Robot Confidence And Trust Alignment

Poornima Kaniarasu[1], Aaron Steinfeld[1], Munjal Desai[2], and Holly Yanco[2]
[1]Carnegie Mellon University, Pittsburgh, PA 15213
{kpoornima, steinfeld}@cmu.edu
[2]University of Massachusetts Lowell, Lowell, MA 01854
{mdesai, holly}@cs.uml.edu

ABSTRACT
Trust in automation plays a crucial role in human-robot interaction and usually varies during interactions. In scenarios of shared control, the ideal pattern is for the user's real-time trust in the robot to align with robot performance. This should lead to an increased overall efficiency of the system by limiting under-trust and over-trust. However, users sometimes display incorrect trust and the ability to detect and alter user trust is important. This paper describes measures for real-time trust alignment.

Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Human factors
General Terms - Experimentation, performance
Keywords - Trust, automation, experiments, robot confidence

I. INTRODUCTION

For safety concerns and maximum utilization of robot abilities, the user's trust in a robot should be aligned with its competence. The main objective of our work is to model a person's current level of trust in a robot system so in order to design interfaces and robot behaviors that foster appropriate levels of trust [2]. We are particularly interested in mobile robots designed for non-social interaction. These robots are likely to be task-oriented and therefore used for time-dependent activities, capable of damaging objects and hurting people, and unable to express their intent to bystanders (e.g., [6]).

There are numerous studies in identifying factors affecting trust (e.g., [4-5]). In [3] we suggested real-time factors that could potentially allow robots to recognize that a human has lost trust, thereby allowing the robot to adjust its behavior accordingly. But all effective communications are two-way interactions. Therefore, in this work we examine if the robot conveying information back to the human about internal states to help the user adjust trust accordingly. In short, we provide the user with a real-time feedback of the robot's confidence and analyze the user's trust on the robot system in real-time.

II. METHOD

For this paper, we use data collected from a recent experiment. Details about the robot, experimental design, interface (figure 1) and data collection methods can be found in Desai et al. [2]. This study collected real-time trust from the user every 20 seconds as the robot traversed a slalom course under varying reliability. At every trust prompt user was asked to press buttons corresponding to trust increasing (↑), decreasing (↓) or remaining the same (↔). The robot reliability changed between high and low through the course and users were allowed to dynamically shift between a fully autonomous mode and a robot-assisted shared mode as much as they wished during each run. The experiment was designed to have high workload and a performance-driven financial bonus. These implicitly encouraged use of the autonomous mode.

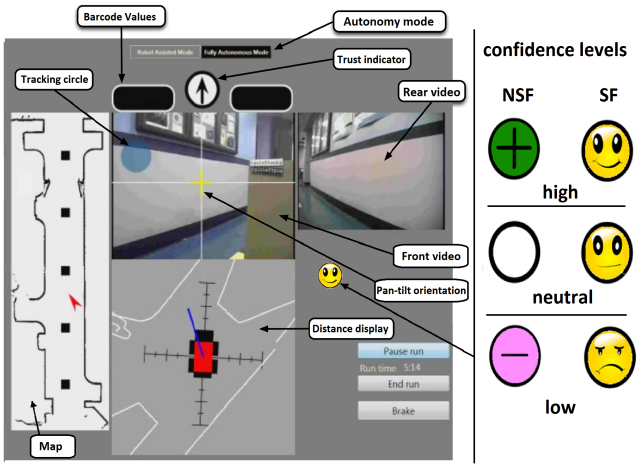

Figure 1: The user interface used to control the robot, displays the confidence indicator right below back video feed (is on the left). The right side shows 3 levels of confidence indicators for both semantic (SF) and non-semantic feedback (NSF).

We examined data from a between subject study for 3 conditions, one with semantic feedback, one with non-semantic feedback and one without feedback. Feedback was provided via a robot confidence indicator in the user interface. The confidence indicator expressed 3 levels of confidence (Figure 1). The robot expressed low confidence when it was in a low reliability (LR) zone, neutral when transitioning between high and low reliability, and high confidence when in high reliability (HR) zone.

To quantify alignment we define two measures (Figure 2):

1. *Trust Mismatch (TM):* The degree of alignment of user's trust with the robot's current reliability. Each time the user is prompted for a trust input, if the robot's performance is high (HR) and user's trust decreases or the robot's performance is low (LR) and trust increases then prompt mismatch is labeled 1 (red arrow in figure). It is labeled 0 (green arrow in figure) for every other case. TM is calculated by summing up the prompt mismatch over the whole run i.e. TM is

calculated by summing up the number of inappropriate trust shifts (count of red arrows) along the course. High TM implies that user's trust is frequently out of phase with robot performance.

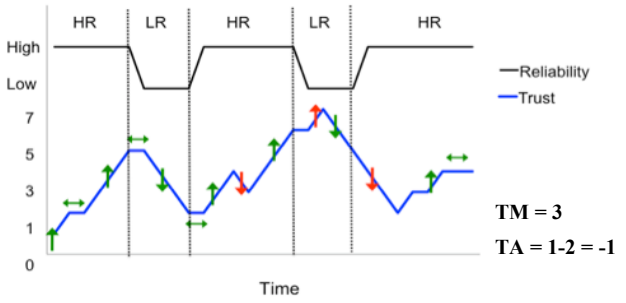

Figure 2: The timeline of the whole run is split into regions of high reliability (HR) and low reliability (LR). The graph shows trust as a function of time. An arrow on it represents a trust prompt. Green arrows are appropriate trust shifts and red arrows are inappropriate trust shifts.

2. *Trust Alignment (TA):* To analyze if people generally under-trusted or over-trusted the robot we incorporated the direction of mismatch into this measure. An increase in trust in LR region was scored +1 and decrease in trust in HR region was scored -1. TA is calculated by summing up the scores. In other words, from figure:

   TA = Count of red↑s – Count of red↓s

   Zero TA would mean trust was in phase with robot's performance, positive values imply over-trust and negative values imply under-trust.

III. RESULTS

A. Effect Of Feedback On Trust Metrics:

It was hypothesized that providing the user with more information on internal states of the robot would help them anticipate robot failures, adjust their trust accordingly and allocate control thus minimizing errors. ANOVA analysis across the 3 feedback conditions viz. no feedback, with SF and with NSF gave a significant main effect $F(2,108)=5.19$, $p < 0.01$. Post hoc analysis using Student's-t indicated that TM values for no feedback condition ($\mu=5.31$, $\sigma=0.59$) were significantly higher (higher values indicate less trust alignment) than the condition with SF ($\mu=3.12$, $\sigma=0.73$) and the condition with NSF ($\mu=2.15$, $\sigma=0.72$). These results align with our hypothesis.

ANOVA analysis of TA showed a significant main effect of $F(2,138)=3.62$, $p < 0.05$ for the feedback conditions. Pair-wise comparisons using Student's-t analysis showed no feedback condition ($\mu=0.15$, $\sigma=0.04$) was higher, whereas both the conditions with feedback SF ($\mu=0.02$, $\sigma=0.05$) and NSF ($\mu=0.008$, $\sigma=0.05$) were significantly lower in magnitude and approximately zero. This implies that without feedback people over-trusted the robot. Correlation analysis for TA with Muir survey results [1] collected after every run strengthens this finding as TA was significantly and positively correlated (0.220, $p < 0.05$) with trust rating from the Muir survey questions [1].

B. Autonomy Preference

Post every run, participants were asked which autonomy mode they preferred for that run. ANOVA analysis showed that TM had a significant main effect for mode preference $F(1,109)=5.09$, $p < 0.05$. People who preferred autonomous mode ($\mu=3.15$, $\sigma=0.51$) had significantly more appropriate levels of trust (lesser TM values) than those who preferred robot-assisted mode ($\mu=5$, $\sigma=0.62$). This may be due to the fact that people who preferred assisted mode for a run were driving manually for most of the run and were not able to effectively judge the reliability of the robot autonomy.

IV. CONCLUSIONS

The results from analysis of TM across feedback conditions shows that as people could detect when the robot dropped down in performance and when it would recover back, they tuned their trust accordingly. This reaffirms our earlier findings [2] that error warnings decrease when users are provided with feedback. The results were not significant for the modality of the confidence indicators (SF & NSF). Hence we can observe that modality of feedback probably does not impact the alignment of trust.

The TA analysis indicates that people generally trust automation more than they should in absence of feedback, (i.e.) when they have little knowledge of internal states of the robot. Thus, introduction of confidence indicators or other reliability feedback can minimize accidents due to robot malfunction.

V. ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation (IIS-0905228 and IIS-0905148). Thanks to the larger team for all the hard work behind the design and execution of the source study.

VI. REFERENCES

[1] B. M. Muir, "Operators' trust in and use of automatic controllers in a supervisory process control task." Doctoral Dissertation, University of Toronto, Canada, 1983.

[2] M. Desai, P. Kaniarasu, C. Medvedev, A. Steinfeld, and H. Yanco. Effects of changing reliability on trust of robot systems. In *HRI '13: Proceedings of the 8th ACM/IEEE Int'l Conference on Human Robot Interaction*, 2013.

[3] P. Kaniarasu, A. Steinfeld, M. Desai and H. Yanco. Potential measures for detecting trust changes. In *HRI '12: Proceedings of the 7th ACM/IEEE Int'l Conference on Human Robot Interaction*, 2012.

[4] A. Freedy, E. DeVisser, and G. Weltman. Measurement of trust in human-robot collaboration. *Collaborative Technologies and Systems,* Jan 2007.

[5] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(5): 517-527, 2011.

[6] B. Mutlu and J. Forlizzi. Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *HRI '08: Proceedings of the 3rd ACM/IEEE Int'l Conf.erence on Human Robot Interaction,* pages 287–294, 2008