Potential Measures for Detecting Trust Changes

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
It is challenging to quantitatively measure a user's trust in a robot system using traditional survey methods due to their invasiveness and tendency to disrupt the flow of operation. Therefore, we analyzed data from an existing experiment to identify measures which (1) have face validity for measuring trust and (2) align with the collected post-run trust measures. Two measures are promising as real-time indications of a drop in trust. The first is the time between the most recent warning and when the participant reduces the robot's autonomy level. The second is the number of warnings prior to the reduction of the autonomy level.

Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Human factors

General Terms
Experimentation, performance

Keywords
Trust, automation, experiments

1. INTRODUCTION
One of the key factors for the acceptance and safe deployment of robots is the degree to which a user trusts the robot. A main goal of our work is to model a person's current level of trust in a robot system so that it can be used to design robot interfaces and behaviors that foster appropriate levels of trust [2]. Mobile robots, especially those not designed for social interaction, are particularly interesting since they 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 is an increasing amount of work exploring what factors are important for human trust in robots (e.g., [1-4]). In this work, we are explicitly examining two measures as potential real-time indicators that an operator has reduced their trust of a task-oriented robot. Specifically, we are focused on measures that can be internally observed by a robot without the use of traditional trust surveys. With these factors, on-line techniques will allow robots to recognize the human has lost trust and act accordingly.

2. METHOD
Our approach is to analyze data collected under an existing study. Details about the robot, interface, measures, experimental design, and results can be found in Desai et al. [2]. The study investigated how changing the robot's reliability influences people's use of robot autonomy and their trust in the robot system through experiments with participants operating a real robot through a slalom course. The experiment was designed to have a high workload so that the participants would need to use the autonomous capabilities of the robot in order to complete the task in time and to be able to complete the secondary task. We hypothesized that people would trust a robot system less when its reliability in autonomous mode decreased, switching to a shared mode. Therefore, the robot's reliability fluctuated between high and low levels during the experiment.

For this work, we examined the three conditions where users encountered reliability changes. These changes consisted of controlled dips in reliability based on pre-set locations (Figure 1). Audio messages indicating the cumulative number of wrong turns during the run and their penalties were issued as warnings to participants immediately after each navigation error occurred. These conditions were counterbalanced within the larger experiment. One of the measures collected at the end of each run was an autonomy trust survey developed by Jian, et al. [5].

Two potential measures were identified as possible indicators for a loss of trust. These were:

1. Time Between Events (*TBE*): The time between two events, whether a warning or a switch. This measure gives the time between two consecutive warnings (e.g., TBE = t_8-t_6 in Figure 2) or, in the case of a warned switch, the time between the most recent warning and a participant changing the autonomy level (TBE = t_3-t_2, in Figure 2).

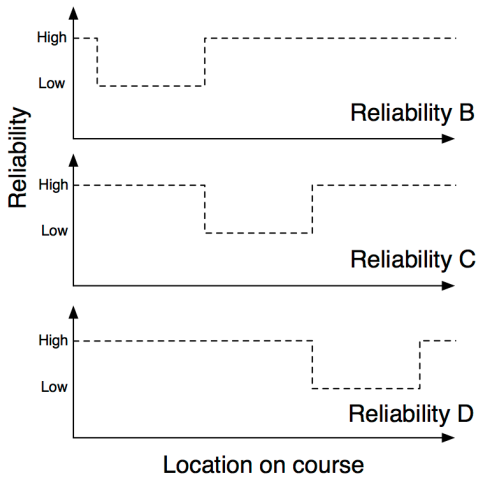

Figure 1. Reliability patterns for the analyzed conditions

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.
ACM 978-1-4503-1063-9/12/03.

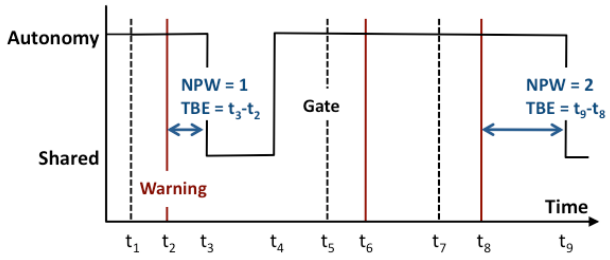

Figure 2. Timeline graph of the metrics TBE and NPW

2. Number of Prior Warnings (*NPW*): How many times the participant was warned prior to a switch. Every time a switch occurs, the NPW counter is reset to zero. If there is a warning but no switch, it accumulates (e.g., t_9).

3. RESULTS

3.1 Switching

We looked at how warnings impacted switching, which revealed that most of the switches were not due to a prior warning (Table 1). Also, the timing of when reliability dropped appeared to have an impact on switching behavior. This result aligns with other analyses by the team on switching behavior which showed that reliability drops in the middle of the run (C) led to increases in the number of mode switches. Additionally, switches away from autonomy were twice as slow as those seen for early drops in reliability (B) [2].

3.2 Warned switches

Next, we focused only on autonomy switches that were spurred by a warning. We also omitted switches where a user made a turn between the most recent warning and the switch since the warning would have been for the preceding turn. Participants could switch in two directions, either into or out of autonomy. ANOVA models were run for the Direction (Autonomy, Shared) and Reliability (B, C, D) on TBE and NPW. There was a significant interaction for TBE ($F(2, 26) = 4.3$, $p < 0.05$) where TBE was significantly higher for Auto-B than other combinations, showing that people took longer to recognize a reliability drop when it occurred near the start of the robot's use. There was also a marginal result for the main effect of Reliability ($F(2, 26) = 2.8$, $p < 0.1$) with reliability B being higher than the other two conditions. A quick power calculation revealed that this would likely become significant with a few more samples. This result aligned with a significant main effect for Reliability on NPW ($F(2, 26) = 5.7$, $p < 0.01$) where reliability B was higher than the other two conditions. No other effects were significant.

Correlation analyses were run for the two measures, the post-condition trust survey, and the percentage of time spent in autonomous mode during the condition. There were significant correlations for: NPW and TBE (0.475, $p < 0.05$), NPW and trust (-0.40, $p < 0.05$), and NPW and percent time using full autonomy (0.42, $p < 0.05$). In other words, for participants who had switched as a result of a warning, these results show that, (1) participants who accumulated more warnings responded slower than their peers to the most recent warning, (2) participants who accumulated more warnings had lower trust, and (3) participants who accumulated more warnings used more autonomy. While this last result seems to conflict with the second, we hypothesize that, when faced with a system that is repeatedly failing, users cannot distinguish between the robot's failures and their own, leading them to accept the easier control method. This result stresses the need to model a person's trust in a robot system, in order to find ways to encourage people to use the best autonomy mode for the current situation.

Table 1. Switching behavior

Condition	Warned	N (% of subtotal)
B	No	114
	Yes	11 (8.8%)
C	No	164
	Yes	17 (9.4%)
D	No	83
	Yes	11 (11.7%)

4. CONCLUSIONS

The impact of warnings on switching behavior seems to reinforce the earlier findings in Desai et al. [2] that early drops in reliability incur different behavior than middle or late reliability drops.

The correlation results have strong face validity. In the face of repeated alarms, a user will stop attending to them as quickly (NPW & TBE). Likewise, repeated alarms imply a system is untrustworthy. Therefore, Number of Prior Warnings and the user's response time to the warnings have the potential as real-time measures of a drop in trust.

5. 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.

6. REFERENCES

[1] I. Dassonville, D. Jolly, and A. Desodt. Trust between man and machine in a teleoperation system. *Reliability Engineering & Systems Safety,* 53(3):319–325, 1996.

[2] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco. Effects of changing reliability on trust of robot systems. In *HRI '12: Proceedings of the 7th ACM/IEEE Int'l Conference on Human Robot Interaction*, 2012.

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

[4] 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.

[5] J. Jian, A. Bisantz, and C. Drury. Foundations for an empirically determined scale of trust in automated systems. *Int'l Journal of Cognitive Ergonomics,* 4(1):53–71, 2000.

[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.