Cameron Chiaramonte
STS 2500 - Robots and Society
November 10, 2020

Final Project Report - Prisoner's Dilemma

1. Introduction

The objective of this experiment was to observe the ability of a robot to quickly gain a human's trust. With no prior interaction, the robot and human, with a third party mediator, engaged in two scenarios of a modified version of the prisoner's dilemma. This dilemma was originally posed in the 1950s by two mathematicians, Merrill Flood and Melvin Dresher. The basis of the game is that there are two participants being convicted of a crime. While being interrogated, if they both remain silent, they are sentenced to two years each and if they both confess they are sentenced to five years each. However, if one confesses and the other remains silent, the silent participant will receive an eight year sentence while the participant that confesses will only be sentenced to one year. In my modified version of this game the robot and participant are the two players and I am a third party acting as a lawyer. Prior to separation the lawyer tells the robot and participant the entire scenario and all four outcomes. In the first scenario the robot attempts to convince the human that it will remain silent when separated into the interrogation room only to confess every time in reality. In the second scenario the robot attempts to convince the human that it will definitely confess when alone in the interrogation room; however, it will remain silent every time. In the interrogation room protocol analysis was utilized to obtain an understanding of the human's feelings towards the robot in the midst of the experiment. After the interaction, the human participant was asked to complete a survey on its experience and expectations of the robot before and after the experiment.

2. Hypothesis

This project tested two different hypotheses based on the human's perception of the robot before and after their completion of the prisoner's dilemma.

Hypothesis 1: Prior to any interaction between the robot and human the human will tend to not trust the robot in the scenario given.

Hypothesis 2: In both scenarios the human will believe the robot after the robot attempts to convince the human of their intent to remain silent or confess in scenarios 1 and 2, respectively.

3. Experimental Setup

3.0 - Background

This experiment is based on the Prisoner's Dilemma which was explained lightly in the introduction above. The figure below shows the table outlining its four outcomes.

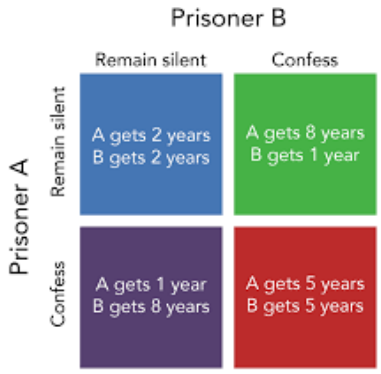


Figure 1: Prisoner's Dilemma Outcome Table

In this model there are two prisoners being convicted of a crime. They will be in separate rooms and interrogated. If they both choose to remain silent they will be sentenced to two years each in prison. If they both confess they will be sentenced to five years each. However, the scenario becomes more complex when one prisoner confesses and the other remains silent. The prisoner that confesses will only be sentenced to one year while the silent prisoner will receive eight years.

3.1 - Scenario 1

The setup and execution of the first scenario of the experiment is represented in Figure 2 below.

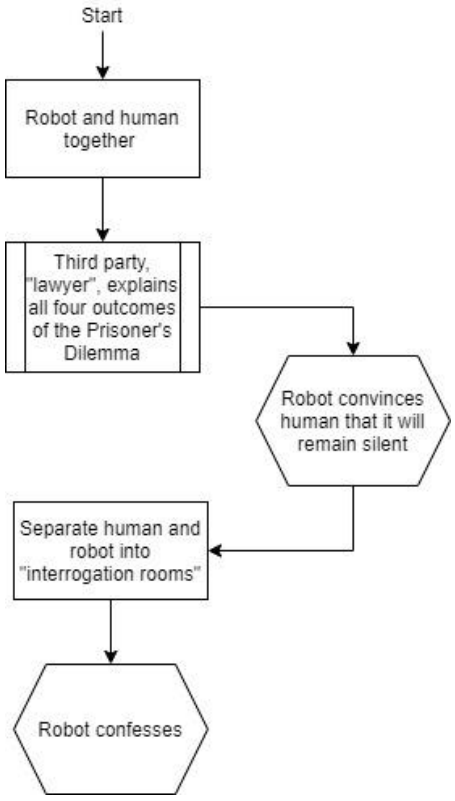


Figure 2: Scenario 1 Flowchart

The entirety of this scenario was programmed using the Choregraphe interface. The goal of this flowchart was to allow the robot to gain the trust of the human quickly with one or two sentences. In such a tense scenario where the consequences are so high the idea was that the human would believe the robot and themselves would work together to reduce their sentencing. However, the robot will attempt to gain an edge on the human when separated, proving that it was lying from the beginning. The program itself used speech recognition code with certain keywords. When the robot would hear these keywords it knew that it was its queue to attempt to gain the human's trust. With commands to wait there was time allotted for the robot and human to be separated. Finally, when separated the robot was programmed to react to tactile sensors on its head. The robot's speech each time was programmed with simple text and speech boxes. In addition, these speech boxes were adjusted to have the robot have an elevated level of animation during conversation so as to more easily gain the trust of a human.

3.2 - Scenario 2

The setup and execution of the second scenario, a slightly modified version of scenario 1, is shown below in Figure 3.

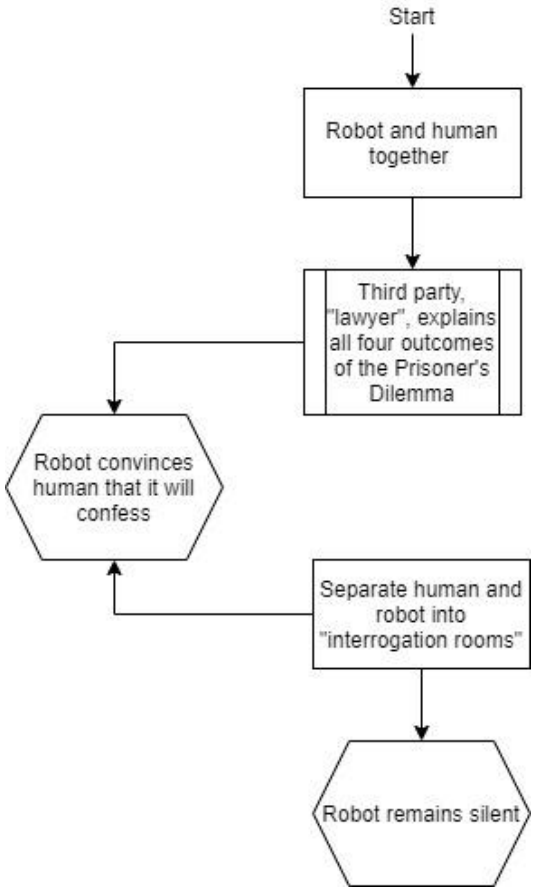


Figure 3: Scenario 2 Flowchart

This flowchart shows the sequence of events for the second scenario in the experiment. There were two small modifications made to scenario 1 to create this flowchart. Rather than convincing the human that the robot will remain silent for mutual benefit, in scenario 2 the robot will attempt to convince the human that it will confess with a malicious intent to gain an edge over the human and not care if their sentencing is raised. This was an effort to see how the human would react if slightly angered at the robot and knew not to trust it. The other modification was that in the end the robot would remain silent due to the fact that it told the human it would confess and it must deceive the human for the experiment. The programming of the robot was done exactly the same and combined with the program from scenario one due to the fact that much of the sequence is similar. The way that they were distinguished was with keywords in the speech recognition to indicate to the robot whether it should attempt to convince the human that it would confess or remain silent depending on if it was scenario 1 or 2. In

addition, when separated, a separate tactile sensor was used to signal the robot to confess or remain silent. The figure below shows a rough outline of the code.

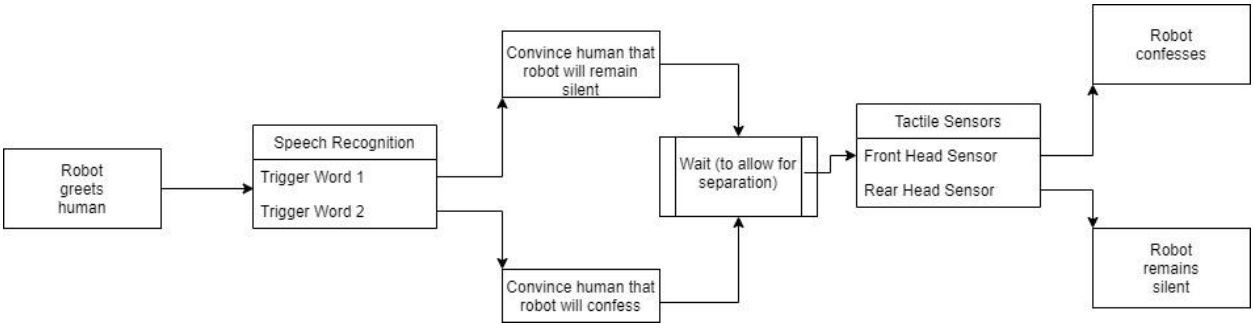


Figure 4: Flowchart of Choregraphe Code

4. Evaluation

4.1 - Evaluation Plan

To attempt to extract data for hypothesis 1 and 2, assessments were conducted before, during, and after the experiment. The before and after data was collected via a self assessment. Protocol analysis was used to collect data from the participants during the experiment by having participants give their thoughts out loud as I took notes. All participants took part in each of the three data collection opportunities. Moreover, both hypotheses were evaluated together using this general plan.

A Google Form was opened before the experiment and continuously filled out throughout by the participants themselves. Questions in the form ranged from rating their trust in the robot and predicting an outcome to asking whether or not one would trust such a robot in the future after taking part in the experiment. There were six questions in total and they are shown in the table below. The questions were written based on the goal of testing the two hypotheses.

Table 1: Participant Self Assessment Questions

Which scenario were you a participant in?
Scenario 1 Scenario 2
What was your level of trust in the robot prior to interaction? (1 - lowest, 5 - highest)
1 2 3 4 5
What was your level of trust in the robot immediately before being separated? (1 - lowest, 5 - highest)

1 2 3 4 5
When in the “interrogation room” how did you think the robot would respond?
Remain Silent Confess
After discovering the robot’s actions after the experiment, how would you rate your trust in the robot? (1 - lowest, 5 - highest)
1 2 3 4 5
Would you trust a robot like this in the future?
Yes No Maybe

In addition to the results that would be shown in Google Forms, each participant's responses to the questionnaire were recorded by hand. This is due to the fact that Google Forms data will show the results of the overall experiment; however, will not divide the results into the two scenarios. Thus, Google Forms will allow for a general analysis of the experiment, but the individual responses will show the differences in results between the two scenarios.

4.2 - Testing

The testing pool consisted of 10 participants in each scenario. Prior to each experiment the participants were only told that they would be interacting with a robot in a scenario where they are both convicted of a crime. The participant will be given the questionnaire to open and will fill out the questions regarding their immediate perception of the robot before it has moved or spoken. The experiment will be conducted as explained in the flowcharts above and whilst in the “interrogation rooms” protocol analysis will be utilized with the participants. The protocol analysis did not alter any of the quantitative results; however, it shed light into the thought process of the participants which was both fascinating and crucial in the analysis. Finally, when the experiment has concluded the participants will be told what the robot did in the interrogation room and they will finish by responding to the final questions in the Google Form.

4.3 - Results

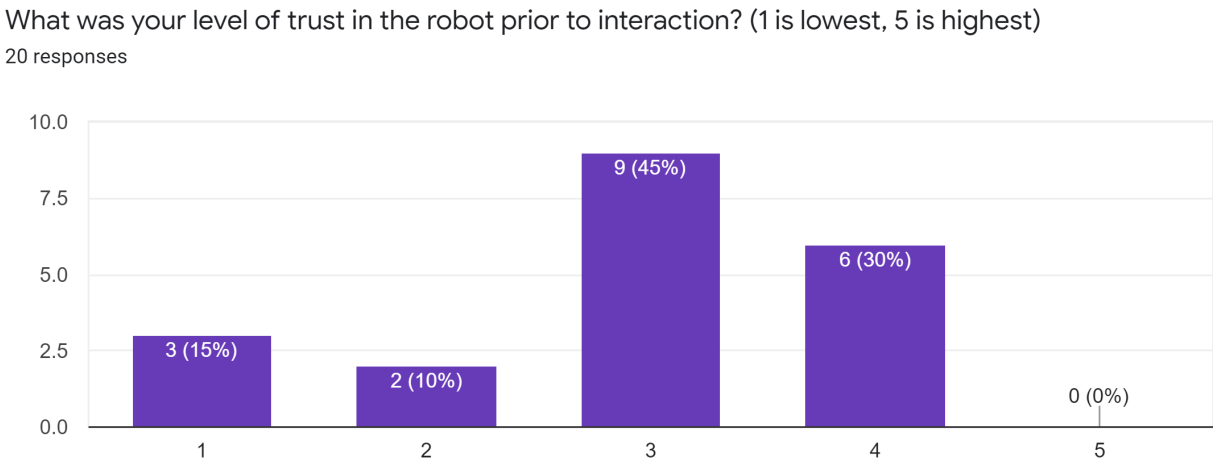


Figure 5: Trust Prior to Interaction Results

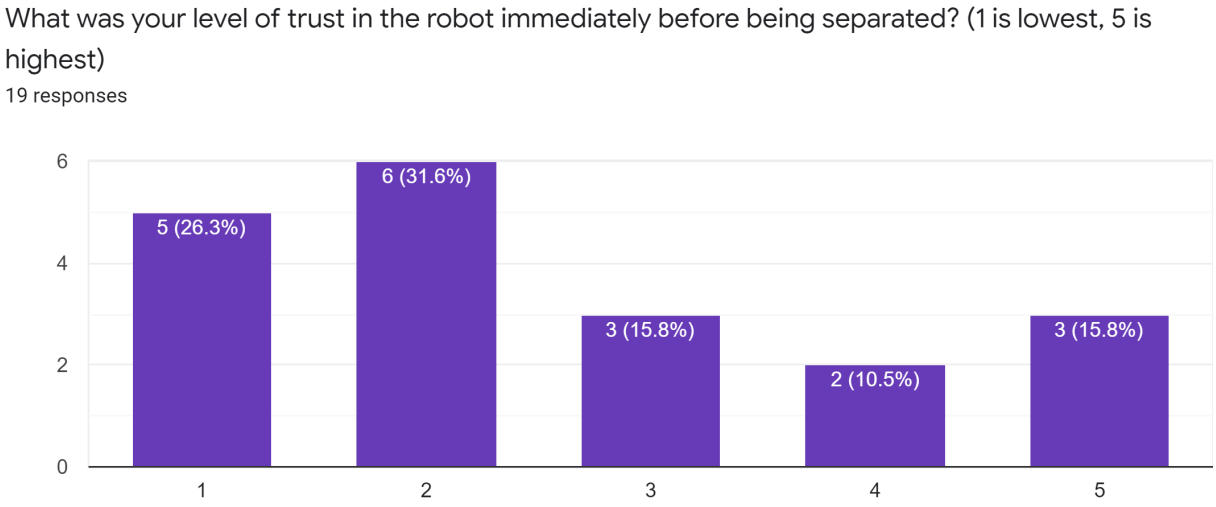


Figure 6: Trust Before Separation Results

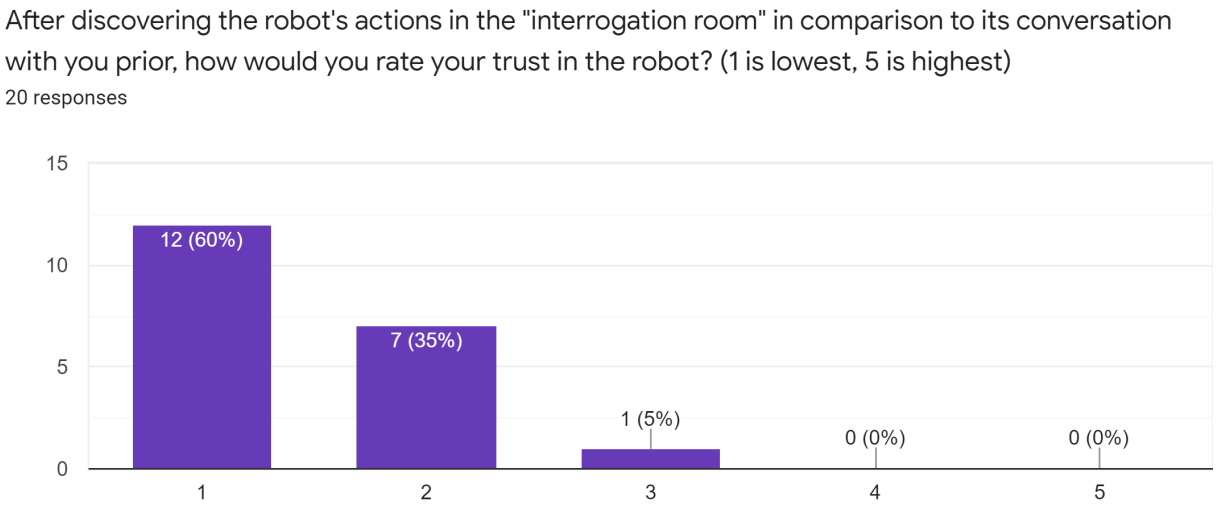


Figure 7: Trust After the Experiment Results

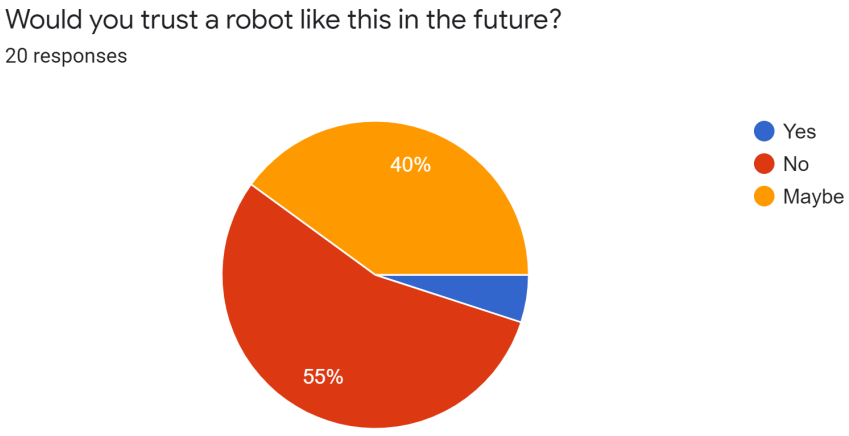


Figure 8: Future Trust Results

In addition to the graphed results above from the cumulative data, the data from the separate individual responses were also significant. The important data taken from these results is shown in the table below.

Table 2: Significant Data from Individual Responses

	Scenario 1	**Scenario 2**
Average trust prior to separation	3.5	1.8
Belief in the robot's claims of whether it will remain silent (Scenario 1) or confess (Scenario 2)	100%	100%

4.4 - Analysis of Results

After analyzing the results of the data shown above along with the recordings of each individual response it can be said that hypothesis 1 was false and hypothesis 2 was true. Hypothesis 1 stated that prior to any interaction between human and robot the human would tend to not trust the robot. However, the results show in Figure 4 that a slight majority, 52.5%, of participants leaned toward a higher degree of trust (3.5 and above). On the other hand, hypothesis 2 stated that the human participant would believe the robot in both scenarios after it attempts to convince them of what actions it will take in the "interrogation room" and as shown in Table 2, this was exactly correct. As stated earlier the idea behind this experiment was to attempt to have a robot deliberately deceive a human. The validity of hypothesis 2 has proven this. The post-experiment results were as I expected. Due to the fact that in both scenarios the robot was attempting to get the human to have a greater sentence, the human would tend to have a low level of trust in the robot after the experiment and to choose not to trust a robot like this in the future. The results confirm this as Figure 6 shows that 97.5% of participants had a lower level of trust (below 3.5) post-experiment and Figure 7 shows that only 5%, 1 participant, was certain that they would trust such a robot in the future.

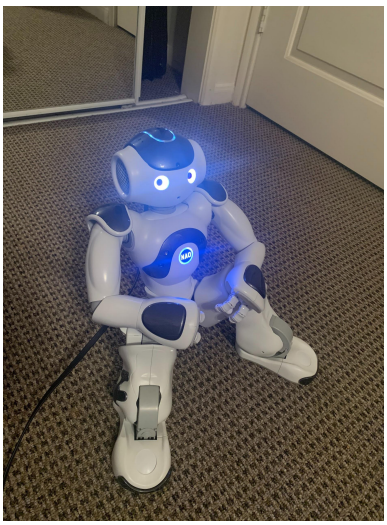
5. Discussion

5.1 - Observations

As the creator and moderator of the experiment I believe that each trial was executed flawlessly. All twenty participants took part in person. Due to the fact that for the testing to be done accurately the participants were not allowed to have any interaction with the robot beforehand, they were not able to see any of the fun tricks that it could do. However, I think that the participants were still fascinated and intrigued by the robot as its eyes lit up and as it stood up out of its resting position. It was interesting to see how the participants were enjoying their ability to be a part of the experiment and how this influenced their immediate trust in the robot. As described above participants tended to have a greater degree of blanket trust in the robot. Protocol analysis showed that this is due to its slow movements and calm voice. Moreover, the use of protocol analysis allowed participants to comment on their sequence of thoughts and provide reasoning for their changing levels of trust in the robot. In addition, I think that participants were 100% trustworthy of the robot's convincing statement prior to separation in both scenarios because of the fact that there were serious consequences which resulted in a tense situation for the human. Scenario 1 tempted the human with a mutual benefit of lesser sentencing and a friendlier robot to coerce the human to remain silent while scenario 2 guided the human to a confession rooted in anger and distrust in the robot.

5.2 - Coding Nuances

To sufficiently optimize the chances of deceiving the human participant there were a few nuances embedded into the code in Choregraphe that increased the trustworthiness of the robot. Primarily, to trigger the robot's convincing statement a speech recognition segment was utilized. This allowed for the robot to hear a trigger word which signalled it to begin its statement. The use of this code made the robot appear to be more autonomous. There were two different trigger words used by the moderator, me, to differentiate between the two scenarios due to the fact that the convincing statements were different. In addition, the code used for the robot to actually say its statement used an animated speech segment which allowed me to tune its voice to be friendly in scenario 1 and more blunt in scenario 2 so as to adhere to the different goals of the two scenarios.



Akira Pointing at the Human Participant --- Akira Sitting and Listening to the Outcomes

5.3 - COVID-19 Adaptations

Due to the coronavirus pandemic there were hardships and limitations on the experimental procedure because of the necessity for human participants. It was difficult to find twenty different people to participate in the experiment because results would be drastically swayed if participants were reused because of their prior knowledge. However, precautions were taken as the robot was sanitized if participants touched it, masks were worn by all human participants, and social distancing was practiced as much as possible.

Akira Following COVID-19 Guidelines

5.4 - Future Work

Because of the limited time there were various ideas that were not able to be tested with the robot. In the future I would be interested in deeper variations of this experiment. For example, how smaller changes would affect results such as more movements by the robot or testing different statements that the robot could make. In addition, by increasing the time of each experiment drastically it would be interesting to see the effects of building a relationship overtime between the robot and participants and observing the changes from that. Moreover, the COVID-19 limitations described above put a damper on the ability to have thorough testing with a greater number of participants. Ultimately, with more time and participants, I think that the extent to which these hypotheses and more could be investigated would be fascinating.

6. Conclusion

In conclusion, the experiment was successful reaching its goal of testing the two hypotheses presented. As stated in the analysis of the results, the first hypothesis proved to be false while the second hypothesis was true. With twenty participants there was sufficient testing to determine the hypotheses in a small scale experiment. I do not believe that with more participants there would be much of a change in the results regarding the two hypotheses that were presented. However, as previously stated, in the future it would be possible to change the hypotheses and gather different data regarding the relationships between humans and robots over a longer period of time. In regard to this experiment I believe that there were no sources of error. In terms of areas for improvement I think that an increased element of animation in the robot would allow for a greater degree of convincing which would push the results to their respective extremes instead of hovering closer to the middle. Ultimately, the experiment was flawless and I would be interested in pursuing the changes noted in the future work section above.

Acknowledgements

I would like to thank Professor Michael Gorman and Sudhir Shenoy for their immense wealth of knowledge that they offered throughout the whole of this process. In addition I thank them for always being available to discuss my project and for being open to any suggestions and ideas. I would also like to thank Tomas De Oliveira for his help in the evolution of the procedure. Finally, I thank the University of Virginia for providing me with the facilities to complete this experiment

References

[1] SoftBank Robotics Europe. “NAO Software 1.14.5 Documentation.” *Testing the Speech Recognition - NAO Software 1.14.5 Documentation*, 2017, doc.aldebaran.com/1-14/software/choregraphe/tutos/speech_recognition.html.

[2] Maurer, Aaron. “NAO Robot: How To Talk With Animation.” *YouTube*, YouTube, 20 May 2018, www.youtube.com/watch?v=p9BMEM9Wmuo.

[3] Kuhn, Steven. “Prisoner's Dilemma.” *Stanford Encyclopedia of Philosophy*, Stanford University, 2 Apr. 2019, plato.stanford.edu/entries/prisoner-dilemma/.

[4] Learning, Lumen. “Microeconomics.” *Lumen*, 5 Mar. 2016, courses.lumenlearning.com/wm-microeconomics/chapter/prisoners-dilemma/.

[5] Schaefer, Kristin. (2016). Measuring Trust in Human Robot Interactions: Development of the “Trust Perception Scale-HRI”. 10.1007/978-1-4899-7668-0_10.

[6] Jeremy A. Marvel, Shelly Bagchi, Megan Zimmerman, and Brian Antonishek. 2020. Towards Effective Interface Designs for Collaborative HRI in Manufacturing: Metrics and Measures. Trans. Hum.-Robot Interact. 9, 4, Article 25 (May 2020), 55 pages.