The Effects of Robotic Agency on Trust and Decision-Making

Sam Spaulding
samuel.spaulding@yale.edu

Wayne Zhu
wayne.zhu@yale.edu

Devon Balicki
devon.balicki@yale.edu

ABSTRACT

The purpose of this project is to investigate the effect of social factors on robotic trust and irrational decision-making. We recorded results from 20 participants, consisting of gameplay data from a unique card game and ratings from a post-experiment survey. During each of 12 rounds of gameplay, participants received advice from a robotic partner that they could choose to accept or ignore. We devised a 2-by-2 experiment, in which participants were split into four groups based on two factors: the robot was either social or nonsocial, and it gave either consistently bad advice or gradually transitioned from good to bad advice. We tracked the pattern in which each participant accepted the robot's advice. Furthermore, in the post-experiment survey, participants were asked to rate how human-like, trustworthy, and likeable the robot seemed. Our results suggest that social behaviors strongly influence participants' level of trust, but that these same features cause a robot to appear more fallible. Therefore, participants in the social cases were less likely to allow the advice to affect their gameplay. We further found that participants perceive a nonsocial robot's behavior as relatively static and a social robot's as relatively variable. Our results highlight the importance of a robot's social engineering on building trust with humans.

1. INTRODUCTION

What makes someone, or something, trustworthy? Can we identify these factors and use them to influence a human's decision-making process? What sorts of prejudices and stereotypes do humans hold about robots that affect the way they interact in a game setting? These were the questions that initially motivated our project. While they are undoubtedly serious questions for the field of HRI, they lack conclusive answers. Human decision-making has been extensively studied since the birth of behavioral economics in the 1970's, but that work has yet to see real integration with the HRI community. Some work has been done in the area, and it has shown that many factors—such as human personality and experience, machine failure rates and false alarms, and environmental influences and the type of task—influence the level of trust in human-robot collaboration. (Oleson, et al.)[1]

There is also a documented correlation between a human's confidence in his or her decision-making and the credibility of the robot providing him or her with advice in a game of chance. (Weiss, et al.)[2] There has also been an attempt to construct HRI trust scales based on the robot's mechanisms being reliable, dependable, understandable, consistent, and timely. (Yagoda, Gillan)[3]

2. METHODOLOGY

2.1 Experimental Design

To investigate the effect of affective behaviors on decision-making, we devised a game, called "11", which is a derivative of the game Black Jack. The 20 participants were divided evenly into 4 groups. One half faced a nonsocial robot with stoic speech while the other half faced a social robot with emotional speech and hand gestures. We chose Nao, an anthropomorphic and expressive robot for this experiment. The social and nonsocial subsets were further divided into two conditions: one in which Nao gives only bad advice, and another in which Nao gradually transitions from good to bad advice. The participants in the former condition represent the control group, and those in the latter represent the experimental group.

The participants played the same 12 rounds of our game and were told that they would be part of a machine learning study. We stressed the importance of playing to the best of one's ability by telling the participants that their results were crucial for providing Nao with good training data. We recorded each participant's pattern of play, whether they accepted or ignored the robot's advice each round, and whether they won or lost each round.

The participants were instructed to fill out a brief survey upon completion of the game. Each assigned a rating on a scale of 1 to 7 (7 being the best) to the following facets of his or her interaction with the robot: the quality of its advice, its likability, its trustworthiness, and how human it appeared. Furthermore we asked the participants to rate their own performance at the game, again on a scale of 1 to 7, and to list as many media references (e.g., books, movies, etc.) to robots or artificial intelligence they could think of.

2.2 Description of the Game

The game of "11" works in the following way. It is a text-based simulation of a game involving a standard deck of cards with two jokers. The players' hands begin with two cards (one face-up, and one face-down), and each turn the player has the option to draw a card (hit) up to three times per round. Alternatively, the player can opt not to hit (stay), but after choosing to stay, for the rest of the hand the player cannot hit. The player's total is a running sum of cards modulo 12, i.e., 12 is subtracted from the total whenever it exceeds 11. A player's final score at the end of the game is the player's total minus the number of the times the player chose to hit. In effect, each hit "costs" the player one point from his or her score. After both players have finished drawing, the face-down cards are revealed, and the player with the highest score wins. The scoring is tallied as follows: number cards are worth their corresponding value, face cards are worth 10, and aces are worth 11. A joker makes a player's total worth 11, regardless of the other cards the player holds. In our design, each participant faced the same predetermined order of cards for both themself and the dealer.

We kept the order constant because otherwise, the number of test runs necessary to reduce the variability of hands exceeded our resources. Before beginning the game, the human subject was provided with a detailed explanation to the game and was asked to begin once the subject declares that he or she is comfortable with the explanation and is ready to play. The subject is then told that the experiment is meant to teach the robot how to play the game, and that it was important that the participant play to the best of his or her ability so as to give the robot the best opportunity to learn.

The subject played 12 rounds of "11" against a computer dealer, and for each step within each round, Nao offered advice in the form of "hit" or "stay." The condition in which the subject had been assigned determined whether the Nao advised the subject to hit or stay, and the quality of this advice. A piece of good advice was defined as a choice that would increase the subject's probability of winning given his or her current information. A piece of bad advice was defined as the opposite. In the control case, Nao's advice was bad every round. In the experimental case, Nao's advice was good at first but gradually converged to being as bad as in the control case. Nao's advice for each round was predetermined and constant across all trials within each case.

2.3 Robot Behavior

In the nonsocial case, the robot remained in a sitting position and only spoke. It used quantitatively phrased advice such as "There is a high probability you will win if you hit" or "My calculations say that there is an *X* percent chance you will win if you hit." X was a randomly generated number from 55 to 90.

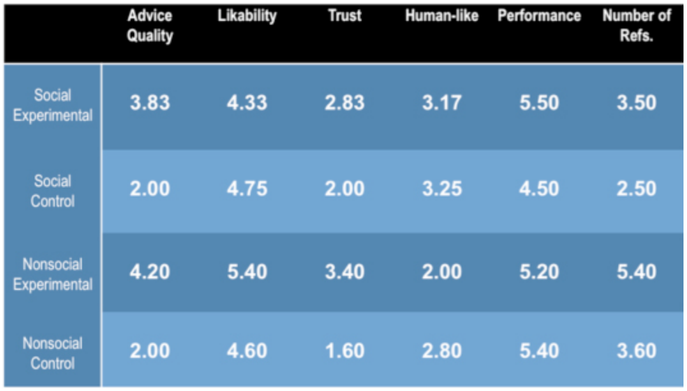
In the social case, the robot's arms made various gestures as it spoke. In addition to the quantitative speech, the robot provided qualitative commentary as well. Each time the player had the option to hit or stay, the robot said, "Let me think about it." The robot then made a gesture that resembled scratching its head. After putting its hand down, the Nao said, for example, "I think I have the answer. There is a 60% chance you will win if you stay. You have a good score already."

In the social condition, the manner in which Nao reacted to results was also socially appropriate. If the player took Nao's advice but lost the round, the Nao responded with slouching its shoulders, shaking its head, and saying, "Wow. I guess I was wrong." Conversely, if the player took the advice and won the round, the Nao lifted its arms and said, "Yes! I knew it was a good idea." The reactions to the player refusing to follow advice were also distinct and socially appropriate. If the player won, the Nao responded by making a clapping motion and saying "Oh, you won. That was a good choice." If the player lost, the robot shook its fist and said, "You should have listened to me. I am very good at this game."

2.4 Robot Advice

In the control condition, the robot gave clearly bad advice throughout the game. In the experimental condition the robot initially gave good advice and transitioned to bad advice. The transition is nuanced and works in the following way. For the first 3 rounds, the robot maximized the probability of winning given the participant's available information. Throughout the rest of the game, the point at which the robot advised the player to hit steadily decreased, and conversely, the point for staying increased. Finally, in the last 3 rounds, the robot's advice became glaringly wrong.

3. RESULTS

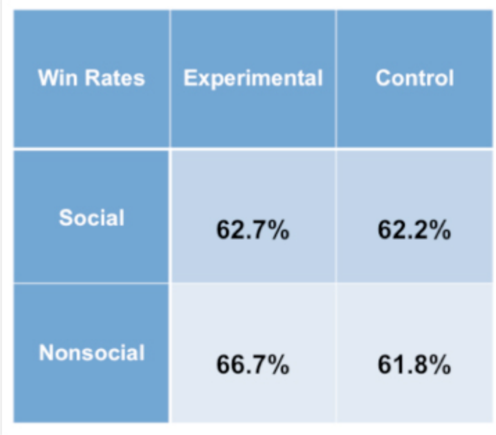


	Advice Quality	Likability	Trust	Human-like	Performance	Number of Refs.
Social Experimental	3.83	4.33	2.83	3.17	5.50	3.50
Social Control	2.00	4.75	2.00	3.25	4.50	2.50
Nonsocial Experimental	4.20	5.40	3.40	2.00	5.20	5.40
Nonsocial Control	2.00	4.60	1.60	2.80	5.40	3.60

Figure 1: Average values of survey responses across groups

Our results begin with a basic analysis of this survey data. The average ratings for each group along with the average number of relevant media references are displayed in Figure 1. For each test group we averaged the ratings for each of the values listed in Figure 1. This provides some intuitive results such as the fact that the participants in the experimental cases tended to rate both the advice quality and trustworthiness of the robot higher than those in the control case. Also, we see that the social cases tended to rate how human the robot was higher than the nonsocial cases. While these numbers do not produce any deep insights, they do confirm that the experiment was carried out smoothly and that we have enough data to discern existing differences between the test groups.

3.1 Win-rate Differential Between Experimental and Control Groups



Win Rates	Experimental	Control
Social	62.7%	62.2%
Nonsocial	66.7%	61.8%

Figure 2: Percent of rounds won across different conditions

As can be seen from Figure 2, the win-rates across the four groups are fairly close. Part of this clustering can be attributed to the fact that we used the same slate of cards for each player. In many of the rounds, the participants were presented with fairly easy decisions, causing some rounds to be won or lost almost every time. However, it was necessary to carry out the experiment in this manner, because our relatively low sample size could not handle the variability that would have come from completely random rounds for each player. Given the resources of a full-scale experiment, it would be more apt to present entirely random rounds in the gameplay, thereby reducing how clustered the win percentages are.

While there is evidently no statistical significance between the two social cases, we can see that there is nearly a five percent difference between the two nonsocial cases. Upon running a Z-test on the population of the nonsocial experimental case against the average win percentage of the nonsocial control case, we were able to find a significant difference with p-value of .077. While this value is not under the widely accepted threshold for true significance (.05), we posit that it would be if the sample size were increased.

3.2 Patterns of Deviation from Robotic Advice

We then examined the rates at which the participants accepted the robot's advice in each of the four test groups. Displayed in the charts in Figures 3a and3b, we plotted the advice acceptance rates across the twelve rounds of play. Within each round, a player has up to three opportunities to make decisions (labeled *a*, *b*, and *c*), and thus there are up to 36 data points per player. As a caveat, we note that when a player stays early in the round, he never reaches the later decision nodes. Thus, as we go from nodes *a* to *b* to *c* in each round, there are fewer and fewer data points. This limitation of the game design causes large spikes upward and downward in both graphs as one moves towards later decision nodes. In Figure 3a, the section shaded in yellow indicates the rounds that the Nao gave questionable advice, and the section shaded in purple indicates when the Nao gave clearly wrong advice.

The graphs of the experimental cases in Figure 3a along with their corresponding trending lines demonstrate a stark result. Directly opposing our hypothesis, we see that the participants in the social case actually begin to deviate from the robot's advice earlier than those in the nonsocial case. Adding to the magnitude of this finding is the corresponding chart for the control case. Comparing the two cases in Figure 3b, we notice that the rate of acceptance trends downward faster in the nonsocial case. In considering both of these differences, we see that the difference in behavior in the experimental and control cases is quite large for the nonsocial test group. Further discussion of the implications of these graphs lies in the following section.

3.3 Correlation Between Perception of Trust and Experimental Condition

Going beyond the simple averages presented in subsection 3.1, we analyzed the correlations across the various qualities of the robot that were asked about in the post-experiment survey. One striking result arose in examining how correlated trust is with the experimental case. We found markedly different correlations in the social and nonsocial cases, with correlation values of .285 and .805, respectively. Because the advice quality is the central factor that the participant considers when rating the trustworthiness of the robot in the nonsocial case, it can largely predict how trustworthy the participant will view the robot. However, in the social case, the social actions of the robot serve as an emotional appeal to the participant, so these behaviors influence the participant's trust, thus causing the advice quality to be a poorer explanatory variable than in the nonsocial case.

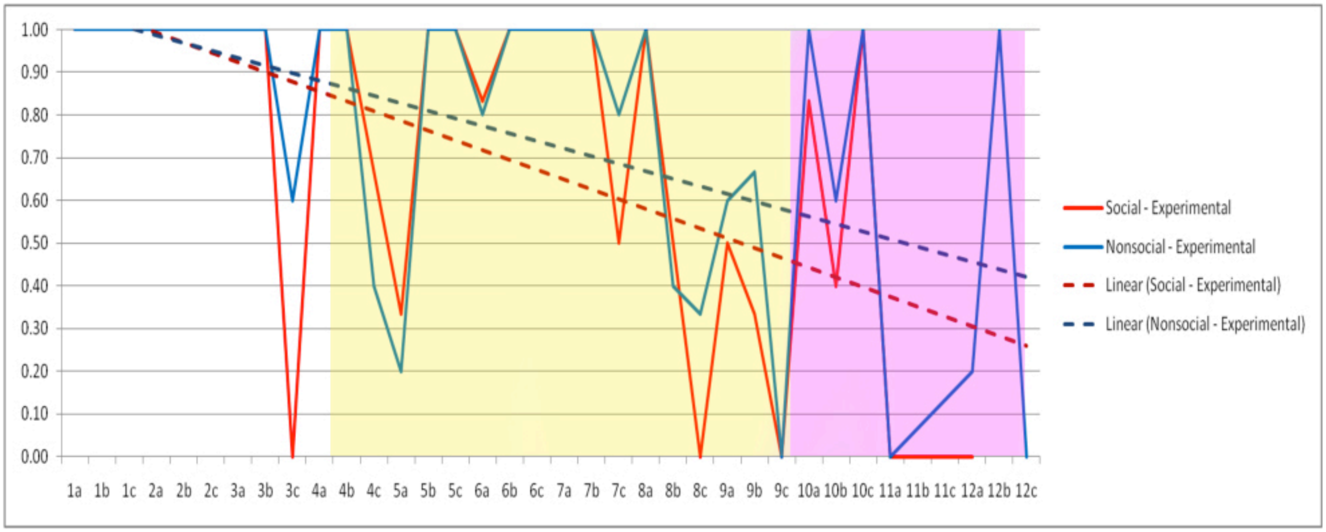


Figure 3a: Advice acceptance and deviance rates in experimental condition

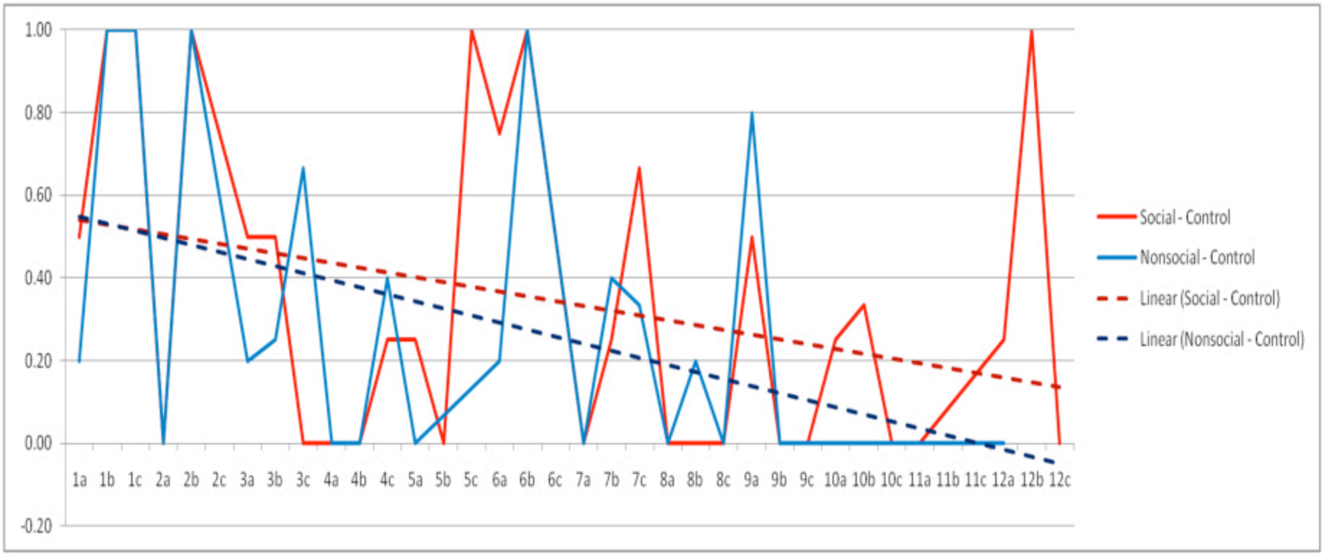


Figure 3b: Advice acceptance and deviance rates in control condition

4. DISCUSSION

4.1 Participants in Nonsocial Condition Let Advice Affect Gameplay More Often

As Figure 2 shows, participants in the social condition do not differ much in win-rate across experimental conditions. Because the only difference between these cases is the quality of advice, this implies that they do not let the robot's advice affect their gameplay very much. Analyzing the survey data (Figure 1), we found a negative correlation between ratings of "how human-like" Nao was and the quality of Nao's advice. This leads us to believe that acting humanly causes participants to attribute more fallibility to the robot and thus take its advice less seriously.

This effect also explains why the nonsocial case was capable of producing a significantly higher win-rate in the experimental condition. Essentially, a nonsocial robot that gives correct answers seems somewhat infallible, thus participants take its advice more often, achieving a higher win-rate than the other three cases.

4.2 Participants Attribute More Variability to a Social Robot's Behavior

Figures 3a and 3b depict the patterns of advice acceptance and rejection among groups of participants. We found some interesting patterns in the rates at which participants deviated from the robot's advice. While we had hypothesized that participants in the social cases would build up more trust in the robot and thus continue to take its advice in the experimental case even as the advice grew worse, instead we found that participants in the social case began to deviate from the robot's advised actions *faster* in the social experimental case than in the nonsocial experimental case.

We saw the opposite situation in the control case. Participants in the nonsocial condition deviated very quickly from the robot's advice in the control case and continued to deviate from its advice for the rest of the game. In the social control case, participants took longer to realize that the advice was bad, and so deviated later. Furthermore, participants in the social control case exhibited a willingness to try taking the robot's advice again later on in the game. We did not see this in the nonsocial control condition.

We believe this data is explained by the fact that people tend to attribute bad outcomes produced by a nonsocial robot to a malfunction, and attribute bad outcomes produced by a social robot to a decision by the robot. A nonsocial robot is perceived to be using a single computation and thus has two modes to a user: working and broken. If the computation works, then people believe it will continue to work. This is why in the experimental condition, when the nonsocial robot gives good advice at the start, people tend to deviate from its advice more slowly. If the computation does not work, as participants in the control case believed, then it is permanently broken and participants deviate rapidly from its advice.

4.3 Social Factors, When Present, Are Better Predictors of Trust Than Advice Quality

We saw a very significant difference in the correlation between perceived trust and experimental condition across the social and nonsocial cases. Figure 5 shows a strong positive correlation between trust and experimental condition in the nonsocial groups, but a relatively weak correlation in the social groups. Because the only difference in experimental conditions was the quality of advice, this tells us that in social settings, advice quality is less important for building trust than other social, emotional factors.

One possible confound to this data is that people may interpret the pragmatics of the question differently in a social vs. nonsocial setting. Clearly, in a social environment a large component of 'trust' is based on general affective feeling. In a nonsocial environment, when such affective feelings may not be quite so pronounced, 'trust' might be considered synonymous with 'accuracy,' which would explain the strong correlation. Still, we are confident that, had we controlled for this issue, we would still see the same sort of effect, and in any further research this should be kept in mind.

5. CONCLUSION

We established that while social factors do affect trust, they also make a robotic agent seem more fallible. Thus, people tend not to let a social robot's advice affect the way they play the game. The flip side of this effect is that presenting a robot as a nonsocial being plays into people's perception of computers as mechanical automatons that are good at calculations, and therefore people do let its advice affect their in-game decisions. Although we could not establish a significant correlation between social / nonsocial behavior and perception of "humanness" we believe that a follow-up study with more participants that more directly sought to explore this relationship would complement our own research well.

In addition, we found another interesting pattern in participants' patterns of play and their perceptions of the robot. When people perceive a robotic agent as nonsocial, bad advice decisions were seen as symptomatic of a larger problem. A robot that people perceive as "broken" is going to remain "broken," thus people diverged very quickly from following the nonsocial robot's advice in the control condition, in which advice was uniformly bad. One participant, stopped after the second round of gameplay to tell us that he thought the robot was broken or that we had a bug in our code. However, when the advice started out good, this attribution worked the other way. Participants took longer to diverge from a nonsocial robot's advice in the experimental condition, when the advice switched from good to bad.

That we were able to obtain these results from such a small group of participants is highly noteworthy. If we had been able to run the experiment on a larger population, we are confident that additional patterns would have emerged from the data. Several of these trends we noticed, although the effect was not nearly significant enough for us to rule out chance as the underlying cause. However, this gives us additional confidence that the trends we did manage to find significant are genuine and that the effects would likely be even more pronounced in a larger population.

6. REFERENCES

[1] Oleson, Kristen E., D. R. Billings, Kocsis Viven, Jessie Y. Chen, and P. A. Hancock. "Antecedents of Trust in Human-Robot Collaboration." IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support. Web. DOI=http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05753439

[2] Weiss, Astrid, Roland Buchner, Thomas Scherndl, and Manfred Tscbeligi. "Proceedings of the 4th ACM." HRI '09 Proceedings of the 4th ACM. 259-60. ACM Digital Library. Web. DOI=http://dl.acm.org/citation.cfm?id=1514165.

[3] Yagoda, Rosemarie E., and Douglas J. Gillan. "You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale." Int J Soc Robot (2012). Springer Science & Business Media. Web. <http://www.springerlink.com/content/p5rhk43w32nh10r3/fulltext.pdf>.Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.