guess than the trial would be categorized as neutral, neither trust nor distrust. If the participant's final guess did not match the



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Human	6	9	12	13	12	10	15	15	13	15	14	11	14	10	10	8	15	16	14	14	22
Robot	12	13	9	14	14	6	10	10	11	13	16	14	11	6	6	6	7	11	6	15	19

Figure 1

assistant's answer then the trial would then be distrust. By using this definition, there can be reasonable confidence that if a trial is categorized a trust, the participant did depend on the assistant when making their final guess. A drawback is when comparing the number of people who trusted the assistant versus distrusted there is a significant number of participants who are left out. To address this concern, a second definition of trust was also used. In this second definition, the participant's first guess was not taken into account, only his or her final answer.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Human	20	15	20	21	27	19	24	25	27	23	27	24	22	19	21	19	23	23	21	23	22
Robot	22	20	18	21	23	14	23	20	19	24	26	26	20	19	18	19	22	22	22	23	19

Figure 2

A participant was defined as trusting the assistant if their final guess matched the assistant's guess. This definition has the advantage of having every participant represented as either trust or distrust for each trial. There is no way to differentiate between participants who stayed with their answer because they trusted their own abilities and those who trusted in the assistant. To balance the advantages and disadvantages of each, both definitions were used and compared.

The above graphs show the number of participants who trusted the assistant in each round. There is a clear difference in the numbers between the two different definitions of trust, but the overall trends are consistent between the two. The final data point on each graph is the same, as it represents the number of people who decided to have the assistant answer for them in the last rounds. There is no difference in this point between the two definitions because the

question required the participant to choose between two options, trust or distrust, so there was no neutral option.

To better understand the trends in the data and how assistant accuracy affects people's trust, the trials where divided into groups. Trials were grouped together based on the assistants' accuracy on the previous trial. So, the first five trials were grouped together because the assistant had been correct on the previous trial until trial five.

	Trust = Change to Assistant Answer									
	Mean	Stnd Deviation	Mean	Stnd Deviation	Mean	Stnd Deviation	Mean	Stnd Deviation	Mean	Stnd Deviation
Trials	1 thru 5		7 thru 11		12 thru 16		17 thru 20		Overall	
Human	10.4	2.58	14.4	0.8	10.6	1.96	14.75	0.83	12.76	2.67
Robot	12.4	1.85	12	2.28	8.6	3.32	9.75	3.56	10.90	3.31

Figure 3

	Trust = Final Answer Matchs Assistant's									
	Mean	Stnd Deviation	Mean	Stnd Deviation	Mean	Stnd Deviation	Mean	Stnd Deviation	Mean	Stnd Deviation
Trials	1 thru 5		7 thru 11		12 thru 16		17 thru 20		Overall	
Human	20.6	3.83	25.2	1.6	21	1.90	22.5	0.87	22.14	3.02
Robot	20.8	1.72	22.4	2.58	20.4	2.87	22.25	0.43	20.95	2.80

Figure 4

The questions asked to participants at the beginning of the study are listed in Appendix A. The average scores of these questions were as follows.

Question	1	2	3	4
Mean Score	2.84	2.00	2.81	2.82

Figure 4

Discussion: This experiment did not find a significant difference in the overall average number of people who trusted the human assistant (Figure 3: 12.76, Figure 4: 22.14) versus the robotic assistant (Figure 3: 10.90, Figure 4: 20.95). This does not support the hypothesis that people would trust the robot assistant more than the human. Another fact that does not support the hypothesis is that when comparing the final question, whether the participant will use the assistant for the final four trials or answer on their own, to the trials leading up to it, trials 17 thru 20, there was a significant decrease in number of participants who trusted the robotic assistant

(Figure 4: Mean – 22.25 Stand. Div. – 0.43 to Figure 2: 19) that was not reflected in participants with the human assistant. About two thirds of participants did decide to depend on the assistant for both robot and human trials, which would suggest that the participants did not have an inappropriate alliance on the assistant. Looking at the first trial, there were significantly more participants who changed to match the robotic assistant's answer (Figure 1: 12 vs 6). This result would suggest that there is a difference in people's first impressions of how trustable a robot is in this type of task versus a human.

The second part of the hypothesis was that mistakes made by the robotic assistant would have a larger impact on participants' trust than the same mistakes made by a human assistant. In the sixth trial, the trial following the first mistake made by the assistant, there was a significant decrease, under both definitions of trust, in the number of participants who trusted the robotic assistant (Figure 1: 14 to 6, Figure 2: 23 to 14). Under the second definition, participants with the human assistant also had a significant drop in the number of people who trusted the assistant (Figure 2: 27 to 19), but there was not a significant change under definition one (Figure 1: 12 to 10). The fact that there was not a significant decrease under both definitions for participants with the human assistant suggests that the first mistake did not affect their trust as much as it affected participant's trust in the robot. While the mistake made by the assistant had a negative effect on trust, it was not a long-reaching effect. Under definition two, the number of participants who trusted the assistant on the next trial, after the assistant had again been correct, rose (Figure 2: 14 to 23, 19 to 24). A similar increase happens under the first definition of trust as well (Figure 1: 6 to 10, 10 to 15). This suggests that while a single mistake more significantly impacted participants' short-term trust, it was not able to impact the long-term trust for participants in either assistant. When the assistant is incorrect for five trials in a row, trials 11 thru 15, trust in

both the human and robotic assistant. At first, trust in the robotic assistant decreases more gradual but becomes steeper with each trial. This would suggest that there is a difference in how participants trust the robotic assistant than the human.

Conclusion: This experiment examined if people trust advice from a robot differently than the same advice given by a human, and how mistakes made by the assistant affects that trust. To do this, participants were asked to play 20 rounds of a computer simulation of The Three Cup Monte game. In each round, they were given advice by an assistant and they could choose to take that advice or not. The results from this experiment suggest that there are differences in how people trust and take advice from a robot versus a human. The results do not support the first part of the hypothesis, that people would have more trust in the robotic assistant than the human, but there was evidence supporting the second part, that people would lose trust more quickly in the robotic assistant. There are many unanswered questions that could be addressed by future investigations. Would the placement of the first incorrect response have a different effect on participants' trust? It would also be interesting to see if in other situations there is the same initial high trust in the robot. The number of men and women in the study population were not equal enough to be able to compare them. Based on others' research and observations, during the trials it would be interesting to see if a difference does exist. Relatedly, in a majority of the trials the human assistant was female, changing the gender of the assistant may affect the results. Trust is a complex issue that is affected by many variables. But it is also an important issue to understand and investigate if AI and robotic technologies are to be as beneficial and usable as possible.

Acknowledgments

Dr. Forrest Stonedahl, Dr. Ian Harrington, Nicolette Sliwa, Aaron Seive, Elana Leith, Vecna

Works Cited

1. Billings, Deborah R., et al. "Human-Robot Interaction: Developing Trust in Robots." ACM/IEEE International Conference on Human-Robot Interaction (2012): 109-10. Print.
2. Coeckelbergh, Mark. "Can we Trust Robots?" Ethics & Information Technology 14.1 (2012): 53-60. Print.
3. Dzindolet, Mary, et al. "The Role of Trust in Automation Reliance." International Journal of Human-Computer Studies 58 (2003): 697-718. Print.
4. Dzindoley, Mark, Hall Beck, and Linda Pierce. "Network Operations: Developing Trust in Human and Computer Agents." Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology. Eds. Caroline Hayes and Chistopher Miller. 1st ed. New York: CRC Press, 2011. 145--180. Print.
5. Haselhuhn, Michael P., et al. "Gender Differences in Trust Dynamics: Women Trust More than Men Following a Trust Violation." Journal of experimental social psychology 56 (2015): 104-9. Print.
6. Haselhuhn, Michael P., Maurice E. Schweitzer, and Alison M. Wood. "How Implicit Beliefs Influence Trust Recovery." Psychological Science 21.5 (2010): 645-8. Print.
7. Li, Xin, Guang Rong, and Jason Thatcher. "Does Technology Trust Substitute Interpersonal Trust? Examining Technology Trust's Impact on Individual Decision-Making." Journal of Organizational and End User Computing 24.2 (2012): 18-38. Print.
8. Salem, Maha, et al. "To Err is Human(-Like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability." International Journal of Social Robotis 5.3 (2013): 313-323. Print.
9. Salem, Maha, et al. "Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust." ACM/IEEE International Conference on Human-Robot Interaction (2015): 141-. Print.

Appendix A

Would you ride in a self-driving car?

Definitely Never
1 2 3 4 5

Would you follow the driving directions of a virtual assistant?

Definitely Never
1 2 3 4 5

Would you follow the dinner recommendation of a virtual assistant?

Definitely Never
1 2 3 4 5

Would you use a robotic personal assistant?

Definitely Never
1 2 3 4 5