Investigating the Effect of Prior Success on Trust in Robots

1st Reza Torbati
Department of Computer Science
Georgia Institute of Technology
Atlanta, Georgia
rtorbati3@gatech.edu

2nd Nicholas Cich
Department of Computer Science
Georgia Institute of Technology
Atlanta, Georgia
ncich3@gatech.edu

3rd Nimisha Pabbichetty
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia
npabbichetty3@gatech.edu

***Abstract*—Maintaining the correct level of trust in a robot during an human-robot interaction is vital to the success of the team. As such, finding ways to control the level of trust people place in a robot is an important problem. We use two key incites to analyze this problem. First, it has been established through psychology research that prior success will affect people's performance in subsequent tasks. Second, the primary indicator of trust that people place in robots in based on the performance of the robot. Therefore, we hypothesize that if users are given fake performance scores while interacting with a robot, we will be able to control their perception of the robot's performance and thus control their level of trust in the robot.**

***Index Terms*—HRI, trust, performance, prior success**

I. Introduction

The level of trust that a human places in a robot is an important factor in how well the team will perform. If the human places too little trust in the robot then they may not allow the robot to perform to the best of its capabilities. However, if the human places too much trust in the robot then they may over-rely on the robot which can also lead to lower team performance [3]. Therefore, for the team to perform optimally, the amount of trust a human has in a robot needs to be controlled. However, it is not clear how this can be done.

In an effort to establish a method to control the level of trust a human has in a robot, we combined two ideas: one from psychology and one from robotics. In psychology, it has been established that if a person has succeeded on a task in the past then they will typically outperform people who has failed on the task in subsequent tasks [2]. Additionally, fake percentiles are an established method to make people think they succeeded or failed at a task [4].

We combined these facts with the fact that performance of a robot is the primary indicator of how much a person will trust it [3] to create our user study. In our study, we created three practice rounds to familiarize participants with our robot. At the end of each practice round, we gave the participants a percentile score based on how quickly they completed the round. However, the score was fake and depended on which group we assigned the participant at the beginning of the study. There was a success group that got a high percentile score for each practice round, a failure group that got a low percentile score for each practice round, and a control group that got no score after completing each practice round. We then administered a trust survey from [5] to find how much trust the participants had in the robot after the practice rounds. Finally, we conducted a final round where the participants were scored based on how many tasks they could complete. We also created subroutines that the participant could have the robot execute autonomously to speed up the task completion.

These are the 3 hypotheses that we formulated for this study:

- **Hypothesis 1:** Users that succeed in the practice rounds will have more trust in Vector than either other group.
- **Hypothesis 2:** Users that succeed in the practice rounds will use Vector's autonomous capabilities in the final round more than either other group.
- **Hypothesis 3:** Users that always succeed in the practice rounds will score higher in the final round than the users that always fail in the practice rounds.

II. Related Works

Our study is motivated by two key findings. First, [2] is a study from psychology that demonstrated the effects of prior success and failure on the performance of a person in subsequent tasks. In the study, subjects are split into two groups and asked to take an initial test. The test was easy for the first group while it was challenging for the second group, which led to the first group succeeding and the second group failing at the test. Then, the two groups of subjects took the same test. It was observed that the subjects who took the easy exam outperformed those who took up the challenging exam.

This premise was further extended in [4] where the researchers assigned participants fake percentiles after performing a task to make them believe they succeeded or failed regardless of their actual performance on the task. The researchers observed that the fake percentiles significantly influenced the behavior of the participants.

On the robotics side, [3] conducted an in-depth literature review that showed that the primary indicator of people's trust in robots depends on the performance of the robot. We look to extend this idea to find if we can make participants believe that the robot performed better or worse by combining the ideas from [4] and [3] where we give participants fake percentiles

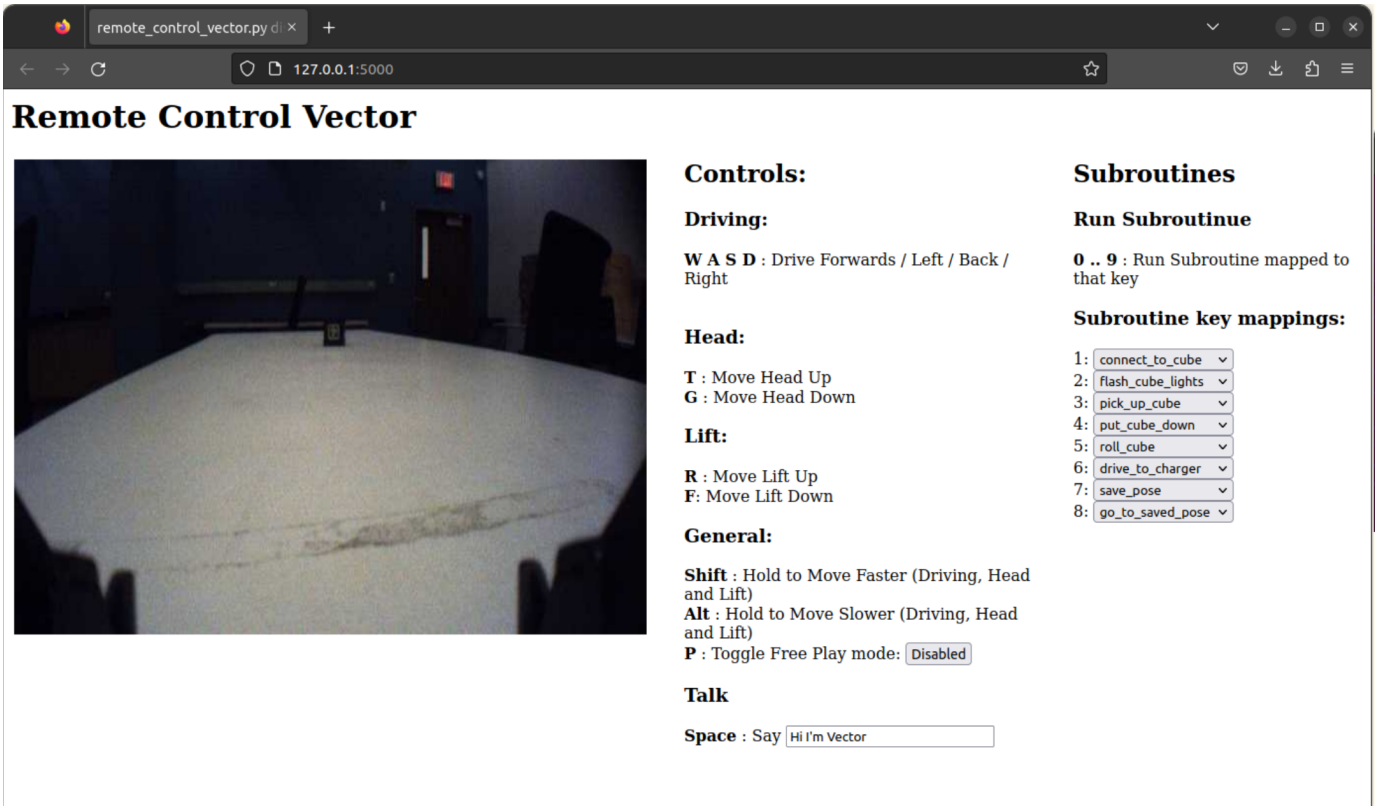

Fig. 1. Our user interface for Vector

ranking their ability to control a robot to see if this will affect their trust in it.

The study described in [1] was the main inspiration for the design of our study. They investigated the effects of drops in reliability on the trust that users placed in autonomous robots. This was quantified by measuring the amount of time the participants spent manually controlling the robot as opposed to using its autonomous driving capabilities. Lack of trust in the robot was characterised by increased manual control. This was the primary motivation for why we measured how often participants utilized our robot's autonomous subroutines.

III. Approach

We recruited 25 participants for our user study to examine how prior success can influence an individual's perception of a robot. Our general approach consisted of two stages: three practice rounds and a final round. In every stage the participants interacted with the Vector robot created by Anki with a custom user-interface as seen in Fig. 1. The participant always has the option to manually control Vector (with W A S D to drive, T G to move Vector's head, and R F to move Vector's lift) or can use subroutines to accomplish tasks such as picking up a cube or driving to its charger.

A. Practice Rounds

Before starting the first practice round, the participants were randomly assigned to one of three groups: group A, group B, and group C. The practice rounds consisted of the same task for each participant group: the first practice round required the participant to drive to a series of locations, the second practice round required the participant to pick up Vector's cube and put it down somewhere else, and the third practice round required the participant to drive Vector to a series of locations with its cube.

The practice rounds had two purposes. The first is to allow the participant to get familiar with using Vector's user interface before the final round. The second purpose was to bias the participant based on their group. For each practice round, the researcher told the participant that they had a limited amount of time to complete the task (2 minutes, 2 minutes, and 5 minutes for round 1, round 2, and round 3 respectively). After the participant finished each round, the researcher pretended to type their score into a database and then told the participant a fake percentile based on how quickly they finished the task as detailed below:

Group A (The success group):

- After round 1, the researcher told the participant "you reached a percentile rank of 90. This means you completed the round faster than 90% of the other participants".
- After round 2, the researcher told the participant "you reached a percentile rank of 92. This means you completed the round faster than 92% of the other participants".

- After round 3, the researcher told the participant "you reached a percentile rank of 88. This means you completed the round faster than 88% of the other participants".

Group B (The failure group):

- After round 1, the researcher told the participant "you reached a percentile rank of 10. This means you completed the round slower than 90% of the other participants".
- After round 2, the researcher told the participant "you reached a percentile rank of 8. This means you completed the round slower than 92% of the other participants".
- After round 3, the researcher told the participant "you reached a percentile rank of 12. This means you completed the round slower than 88% of the other participants".

Group C (The control group):

- The participant was not told a percentile after any practice round.

The theory behind this was to make the users believe that they failed or succeeded after each round in the same way as [4]. From the findings in [2], we hypothesize that this will make the participant believe that Vector will perform better or worse in the future depending on if the participant was in group A or group B which will lead to a corresponding increase or decrease in the participant's trust in Vector. To measure how much the users trusted Vector after the practice rounds, we first administered the 14 question trust survey from [5] which we used to evaluate Hypothesis 1. We then conducted the final round.

B. Final Round

The final round consisted of a series of tasks similar to the practice rounds where the participants got points for each task they successfully completed. Like the practice rounds, the tasks were designed so that they could be completed with the manual controls or with the subroutines. Examples of the tasks include "Roll the cube three times" or "Drive to the charger". We also gave the user a series of multiplication questions that they could solve for a small amount of extra points.

Taking inspiration from [1], we believe that participants that trust Vector more will utilize its subroutines more which forms the inspiration for Hypothesis 2. Finally, because participants that use Vector's subroutines will have more time to solve math questions, we created Hypothesis 3.

IV. Data Collection

While no data was recorded during the three practice rounds, after their completion, participants were given a 14 question trust survey that has been shown to accurately quantify an individual's trust in a robot [5]. The authors of [5] detail how to aggregate the Likert scale responses to each of the 14 questions into a single trust metric, which we perform data analysis on.

In the final round, we additionally collected three other data points for each participant: percentage of time spent

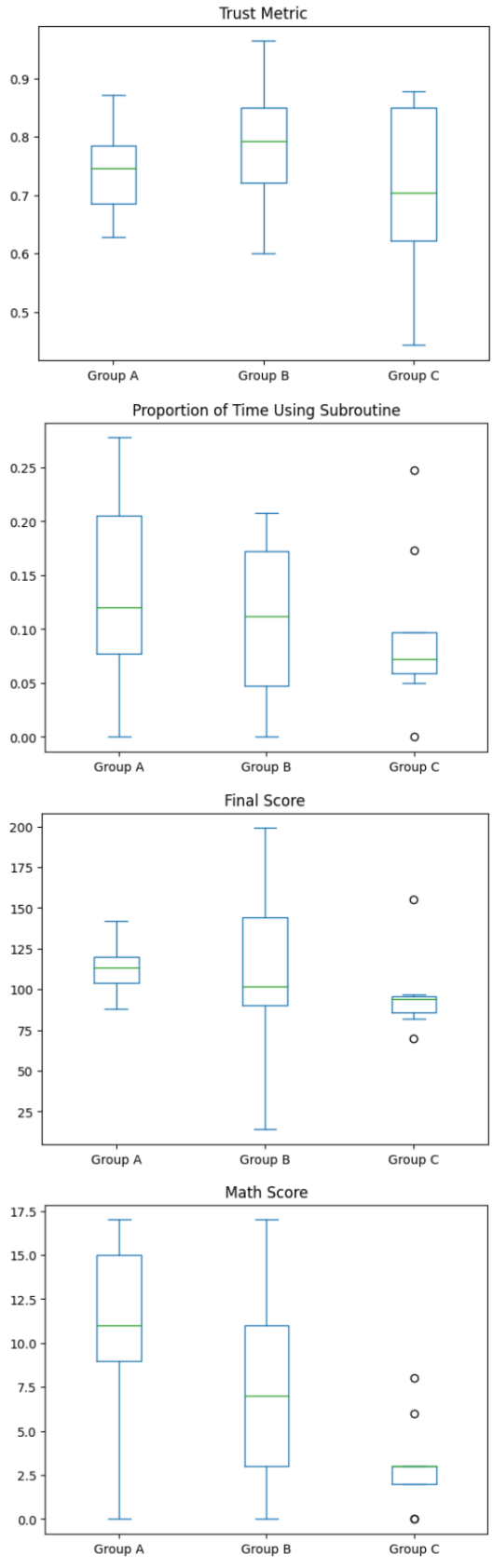

Fig. 2. Box plots showing the collected data for our four key metrics. These include level of trust (top left), percentage of time using autonomous capabilities (top right), final score (bottom left), and number of math problems solved (bottom right).

autonomously controlling the robot, the final score, and the number of multiplication questions solved. The first and third values were calculated automatically by our user interface based on user input, while the second was recorded by the researcher conducting the experiment. The first two metrics, as well as the trust metric, were the primary metrics that we used for testing our hypotheses. Box plots showing the recorded data for each of the three groups can be found in Figure 2.

V. Results

We began by verifying that our data was normal by using `scipy.stats.normaltest`. In all cases, there was insufficient evidence to reject the null hypothesis that the given data was drawn from a normal distribution. We then began to test each of our three hypotheses. For each hypothesis, we ran a one-way ANOVA on the data for the three groups. The p-values found are shown below:

Hypothesis 1	Hypothesis 2	Hypothesis 3
0.57	0.68	0.59

As shown, there was insufficient evidence to reject any of our null hypotheses regarding differences in the means between the three groups. However, when running a one-way ANOVA test on the number of math problems solved for each group, we did see that there was a statistically significant difference in the group means ($p \approx 0.025$). Running a pairwise Tukey Test on the number of math problems solved for each group, we found the following p-values for comparing the mean number of multiplication questions solved:

Groups A and B	Groups A and C	Groups B and C
0.6181	0.0225	0.1303

Thus, there was statistically significant evidence ($p \approx 0.02$) to show that the mean number of math problems solved by individuals in Group A was higher than the mean number of math problems solved by individuals in Group C. As such, while we were unable to find statistically significant evidence to support any of our hypotheses, we did find an interesting insight that suggests that the fake percentile scores did affect how the participants interacted with Vector in the final round, just not in the ways we predicted.

VI. Limitations and Future Work

We believe that there were several issues contributing to the lack of statistical significance of our results. As we were conducting the study, we realized that the feedback we were giving to participants in the practice round was being overshadowed by other factors, preventing us from being able to properly observe its effects.

For instance, we were able to observe that the effects of our feedback were not consistent between users. Many users stated that the percentiles they were being told biased their use of Vector, but not always in the way we were expecting. Some users said that negative feedback made them want to use Vector's subroutines more because they didn't trust themselves, while others said that positive feedback reinforced their initial behaviour. For example, when a user from Group A started out manually operating Vector, the fact that they were scoring so well in the practice rounds just encouraged them to continue manually operating Vector rather than try out the subroutines.

Furthermore, the variance between users and their behaviours was very high. One user reported that they did not use the subroutines at all because they wanted the challenge of operating Vector manually the entire time, while others were frantically solving math problems in an attempt to get the highest possible score. Some were just not interested in solving the math questions and did not attempt to complete even one. Given the small number of participants within each group, these type of outliers had a significant effect on our results.

With these insights in mind, we were able to formulate some potential improvements that we would implement in future iterations of this study. The first change would be to use less aggressive percentiles during the practice rounds. The current set of percentiles seemed quite harsh and may have also impacted our credibility, as some users later told us that they did think that the percentiles seemed a bit fake. Additionally, the fact that we were giving such extreme percentiles three separate times could also have been a part of the problem (especially in the case of a Group B participant who objectively was performing quite well, but were repeatedly being told that they performed so poorly).

The second change would be to record how often Vector was malfunctioning during the practice rounds. Anecdotally, Vector malfunctioned in some manner during approximately one quarter of the trials, and seemed to stop working especially often for participants in Group A. This undoubtedly had an effect on user trust, and recording the malfunctions would've allowed us to do further post-hoc analysis.

The final change would be to refine the script we followed while conducting the study. We did not explicitly explain how each subroutine worked when introducing them to the UI, as we wanted to maintain a neutral position and not bias them into using a subroutine as a result of a thorough explanation. However, this resulted in some users not realising that certain subroutines even existed, so they were not able to use them during the study. The script was more of an outline and did not mandate what needed to be said word for word except for the task descriptions, so it took a few tries to make sure we didn't forget anything and overall ensure a smoother experience.

Overall, we believe that this was a valuable pilot study that highlighted many key issues with our approach. We are confident that the insights we have found would allow us to design a better, full-fledged study with a larger number of participants.

REFERENCES

[1] M. Desai et al., "Effects of changing reliability on trust of robot systems," 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boston, MA, USA, 2012, pp. 73-80, doi: 10.1145/2157689.2157702.

[2] Feather, Norman T. "Effects of prior success and failure on expectations of success and subsequent performance." Journal of personality and social psychology 3.3 (1966): 287.

[3] Hancock, Peter A., et al. "A meta-analysis of factors affecting trust in human-robot interaction." Human factors 53.5 (2011): 517-527.

[4] Grinschgl, S., Meyerhoff, H.S., Schwan, S. et al. From metacognitive beliefs to strategy selection: does fake performance feedback influence cognitive offloading?. Psychological Research 85, 2654–2666 (2021). https://doi.org/10.1007/s00426-020-01435-9

[5] Schaefer, K.E. (2016). Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI". In: Mittu, R., Sofge, D., Wagner, A., Lawless, W. (eds) Robust Intelligence and Trust in Autonomous Systems. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7668-0_10