Impact of Robot Failures and Feedback on Real-Time Trust

Munjal Desai[1], Poornima Kaniarasu[2], Mikhail Medvedev[1], Aaron Steinfeld[2], and Holly Yanco[1]
[1]University of Massachusetts Lowell
email: {mdesai, mmedvede, holly}@cs.uml.edu
[2]Carnegie Mellon University
email: {kpoornima, steinfeld}@cmu.edu

***Abstract*—Prior work in human trust of autonomous robots suggests the timing of reliability drops impact trust and control allocation strategies. However, trust is traditionally measured post-run, thereby masking the real-time changes in trust, reducing sensitivity to factors like inertia, and subjecting the measure to biases like the primacy-recency effect. Likewise, little is known on how feedback of robot confidence interacts in real-time with trust and control allocation strategies. An experiment to examine these issues showed trust loss due to early reliability drops is masked in traditional post-run measures, trust demonstrates inertia, and feedback alters allocation strategies independent of trust. The implications of specific findings on development of trust models and robot design are also discussed.**

I. Introduction

During an operator's interaction with an autonomous system, a key issue is how the operator uses the available autonomy levels, often referred to as the control allocation strategy [1]. Inappropriate control allocation strategies can result in over-reliance or under-reliance on the automated system [2]. One of the known contributing factors to improper reliance on automation is trust (e.g., [3], [4]). While it is difficult to conclusively state the root cause, over-reliance or under-reliance on automated systems due to miscalibrated trust can often be inferred in incident reports from the aviation industry. For example, while using the flight management system (FMS) to navigate to Cali, Colombia, the crew of American Airlines Flight 965 entered the first few characters for their destination. Accustomed to selecting the first option, the crew selected a destination that happened to be a few miles behind them rather than the intended destination, which was not the top option in this case. The FMS turned the plane around; the plane crashed into a mountain shortly afterwards [5].

For decades, researchers in the human-automation interaction field have investigated the control allocation strategies of operators under different circumstances (e.g., [6], [7], [8]) and observed how people use, misuse, or disuse automation (e.g., [2], [9], [10]). Specifically, the influence of several factors including reliability on control allocation have been studied by several researchers; a detailed survey was conducted by Wickens and Xu [11]. Researchers have also investigated factors that influence trust and, ultimately, reliance on automated systems (e.g., [3], [4], [7]) in order to prevent accidents and improve the performance of automated systems. Factors such as self-confidence (e.g., [12], [13], [1]), reliability (e.g., [14], [15], [16]), and risk (e.g., [17], [18]) are known to impact an operator's trust of the system. Additional factors such as task complexity, workload, system accuracy have also been hypothesized as contributing factors [7]. Similar attempts to understand how robot operators trust and utilize automated behaviors of robots have been made in the field of human-robot interaction (HRI) (for a survey and analysis of recent research see [19]).

In our prior work, we examined the impact of changing reliability on an operator's trust and control allocation strategy [16]. One of the key contributions of that research was finding the impact of the timing of failures of the autonomous behaviors on operator trust and control allocation. However, one of the limitations of that research, due to the experimental methodology utilized, was the inability to examine how trust evolved during a participant's interaction with a remote robot system and how it was impacted by robot failures at the time of the failure. To investigate the evolution of trust and the impact of varying reliability on real-time trust, we modified the experimental methodology and conducted the research studies described in this paper.

While it is important to understand trust and control allocation strategies, it is equally important to find means to influence them, should the need arise. Research exists where participants were provided information about results of past decisions [20]; however, to our knowledge no research exists that investigates the impact of providing information about the automated system's confidence in its own sensors. Therefore, as part of this research, we also investigated the impact of providing feedback conveying this confidence information on trust and control allocation.

Our long term goal is to understand how different factors impact trust and control allocation and, based on this information, to build a model that can predict an operator's current level of trust so that the system can adjust in ways to increase the current level of trust to prevent inappropriate usage of the autonomy levels. Towards this end, we created a set of research questions that we needed to address:

- *Q1:* How does the timing of periods of low reliability impact real-time trust? Our prior experiments suggest trust of the robot system is influenced by whether a period of low reliability is at the beginning or end of the run (trust in a robot, as measured using a trust scale after the run is complete, drops if the robot is unreliable near the end of a run [16]). We designed this study to investigate how real-time trust is influenced by the timing of reliability drops.

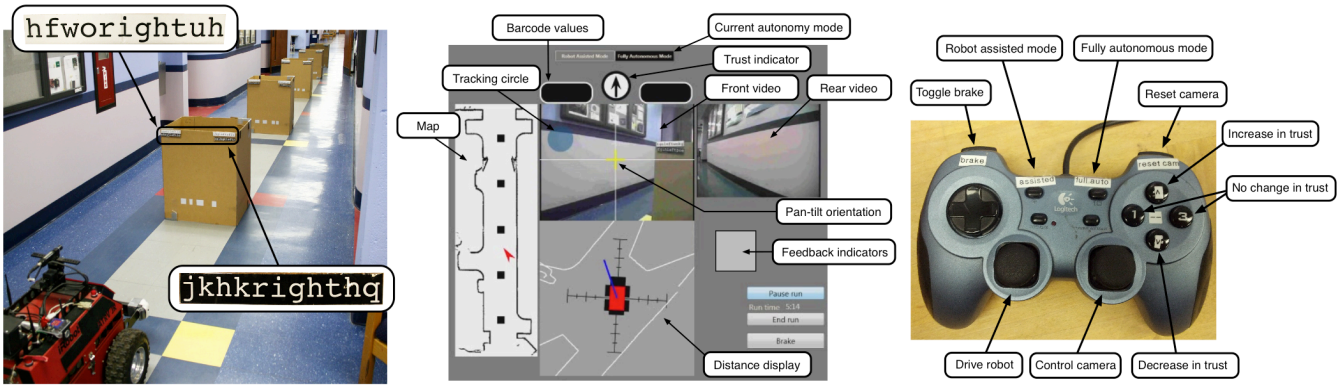

Fig. 1. *Left:* The robot (ATRV-Jr) at the starting position in the course. The boxes have navigation labels on the top and barcodes on the bottom. *Center:* The user interface used to control the robot for the DR group. The DR+F group saw an additional icon from the set of images in Figure 2, displayed where the box marked feedback indicators is shown. *Right:* The gamepad used by participants to control the different aspects of the robot.

- *Q2:* Is real-time trust influenced by feedback from the robot system? Some participants were shown confidence indicators on the interface, which we tied to the reliability drops in the system (i.e., the confidence indicator would drop before the system's reliability dropped and the indicator would rise when the reliability rose). Other participants received no feedback.
- *Q3:* Does the type of feedback matter? Our feedback included two conditions: one with semantic feedback and the other with non-semantic feedback.

The answers to these questions will allow for the design of robot systems that will be able to predict when an operator's trust level is likely to be falling and, thus, to be able to foster appropriate levels of trust in their operators.

II. Methodology

Experiments were conducted at the University of Massachusetts Lowell (UML) and Carnegie Mellon University (CMU)[1]. The goals of this experiment were to investigate the impact of varying reliability and feedback on the evolution of trust and control allocation strategy. Twelve participants were recruited for the varying reliability with no explicit feedback group at UML, henceforth referred to as the 'Dynamic Reliability' (DR) group. Of the twelve participants, six were male and six female; the mean age was 37.4 years (SD=16.3). A total of sixteen participants were recruited for the dynamic reliability with feedback group at CMU, henceforth referred to as the 'Dynamic Reliability + Feedback' (DR+F) group. Eight of the sixteen participants experienced the 'Semantic Feedback' (DR+F:S) condition and the other eight experienced the 'Non-Semantic Feedback' (DR+F:NS) condition. Of the sixteen participants for the DR+F group, eight were male and eight female and the mean age was 22.2 years (SD=4.0). All of the participants were novice users since none of them had prior experience controlling remote robots.

[1]Unless explicitly mentioned, all of the parameters were identical between the two sites.

A. Robot System

Similar iRobot ATRV-JR robots were used at both sites (Figure 1, left). The front camera was mounted on a Directed Perception PTU-D46-17 pan-tilt unit and another camera with a fixed base was mounted on the rear. A SICK LMS200 laser in the front and a Hokuyo URG-04LX laser mounted on the back were used for sensing distance. The robots had computers with similar capabilities and ran the same code base.

B. User Interface (UI)

The video from the front camera was displayed at the center of the UI (Figure 1, center) and the video from the back camera was displayed on the top right (laterally inverted to serve as a rear view mirror). The distance information from both lasers was displayed on the bottom around a graphic of the robot. The map of the course with the pose of the robot was displayed on the left. Participants could use the gamepad (Figure 1, right) to drive the robot, pan and tilt the front camera, select the autonomy modes, toggle the brakes, recenter the camera, and acknowledge the secondary task and trust prompts.

C. Modifications for the feedback condition

The participants in the DR+F group were additionally given feedback (3 levels) that indicated the confidence of the robot in its ability to read barcodes. The interface displayed the confidence indicator just below the rear camera view (Figure 1, center). The robot indicated high levels of confidence for all high reliability regions, except for one box before and one box after the low reliability region where it displayed a neutral state to ensure a gradual transition between the reliability levels. For participants who experienced semantic feedback (DR+F:S), they were shown emoticons to represent the confidence levels, whereas participants who experienced non-semantic feedback (DR+F:NS) were shown green, white and pink lights to indicate high, neutral and low level of confidence respectively. The indicators also had a plus sign for high level and a minus for low level of confidence embedded in the circle (Figure 2) to take color-blind users into consideration.

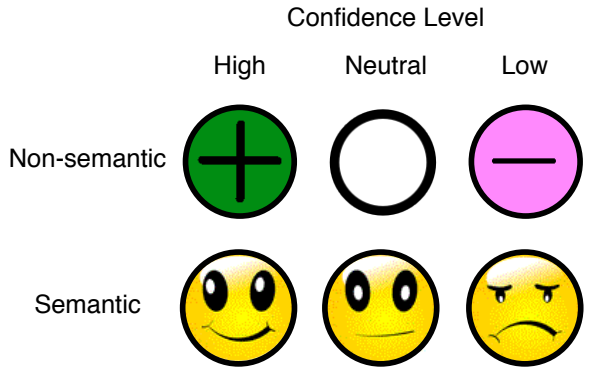

Fig. 2. Semantic and non-semantic indicators. The icons for semantic feedback had yellow backgrounds. The high confidence icon for non-semantic feedback had a green background and the low confidence icon for non-semantic feedback had a pink background.

D. Secondary task

A translucent blue circle 70 pixels in diameter was spawned every 35 seconds on the front video feed. The participants were asked to acknowledge the blue circle by moving the yellow crosshair (that indicated the pan-tilt orientation) over the blue circle. When the yellow crosshair moved over the blue circle, the blue circle disappeared, as it had been acknowledged. The location of the blue circle was always a fixed distance away from the yellow crosshair and in a random direction. The secondary task was designed to provide a consistent and regular workload, unlike the secondary task of searching for a 'victim tag' in [16], which we observed could be found accidentally rather than as a deliberate secondary task. A blue circle can be seen in Figure 1 (center).

E. Test course

A map of the course with five boxes in it is shown on the left side of the UI (Figure 1, center). A photo of the course is also shown (Figure 1, left). The courses were approximately 18 meters long and had 5 obstacles (boxes) placed about 2.75 meters from each other. The clearance on either side of the boxes was 0.9 meters, and the robot was 0.66 meters wide. The start and end positions were the same for each run. For each run, the participants were asked to follow a set path. There were five different paths based on the following criteria:

- The length of each path must be the same (72 meters).
- The number of u-turns in a path must be the same (3).
- The number of transitions from the left side of the course to the right and vice versa must be the same.

As the maps were similar in difficulty and length, we did not counterbalance paths for the participants. Instead, paths were selected based on a randomly generated sequence.

Text labels were placed on top of the boxes to indicate the path ahead (Figure 1). The labels indicated 'left,' 'right,' or 'uturn.' The directions were padded with additional characters to prevent the participants from recognizing the label without reading them. There were two types of label, ones with a white background and others with a black background. The labels with the white background (referred to as white labels) were to be followed for the first half of the entire length and the labels with black background (referred to as black labels) for the second half. The transition from following the white labels to the black labels was indicated to the participants via the UI via audible and visual cues. Two sets of labels were necessary to prevent the participants from driving in an infinite loop.

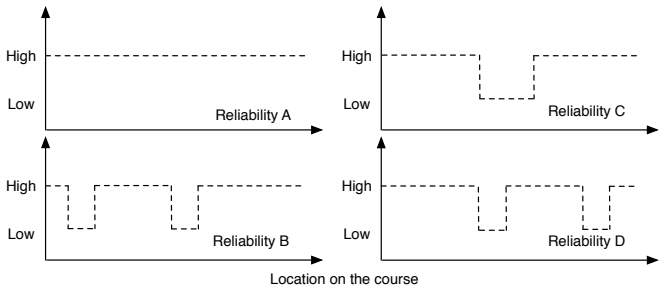

Fig. 3. Reliability conditions that the participants experienced.

The boxes also had barcodes made from retroreflective tape that the robot could read (Figure 1). The robot would display the contents of the barcode on the UI. However, the paths for each run were hard coded because the barcodes could not be consistently read by the robot. While the barcodes were not used by the robot, the participants were told that the robot read the barcodes to determine the path ahead, just like they read the labels. Based on a constant video compression rate, sampling resolution, and the font size, the labels could be read from about 1 meter away. The robot was set to simulate reading the label from approximately the same distance. The participants were informed that, at times, the robot might make a mistake reading the barcodes and that they should ensure that the direction read by the robot was correct. The participants were also told that if the robot did make a mistake in reading the barcode, it would proceed to pass the next box on the incorrect side, resulting in a lower performance score.

F. Autonomy modes

There were two autonomy modes that the participants could select. In the fully autonomous mode, the robot read labels and drove autonomously through the course while avoiding obstacles. In assisted mode, the participants had 75% control over the direction and velocity of the robot[2].The maximum speed of the robot was the same for both modes (0.2 m/s).

G. Reliability patterns

The reliability patterns used for the experiment are shown in Figure 3. The low reliability durations lasted for four boxes ($1/4^{th}$ of the course). Conditions B and D had two sets of reliability drops, each lasting for two gates, whereas condition C had one period of low reliability that lasted for four boxes. The reliability drops for B occurred early on, whereas the drops for D occurred later in the run. The drop in condition C occurred during the middle of the run. The two sets of reliability drops in B and D occurred four boxes apart. We designed the conditions to ensure that the reliability drops were

[2]75% of the final velocity was based on the user's desired velocity and the rest on the robot's desired velocity

not immediately at the beginning or at the very end of a run, so that all participants started and ended their runs with a working robot system.

H. Compensation

Using higher levels of automation reduces workload and hence is desirable, especially under heavy workload from other tasks. To prevent participants from using high levels of autonomy all of the time, regardless of the autonomous system's performance, it is typical to introduce some amount of risk. Hence, in line with similar studies (e.g., [21], [22], [4], [16]), the compensation was based in part on the overall performance. The maximum amount that the participants could earn was $30. Base compensation was $10. Another $10 was based on the performance during the runs. The last $10 was based on the time needed to compete the runs, provided that the performance on those runs was high enough.

The performance for each run was based on multiple factors, with different weights for each of those factors determined before the experiment was run. The participants were told there was a significant penalty for passing a box on the incorrect side, regardless of the autonomy mode. If the participants passed a box on the wrong side, they were heavily penalized (15 points per box). Participants were penalized 3 points per tracking task that was not acknowledged. The participants were not penalized for unacknowledged trust prompts. However, they were told that they had to acknowledge at least 90% of the trust prompts to be eligible for the time bonus.

The scoring details were not revealed to participants, although they were told about the factors that would influence their score. The score for each run was bounded between 0 and 100. If the score was 50 or more, the participants were eligible for a time bonus; if they had completed the runs in under 11:00 minutes, they would receive an additional $10. If they were eligible for the time bonus but took between 11:00 and 12:00 minutes then they would receive $8 for the time bonus, and so on. Participants were told about this interdependence between score and time, which was designed to prevent participants from quickly running through the course, ignoring the tasks, while also providing a signicant motivation to perform the task quickly.

At the end of each run, its score was calculated and the participants were informed about the amount of compensation that could be received based only on that run. The participants were also informed about the number of trust prompts that they acknowledged. At the end of five runs, the average compensation was calculated and given to the participant.

I. Questionnaires

There were three sets of questionnaires. The pre-experiment questionnaire was administered after the participants signed the consent form; it was focused on demographics (i.e., age, familiarity with technology similar to robot user interfaces, tendency towards risky behavior, etc.). The post-run questionnaire was administered immediately after each run; participants were asked to rate their performance, the robot's performance, and the likelihood of not receiving their milestone payment. Participants were also asked to fill out the Muir trust questionnaire [3] and a TLX questionnaire [23][3]. After the last post-run questionnaire, the post-experiment questionnaire was administered, which included questions about wanting to use the robot again and its performance.

J. Real-time trust

Trust questionnaires, such as the Muir questionnaire [3], only provide information about the participant's trust of the robot at the end of each run. In order to examine how trust of the robot is immediately impacted by changes in reliability, participants were asked to respond to prompts. At each prompt, participants were instructed to indicate if their trust of the robot had increased, decreased, or had not changed by pressing buttons on the gamepad (Figure 1, right). Participants were prompted for this trust measure every 25 seconds. We selected a gap of 25 seconds to ensure that participants were not overwhelmed, but that, at the same time, there would be at least one trust prompt between consecutive boxes (which we call gates). When the trust prompts were triggered, the trust indicator circle shown in Figure 1 turned red and an audible beep was sounded. The trust prompt indicator would stay red until the participant recorded his or her trust level. When one of the buttons was pressed, the trust prompt indicator would show an up arrow, down arrow, or a sideways arrow indicating increase, decrease, or no change in trust, respectively.

K. Procedure

The participants started by filling out the demographic questionnaire. There were two trial runs, one in autonomous and one in assisted mode, where the participants were provided help with controls and got comfortable with the robot. Then there were five runs, each following a different map. The first run was always in high reliability (A). The ensuing four runs were counterbalanced using superimposed Latin squares of reliability modes (A, B, C & D) to reduce ordering effects. Participants could switch between the two autonomy modes any number of times they wanted during these five runs.

III. Results and Discussion

Apart from minor differences noted above, the underlying structure of the two groups (DR and DR+F) was similar and similar behavior was observed across both groups. For this reason, the data is reported in aggregate when appropriate and, when differences between the two were observed, these differences are highlighted.

Data from the practice and baseline runs were not included in the analyses. We checked for practice effects (run order) and map effects and did not find any issues. This lack of significant differences for run and map effect suggests the counterbalancing and map designs were adequate.

[3]Due to space limitations the workload data is not reported.

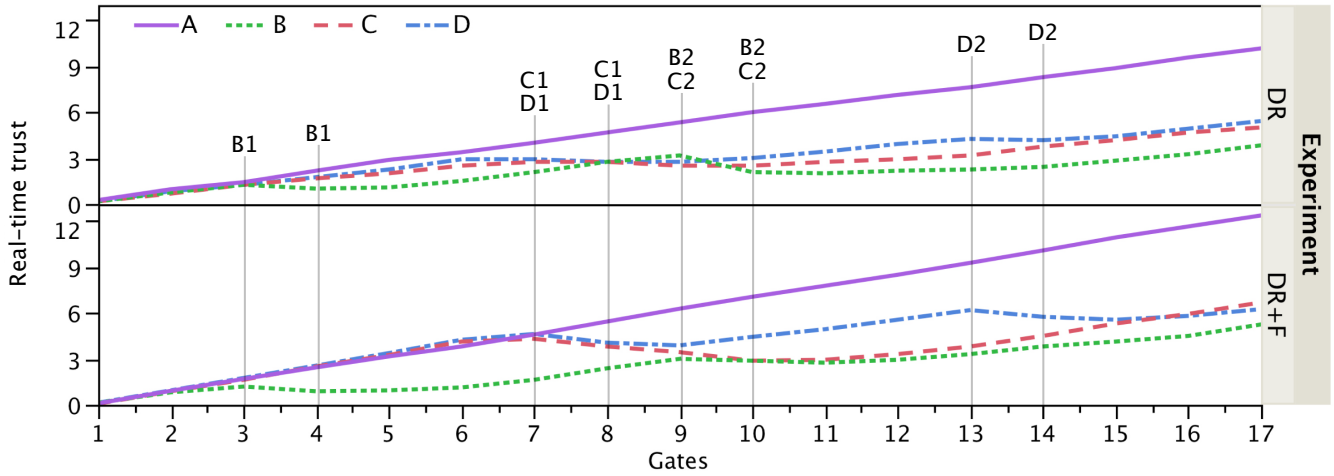

Fig. 4. The evolution of trust. The graph shows the average real-time trust ratings for the two groups. The four line types represent the four reliability levels. The vertical lines indicate the beginning of a reliability drop for one box during a specific reliability condition indicated by the letter. For example, B1 indicates the first reliability drop for Reliability B.

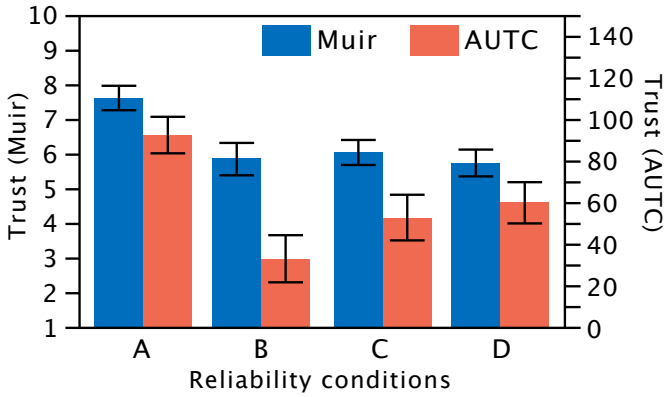

Fig. 5. Trust ratings from the Muir questionnaire (blue; left bar in each pair) and the area under the trust curve (AUTC) (red; right bar in each pair). The mean values are shown along with ± 1 standard error. The scale for the left bars (Muir) is on the left y-axis and the scale for the right bars (AUTC) is on the right y-axis.

A. *Effect of reliability changes on trust*

1) Muir trust questionnaire: To analyze the impact of reliability drops on participants' trust of the robot, we conducted a two-way analysis of variance for trust that yielded a significant main effect of Reliability, $F(3,103)=4.49$, $p<0.01$. However, the main effect of Group type (DR *vs* DR+F), $F(1,103)=0.05$, $p=0.81$ and the interaction of Reliability and Group type was non-significant, $F(3,103)=0.24$, $p=0.86$. Post hoc comparisons for Reliability using Tukey's HSD test indicated that the trust values for Reliability A ($\mu=7.59$, $\sigma=1.82$) were significantly higher (higher values indicate more trust) than Reliability B ($\mu=5.83$, $\sigma=2.34$, $p<0.05$), C ($\mu=5.97$, $\sigma=2.03$, $p<0.05$), and D ($\mu=5.79$, $\sigma=1.95$, $p<0.05$) (Figure 5). The data indicates that the participants' trust of the robot was higher when the robot operated reliably and lower when the robot's reliability dropped during the runs. However, the Muir trust questionnaire was not able to discern between the different reliability conditions and confirms the findings of our earlier experiments [16].

2) Real-time trust: To better examine the impact of differing reliability conditions on trust, we gathered and analyzed real-time trust data. Figure 4 shows how trust evolved during the four reliability conditions. The graphs show an overall increasing trend in trust. As expected, trust for Reliability A monotonically increases while trust for Reliability B, C, and D does not. There are noticeable drops in trust when reliability decreases and, once reliability recovers, trust again starts to increase monotonically. We calculated the area under the trust curve (AUTC) to analyze this data.

The AUTC data highlights the impact of timing of low reliability periods on trust. Figure 5 shows the mean AUTC values along with Muir trust values. A two-way analysis of variance for AUTC yielded a significant main effect of Reliability, $F(3,99)=5.66$, $p<0.01$; however, the main effect of Group type was not significant, $F(1,99)=0.54$, $p=0.46$. The interaction of Reliability and Group type was also not significant, $F(3,99)=0.03$, $p=0.99$. Post hoc comparison for Reliability using Tukey's HSD test indicated that the trust values for Reliability A ($\mu=92.0$, $\sigma=45.7$) were significantly higher than Reliability B ($\mu=32.7$, $\sigma=58.7$, $p<0.01$) and C ($\mu=52.3$, $\sigma=54.9$, $p<0.05$).

An important observation can be made based on this data: periods of low reliability earlier during the interaction have a more detrimental impact on overall trust than periods of low reliability later in the interaction. While this analysis addresses *Q1*, it does not explain how periods of low reliability early in the interaction impact trust. To understand this we investigated how low reliability impacted the recovery of trust (Section III-A3) and how trust changes during a period of low reliability (Section III-A4).

3) Recovery of trust: We looked at normalized[4] AUTC before and after the period of low reliability for C (6 gates each) to investigate trust recovery. Due to the placement of the reliability drops in both B and D, there were not as many gates before and after each reliability drop, meaning there were shorter periods in which to assess trust recovery, so B and D

[4]Normalization occured by setting the y-axis to 0 at the start of the period and calculating AUTC between this start and the end of the period. This allows comparisons without skew from the entering cumulative trust level.

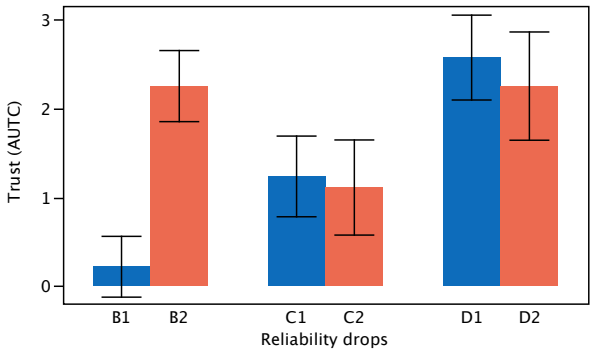

Fig. 6. Area under the trust curve (AUTC) for the following drops in reliability: first drop in B, second drop in B, first half of drop in C, second half of drop in C, first drop in D, and second drop in D.

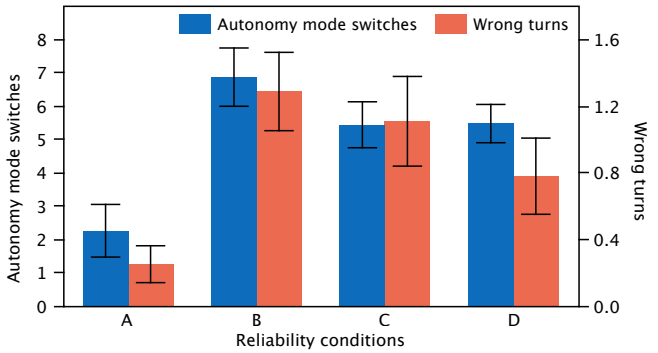

Fig. 7. Autonomy mode switches and wrong turns during different reliability conditions.

were omitted from this analysis. Using a paired two-tailed t-test, we found that the post-C value (μ=8.33, σ=8.27) was significantly lower than pre-C (μ=11.82, σ=7.87, t(27)=3.12, $p<0.01$). According to this data and the corresponding analysis, we find that the recovery of trust after a reliability drop occurs at a slower pace than the pace at which trust develops before reliability drops or in condition A, which has no reliability drops.

4) Trust during low reliability periods: We compared the normalized AUTC during periods of low reliability to determine how real-time trust was impacted by these periods (Figure 6). The low reliability periods were two gates long at two separate instances in reliability B (B1 and B2) and D (D1 and D2). To ensure consistency, the four consecutive low reliability periods in C were split into C1 and C2. Three two-tailed paired t-tests were conducted on the AUTC values for the three pairs of reliability drops. The AUTC value for B1 (μ=0.21, σ=1.81) was significantly less than that of B2 (μ=2.25, σ=2.11, $p<0.01$). No statistically significant difference was found for the other two pairs.

Data about trust during periods of low reliability and the recovery thereafter indicates that an early period of low reliability impacts trust in two ways: 1. It reduces an operator's trust more than it would if low reliability occurred later on, and 2. The rate of recovery is lower than it would otherwise be, thus addressing the first half of *Q1*. Early periods of low reliability not only impact trust, but also take a toll on the operators' control allocation strategy as shown in Figure 7 and further explained in Section III-B.

As suspected, post-run Muir trust ratings appear to be biased by a primacy-recency effect. If Muir was more representative of the whole experience, then the persistently low C1 and C2 AUTC levels would suppress Muir dramatically. As is typical, recency appears to be stronger than primacy. Our prior study, which had 13 gates and only one reliability drop, showed that Muir was less negatively affected by an early drop [16]. This aligns with our findings here, which show the early AUTC B1 drop being washed out in Muir by later gains. This study is longer and allows more early trust accumulation in C and D, thereby suggesting that the primacy-recency effect is amplified for longer experiences. The lack of impact from the late D2 drop is contrary to the primacy-recency theory, but the upwards AUTC slope for the last three gates (Figure 4) suggest that a positive primacy experience and ending strong can mask problems. This, combined with the relatively small drops in trust seen during low reliability, implies that trust can develop inertia over longer experiences.

B. Changing autonomy levels

To examine the impact of reliability on the participants' control allocation strategy, we conducted a two-way analysis of variance for autonomy mode switches that yielded a significant main effect of Reliability, $F(3,104)=6.68$, $p<0.05$ and Group type, $F(1,104)=6.64$, $p<0.01$. However, the interaction of Reliability and Group type was not significant, $F(3,104)=0.32$, $p=0.80$. Post hoc comparison for Reliability using Tukey's HSD test indicated that the autonomy mode switches for Reliability A (μ=2.25, σ=4.17) were significantly lower than Reliability B (μ=6.85, σ=4.62, $p<0.01$), C (μ=5.42, σ=3.64, $p<0.05$), and D (μ=5.46, σ=3.03, $p<0.05$) (Figure 7).

The difference in autonomy mode switches between reliability condition A and reliability conditions B, C, and D indicates that the participants noticed the changes in reliability, its potential for impact on performance, and adjusted their control allocation strategy accordingly. Based on this finding, we expected the number of incorrect passes (wrong turns) to be similar across all reliability conditions, especially for the low reliability conditions. However, the results of a one-way analysis of variance for wrong turns among the reliability conditions showed significant results, $F(3,107)=4.35$, $p<0.01$. Post hoc comparison using Tukey's HSD test indicated that Reliability A (μ=0.25, σ=0.58) had significantly fewer wrong turns than Reliability B (μ=1.28, σ=1.24, $p<0.01$) and C (μ=1.1, σ=1.42, $p<0.05$), but not D (μ=0.77, σ=1.18, $p=0.33$) (Figure 7). The proportional relationship between autonomy mode switches and wrong turns (though not significant), highlights the possibility that periods of low reliability early in the interaction can confuse operators and can result in suboptimal control allocation strategy.

C. Effect of feedback

As stated in Q2 above, we wanted to examine the impact of providing feedback about the robot's confidence in its sensors

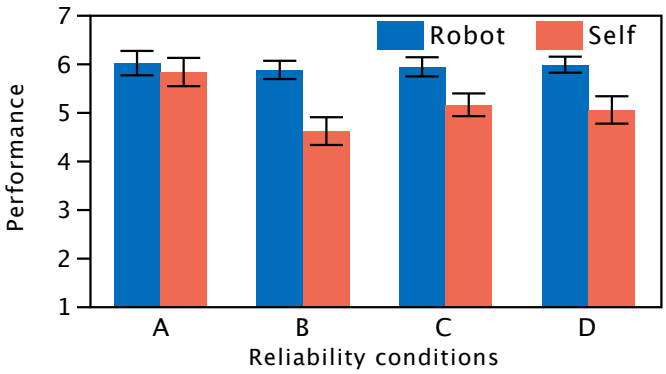

Fig. 8. The robot's perceived performance rating and self performance rating on a semantic differential scale (7=excellent and 1=poor).

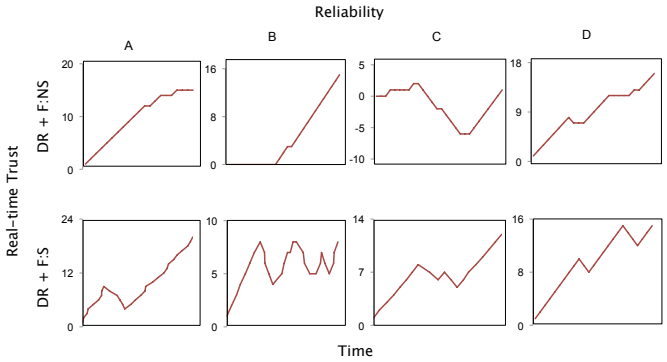

Fig. 9. Sample trust curves for DR+F:NS (participant 1) and DR+F:S (participant 13). Axes are of different scales to better visualize the jerks in the trust curves.

on participants' trust and control allocation. The real-time trust data from Section III-A2 showed a non-significant effect of Group type on AUTC trust; however, the results from Section III-B showed a significant effect of Group type on control allocation. Hence we conducted a post hoc comparison for autonomy mode switches by Group type using a Student's t-test. The result indicated that DR+F (μ=5.81, σ=4.63) had significantly more autonomy mode switches than DR (μ=3.91, σ=3.33, $p<0.05$).

Participants who received feedback switched into assisted mode and back significantly more to correctly pass gates during low reliability. However, it was also observed that participants often switched into assisted mode whenever there was a drop in the robot's confidence, even in high reliability regions when the confidence level changed from high to neutral. We speculate that these changes were due to participants anticipating a robot failure after seeing the robot's confidence drop. Overall, this behavior resulted in fewer wrong turns for DR+F. An unpaired one-tailed t-test was conducted to verify the effect of Group type on wrong turns. As expected, the result indicated that the DR+F group (μ=0.7, σ=1.00) had fewer wrong turns (marginally significant) than the DR group (μ=1.06, σ=1.42, p=0.07).

This result implies that an operator's control allocation strategy can be altered by providing information about the robot's confidence. However, the information should be provided only when appropriate to avoid unwanted side effects.

1) Feedback type: We hypothesized that semantic feedback would increase the perceived intelligence of the robot and therefore produce more dramatic trust responses when the robot changed confidence. Figure 9 shows a sample set of trust curves, one for DR+F:S (for Participant 1) and DR+F:NS (for Participant 13). When plotting the trust curves we observed that the curves for DR+F:S were more rugged compared to those from the DR+F:NS condition. To investigate this effect, we examined the rate of change of the trust curves by computing the magnitude of its second derivative (i.e., jerk). An unpaired two-tailed t-test analysis showed a significantly higher value for DR+F:S (μ=0.99, σ=0.15) than DR+F:NS (μ=1.43, σ=0.15, $p<0.05$). The higher second derivative value for DR+F:S indicates that trust curves for DR+F:S had significantly more sudden changes then DR+F:NS. This result shows that the modality used to present information to participants plays an important role as well. In applications where we want to build up and maintain trust at a steady level, we should choose non-semantic indicators, whereas we would want to use semantic indicators in manipulative applications which demand more attention from the operator.

D. Predicting trust

The ability to predict an operator's broad control allocation strategy can be very useful in preventing accidents and improving performance. For example, if it can be known prior to an operator interacting with a robot, with some certainty, that he or she is less likely to trust an autonomous robot, then it might be worth taking additional steps during training or the initial interaction to mitigate lower trust levels. To examine what factors, if any, can be used to predict an operator's tendency to trust robots, we performed a backward stepwise regression on the data from the DR group (R^2=0.86, $p<0.01$). Table I lists different factors that were considered for the analysis and those that can be used to estimate trust. This is an initial attempt to predicting trust and this data will be augmented with data from future studies.

Most of the factors listed in Table I can be assessed prior to an operator interacting with a robot. The ability to gauge an operator's tendency to trust an autonomous robot can provide valuable information. As in our prior study, most of the factors are not based on the robot's actual performance [16]. This is not to say that the robot's performance does not impact trust; however, it accounts for a smaller variation when trust is aggregated across all reliability levels.

A similar trend was observed when we examined how participants rated the robot's performance. There was no significant difference in the robot's performance for the four reliability levels (Figure 8). However, the robot's performance did impact their rating of self performance. A one-way analysis of variance for the participants' self performance rating yielded a significant main effect of Reliability, $F(3,103)=3.43$, $p<0.05$. Post hoc comparison for Reliability using Tukey's HSD test indicated that the self performance rating for Reliability A (μ=5.82, σ=1.54) was only significantly higher than Reliability B (μ=4.60, σ=1.49, $p<0.05$). A one-way analysis of variance

TABLE I
BACKWARDS STEPWISE REGRESSION RESULTS FOR AUTC TRUST RATINGS FROM THE DR+F GROUP (R^2=0.86).

Effect	Estimate	p
Age	6.21	< 0.01
Exp with robots	-5.61	< 0.01
Exp with RTS games	-5.31	< 0.01
Risk taking Q3	4.78	< 0.01
Exp with FPS games	4.35	< 0.01
Autonomy mode switches	-4.24	< 0.01
Time	3.85	< 0.01
Exp with RC cars	-3.72	< 0.01
TLX	-3.33	< 0.01
Risk taking Q2	-3.32	< 0.01
Wrong turns	2.21	< 0.05
Risk of not receiving bonus	2.13	< 0.05
Self performance rating	removed	–
Robot's performance rating	removed	–
Risk taking Q1	removed	–
Risk taking Q4	removed	–
Scrapes	removed	–
Bumps	removed	–
Pushes	removed	–
Tracking tasks missed	removed	–
Trust prompts missed	removed	–
Time in assisted mode	removed	–
Time in fully autonomous mode	removed	–

for the robot's performance rating yielded a non-significant effect of Reliability.

IV. CONCLUSIONS

The real-time trust results from this study confirm traditional post-run survey approaches for human-robot trust can be masked by primacy-recency bias and demonstrate that early drops in reliability negatively impact real-time trust differently than middle or late drops (*Q1*). In particular, early drops in reliability have dramatically lower real-time trust than later drops and appear to promote suboptimal control allocation strategies.

This study also shows that robot confidence feedback can improve autonomy control allocation during low reliability without altering real-time trust levels (*Q2*). Therefore, warning users of potential robot performance drops can be done without negatively impacting trust in the robot. It should be noted that feedback interface designs using semantic symbols lead to more abrupt real-time trust changes than non-semantic symbols (*Q3*). Designers should match feedback interface style to the prefered response for their application domain.

Finally, this study also shows that predicting real-time trust from user behavior is more feasible than trying to predict post-run survey trust. In our prior work, the only user behavior factors predictive of trust measured with the Muir survey were visual search performance and milestone payment [16]. Neither of these are easily measured internally by a robot, thereby making robot-assessed human trust impossible. In this study, predictive factors of AUTC included autonomy mode switches, elapsed time, and taking wrong turns. While somewhat limited, these support the belief that real-time robot assessment of human trust is possible.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation (IIS-0905228 and IIS-0905148). Thanks to Jordan Allspaw, Christian Bruggeman, Erika Mason, and Anusha Nagabandi for their assistance with robot and experiment logistics, Adam Norton for his assistance with figures, and to all of the experiment participants.

REFERENCES

[1] P. deVries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *Int. J. Human-Computer Studies*, vol. 58, no. 6, 2003.

[2] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 2, pp. 230–253, 1997.

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

[4] J. D. Lee and N. Moray, "Trust, Self-Confidence, and Operators' Adaptation to Automation," *Int. J. Human Computer Studies*, vol. 40, no. 1, pp. 153–184, 1994.

[5] R. O. Phillips, "Investigation of Controlled Flight into Terrain: Descriptions of Flight Paths for Selected Controlled Flight into Terrain (CFIT) Aircraft Accidents, 1985-1997," FAA, Tech. Rep., 1999.

[6] S. Dixon and C. Wickens, "Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload," *Human Factors*, vol. 48, no. 3, pp. 474–486, 2006.

[7] V. Riley, "Operator reliance on automation: Theory and data," *Automation and Human Performance: Theory and Applications*, 1996.

[8] R. Parasuraman, "Adaptive automation and human performance: II. effects of shifts in the level of automation on operator performance," DTIC Document, Tech. Rep., 1991.

[9] M. Dzindolet, L. Pierce, H. Beck, L. Dawe, and B. Anderson, "Predicting misuse and disuse of combat identification systems," *Military Psychology*, vol. 13, no. 3, pp. 147–164, 2001.

[10] J. Bahner, A. Huper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *Int. J. Human-Computer Studies*, vol. 66, no. 9, 2008.

[11] C. Wickens and X. Xu, "How does automation reliability influence workload. 1st annual robotics consortium," *US Army Research Laboratory Collaborative Technology Alliance*, 2003.

[12] J. Lee, "Trust, Self-confidence and Operator's Adaptation to Automation," Ph.D. dissertation, University of Illinois, 1992.

[13] J. Lee and N. Moray, "Trust, self-confidence and supervisory control in a process control simulation," *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pp. 291–295, 1991.

[14] C. Wickens, K. Gempler, and M. Morphew, "Workload and Reliability of Predictor Displays in Aircraft Traffic Avoidance," *Transportation Human Factors*, vol. 2, no. 2, pp. 99–126, 2000.

[15] C. Wickens and X. Xu, "Automation Trust, Reliability and Attention HMI 02-03," *Technical Report AHFD-02-14/MAAD-02-2*, Oct. 2002.

[16] M. Desai, M. Medvedev, M. Vazquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of Changing Reliability on Trust of Robot Systems," in *Proc. Seventh Annual ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2012.

[17] M. T. Dzindolet, H. P. Beck, and L. G. Pierce, "A Framework of Automation Use," Tech. Rep. ARL-TR-24 12, Mar. 2001.

[18] L. Perkins, J. E. Miller, A. Hashemi, and G. Burns, "Designing for Human-Centered Systems: Situational Risk as a Factor of Trust in Automation," in *Human Factors and Ergonomics Society*, 2010.

[19] 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*, vol. 53, no. 5, pp. 517–527, Sep. 2011.

[20] M. Dzindolet, S. Peterson, R. Pomranky, L. Pierce, and H. Beck, "The role of trust in automation reliance," *Int. J. Human Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[21] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe, "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," *Human Factors*, vol. 44, no. 1, p. 79, 2002.

[22] J. D. Lee and N. Moray, "Trust, Control Strategies and Allocation of Function in Human-Machine Systems," *Ergonomics*, vol. 31, no. 10, 1992.

[23] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Human Mental Workload*, vol. 1, no. 3, pp. 139–183, 1988.