Towards Modeling Real-Time Trust in Asymmetric Human-Robot Collaborations

Anqi Xu and Gregory Dudek

Abstract We are interested in enhancing the efficiency of human-robot collaborations, especially in "supervisor-worker" settings where autonomous robots work under the supervision of a human operator. We believe that trust serves a critical role in modeling the interactions within these teams, and also in streamlining their efficiency. We propose an operational formulation of human-robot trust on a short interaction time scale, which is tailored to a practical tele-robotics setting. We also report on a controlled user study that collected interaction data from participants collaborating with an autonomous robot to perform visual navigation tasks. Our analyses quantify key correlations between real-time human-robot trust assessments and diverse factors, including properties of failure events reflecting causal trust attribution, as well as strong influences from each user's personality. We further construct and optimize a predictive model of users' trust responses to discrete events, which provides both insights on this fundamental aspect of real-time human-machine interaction, and also has pragmatic significance for designing trust-aware robot agents.

1 Introduction

In this paper, we consider methods that aim to increase the efficiency of human-robot teams, by optimizing the robot's level of performance with respect to an objective function reflecting human satisfaction: *trust*. The specific class of human-robot collaborations that we address are those with supervisor-worker relationships, where the human oversees and delegates work to a robot "worker", and also has the ability to take over control momentarily to correct the robot's mistakes when necessary. We are motivated to develop techniques to simultaneously improve the quality of the work performed by the robot, and also reduce the human supervisor's

Anqi Xu and Gregory Dudek
School of Computer Science, McGill University, Montreal, Canada,
e-mail: anqixu@cim.mcgill.ca/dudek@cim.mcgill.ca

task load demands. We believe that *trust* is the key underpinning towards realizing these objectives, namely by enabling the robot to sense and adapt to the human's intentions through trust assessments, and by encouraging the building of the human's level of trust in the robot leading to the intrinsic disposition towards task delegation.

This paper describes our latest contributions in quantifying and predicting real-time changes in the human's level of trust in the robot during interaction. This is inspired by our previous research [20] on modeling and making use of human-robot trust within a tele-robotics setting, where a human operator supervises a remotely-located robot that is autonomously tracking terrain boundaries using visual processing. Whereas our previous work characterized the level of trust that the robot *deserves* based on logical reasoning about its task performance, we take a more data-driven approach in this work. Concretely, we attempt to predict the *actual* degree of trust the user has in the robot, from moment to moment during interaction, by studying experience-based metrics as well as human factors that give rise to subjective trust attribution.

We report on a user study that collected empirical human-robot interaction data in reaction to different controlled events. In particular, this study logged all experiences and actions occurred during short interaction sessions, and also elicited questionnaire responses reflecting each operator's mental state when interacting with the robot. The questionnaires were designed to quantify a number of factors that are known to influence human-robot trust, based on existing literature. We also present a descriptive analysis of the resulting dataset that establishes several quantitative characteristics of users' trust responses to different events, which are both logical and consistent with prior research. We further propose a parametric model for predicting reactive changes in the user's trust, and evaluate this novel model's generalizability and ability to predict the real-time progression of human-robot trust.

2 Background

Our work is inspired by an extensive literature across diverse disciplines observing the critical role played by trust in human teams. Trust is a very rich concept in the modeling of human behavior, and it is subject to a multitude of interpretations under different contexts, such as within a society, an organization, or a mutual relationship [15]. Within a mutual human-robot team, trust encompasses two major elements:

- the degree of trust: a quantifiable subjective assessment towards another individual;
- the act of trust: the decision and behavior of relying upon an individual's abilities or services.

In this work, we focus solely on quantifying the *degree of trust*. This measure can then be applied to mechanisms that encourage a human to adopt the *act of trust*, as shown in our prior work [20].

2.1 Related Work

Studies of trust in Human-Robot Interaction (HRI) historically evolved out of a multi-decade literature investigating the interaction between humans and automation. Several measurement scales of a human's degree of trust in computer automation have been previously proposed and evaluated [13, 16]. Other groups have quantified how trust varies with respect to the performance and error rate of the human-robot team [6, 9, 18], the nature of these failures [2], and the user's mental load [6].

Lee and Moray [14] developed a temporal model for characterizing human trust within a human-automation team setting. Our work shares several similarities with the authors' approach, and further extends analysis into the human-robot domain.

Among the earliest studies of trust in HRI, Hall [10] formulated a binary trust measure assigned to each state of the world, and devised an update mechanism for trust based on the robot's experiences. Freedy *et al.* [8] investigated the effects of mixed initiative robot control on a user's trust within a military tele-robotics setting. Hancock *et al.* [11] carried out a meta-analysis of empirical results from the HRI trust literature, and established quantitative estimates of various factors influencing trust across different interaction domains. Yagoda [21] gathered trust assessments from a broad audience by showing videos of human-robot interactions and eliciting users' trust responses through an online crowd-sourcing framework. Arkin *et al.* [1] studied aspects of trust and deception in a tele-robotics context. Our research shares various commonalities with all of these works, in the formulation, instantiation, elicitation, and evaluation of human-robot trust.

Our study of human-robot trust is most similar to the work by Desai *et al.* [4, 5], which carried out a multitude of investigations on trust within a search-and-rescue tele-robotics setting. In particular, the authors' quantified the effects on trust of diverse interaction-level factors, including the level of autonomy, degree of robot reliability, amount of situational awareness, etc. Likewise, in this work we begin with a descriptive approach for studying human-robot trust, though at a finer time scale of the interaction experience. We then expand on the analysis further to extend towards a predictive model for the user's real-time trust assessments.

2.2 Trust Characterization

Several studies have highlighted the influences of various factors on the degree of a human's trust in a robot (e.g. [5, 11, 14]). These factors include:

- human's *demographic attributes*: e.g. age, gender, occupation;
- human's *attitudes and experiences*: e.g. propensity to trust robots, prior experience with robots and with task setting;
- human's perception of *robot attributes*: e.g. adaptability, benevolence;
- robot's *task performance*: e.g. internal automation failures, task errors;
- attributes of the *interaction setting*: e.g. communication quality, task complexity.

A key distinction among these factors is their *bases of trust*, which can be categorized into two main classes: certain trust factors relate to notions of the robot's competence, such as its task accuracy and consistency. These differ conceptually from factors related to the trustees intentions, pertaining for instance to the robots willingness and benevolence. Following the majority of research in robotics and automation, our work will take intention-centric bases of trust for granted, and assume that the robots designers did not include deceptive behaviors into its programming.

Measures of trust have been quantified using a number of different formats in the literature, including a binary representation [10], a continuous bounded measure [8, 14], and a multi-dimensional measure [13, 16]. Each representation has its own merits and drawbacks, and there is no "true" or perfect format unfortunately since trust is fundamentally a non-observable construct. In this paper, we quantify human-robot trust as 1-D bounded continuous measure, which enables both ease of user feedback and the application of standard statistical analysis techniques.

2.3 Interaction Context

Our research revolves around a team that is comprised of an autonomous robot and a human supervisor, collaborating on a common task. Ideally, the software agent governing the robot's autonomous behaviors is responsible for carrying out the bulk of the workload, while the operator predominantly monitors of the task progression. When the autonomous agent makes a mistake however, the human can actively intervene and provide corrective help by overriding the robot's commands. Given the nature of this supervisor-worker relationship, we assume that the human's intervening commands will *always supersede* those generated by the autonomous agent.

We have chosen to study vision-guided navigation as our primary application domain. Concretely, our human-robot setting consists of a human operator sharing control with an autonomous vision-based agent [19] over an aerial vehicle, while being tasked to track different terrain boundaries, such as coastlines and roads. Visual navigation tasks are appealing because humans are naturally inclined to solve them robustly and without effort. In addition, the necessary complexity in autonomous solutions (e.g. [5, 7]) warrants the need of trust. Finally, these setups are relevant to a wide variety of different robot platforms and application contexts.

Within our boundary tracking framework, the human operator is presented with a graphical interface showing the live camera feed from the robot, as seen in Fig. 1. The autonomous agent's internal state is overlaid on top of this view, in the form of the currently-tracked boundary curve and line fit, as well as the current steering command. These overlays provide transparency to the autonomous agent's sensing and control processes, and therefore help the user understand the robot's behaviors more clearly. In addition, whenever the boundary tracking algorithm fails to detect the boundary, the user can also readily perceive such faults by the absence of the boundary-related overlays.

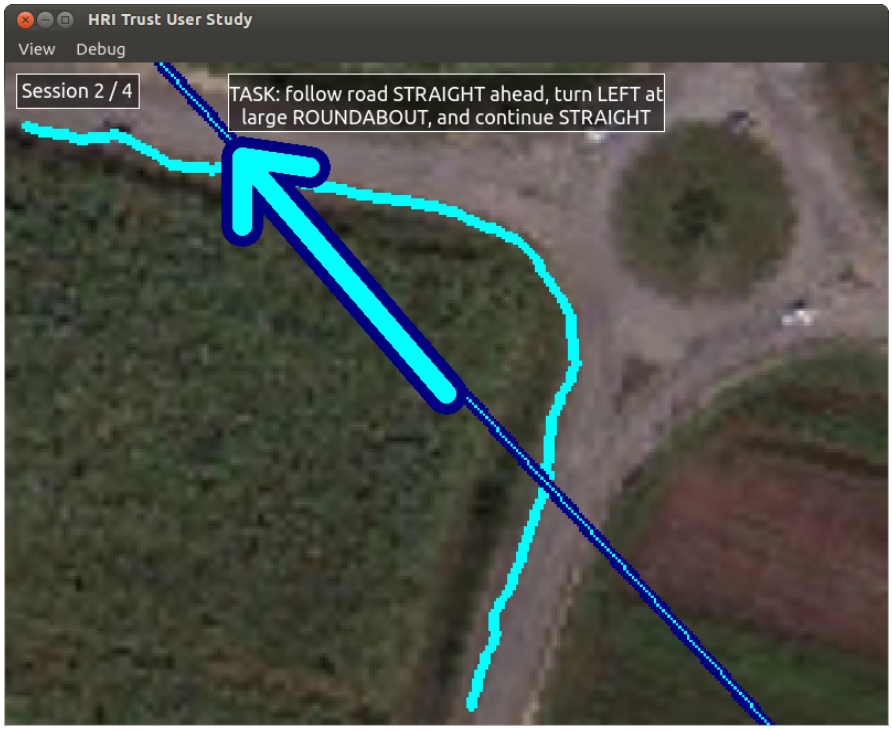

Fig. 1 Our boundary tracking framework provides an interface that overlays the autonomous tracker's internal state representation on a real-time display of the robot's camera view. Additional text overlays pertain to our human-robot interaction study, which are discussed in Sect. 3.

The operator can take over control of the vehicle at any time, by holding down the mouse cursor over the camera display in the desired steering direction; these user interventions are reflected by a change in the steering arrow's color. The autonomous tracker is always running and its internal visualizations are also continuously displayed on-screen, even during periods of manual control. This allows the user to perceive when the agent has regained the tracked target, and thus also when to return control back to the autonomous system.

3 Methodology

We developed a human-robot interaction study to collect empirical data towards the analysis and modeling of real-time human-robot trust. In this study, users interacted with our autonomous boundary tracking agent to control a simulated aerial vehicle. Our simulation framework generates a bird's-eye camera feed of a non-holonomic fixed-wing aerial robot, based on real satellite imagery.

Although our boundary tracker has been previously deployed on both fixed-wing aerial robots [19] and quadrotors [20], we chose deliberately to target an aerial robot simulation framework for our user study in contrast. This setup allowed us to admin-

ister specific interaction experiences to different study participants in a *controlled* and *repeatable* manner, and also eliminated pragmatic concerns such as battery levels and fluctuations in environment conditions.

3.1 Event-Centric Perspective

We perceive a given period of human-robot interaction experience as a sequence of *discrete events*, where each event corresponds to a salient change in the state of the robot and/or the environment. Examples of such events within our visual navigation setting include a sustained period of internal failures in the boundary tracking algorithm, or a strong gust that pushes the aerial robot sideways and causes it to lose track of the target boundary. By measuring the human's *change in trust* in reaction to different types of events, we can quantify the progression of human-robot trust at a small time scale, namely short periods of interaction experience centered around each event. This event-centric view also differentiates our investigations from the majority of studies in the literature, which have characterized impacts on trust from aggregated interaction experiences on longer-term time scales (e.g. [5, 8, 21]).

Concretely, our user study is comprised of a number of short interaction sessions (< 60 seconds each), where in each section the task is to follow a straight road for about 30 seconds till an intersection, make a predetermined turn, and then continue following the new path. We used different road segments and varied the turn directions in each session, in order to introduce diversity in the interaction experience and also prevent potential learning effects. The autonomous boundary tracker is capable of following sides of roads proficiently, but lacks the ability to switch between multiple target boundaries at intersections. Participants in our study were explicitly informed of this limitation prior to commencing the interaction sessions.

In this paper we focus specifically on events corresponding to robot failures, i.e. periods of decreased reliability in the robot's task performance. These drops in reliability are achieved by changing the parameter settings of our boundary tracking algorithm into a *low-reliability* state, where the autonomous agent poorly tracks roadside boundaries and also exhibits frequent internal failures in a non-predictable manner. We devised the following *event scenarios*, by programmatically toggling between reliability modes at different times during the course of each session:

- `Baseline`: the boundary tracking agent is set to high-reliability state throughout the entire session;
- `PoorStart`: the agent starts in the low-reliability state for 10 seconds and is then switched into the high-reliability setting, prior to the intersection,
- `RobotFault`: the agent is momentarily toggled into low-reliability state for a 10-second period in the middle of tracking the first road segment;
- `Limitation`: the agent is switched into low-reliability state at the road intersection, and then switched back into high-reliability state after 10 seconds.

We believe that robot operators typically behave in a *rational* manner, and will attribute blame following robot misbehaviors differently based on the cause of each failure. Our event scenarios are designed to elicit trust changes following different

types of failures, namely poor initial tuning of the autonomous agent (`PoorStart`), algorithmic failure without any noticeable external cause (`RobotFault`), and failure due to limitations in the robot's programming (`Limitation`).

3.2 Trust Factors

The main purpose of our trust modeling work is for an autonomous robot to have the capability of predicting the human's trust in real-time during interaction. Starting from the rich corpus of factors that have been shown in the literature to influence human-robot trust, we exclude factors that are ill-defined, and those that are not possible for the robot to obtain, either via direct observation or by querying the human operator. In addition, since our research context assumes that the robot is always well-intentioned and is never adversarial, we also exclude all intention-based factors from our investigations. Therefore, our study is designed to collect interaction and user feedback data for quantifying the following *experience-based* trust factors:

- the robot's task performance (i.e. distance between robot and tracked target);
- the autonomous agent's internal failure rates;
- the frequency of interventions from the human operator;

In addition, our questionnaires elicit the following *assessment-based* factors:

- a pre-experiment survey: user demographics, general attitudes, and prior experience with robots and remote control (RC) tasks (following [4]);
- post-session questionnaires: assessments of the robot's and user's task performances, as well as the robot's perceived robustness and adaptability;
- a debriefing questionnaire: experiment-wide task load assessments (via Raw TLX [12]), and post-hoc updates on trust propensity towards autonomous robots.

This study design highlights the important characteristic that human-robot trust is dependent on factors at *different time scales*. In particular, we expect responses to both the survey and debriefing questionnaires to be constant throughout the entire study, in contrast to per-session user assessments. These experience-based factors are summarized by statistics aggregated over the entire duration of each session, as well as within a short window of time following an event. The former set of measures reflect a cumulative characterization of the interaction experience during each session, whereas the post-event window-level statistics quantify immediate reactions from both the user and the robot following a discrete event.

3.3 User Assessment Elicitation

In addition to the various questionnaires for gathering assessment-based trust factors, we also asked users to indicate their degree of trust in the robot both before and after each session. The majority of these user queries employed the Visual Analogue Scale (VAS) [17] to elicit unipolar ("Likert-like") and bipolar responses. The VAS format, i.e. a continuous scale without tick marks (Fig. 2), was chosen for its superior metric properties over N-point discrete scales [17]. We have iterated over the design of our questionnaires, in order to mitigate common sources of biases [3].

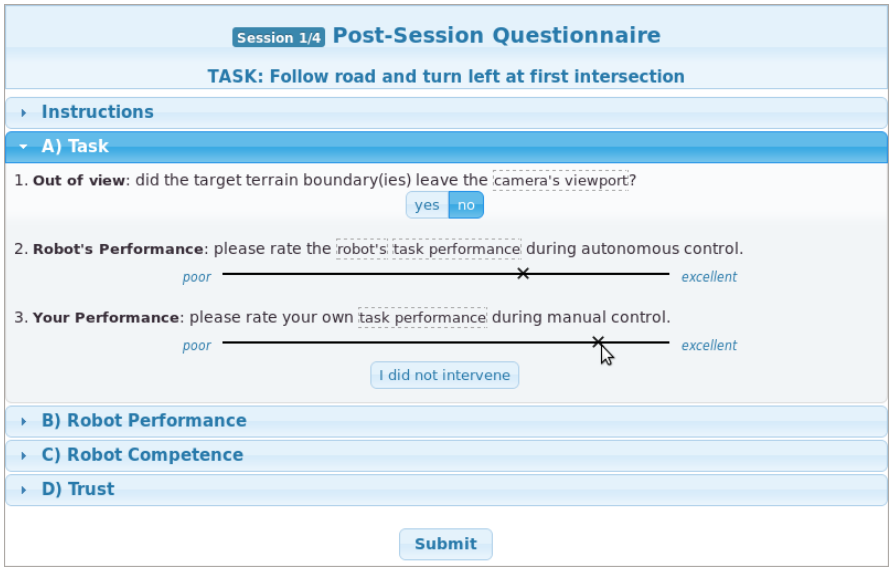

Fig. 2 Questionnaires in our study are comprised predominantly of user assessments using a continuous Visual Analogue Scale (VAS) [17] (i.e. continuous slider without tick marks) as answering format, as well as a few discrete-choice queries.

3.4 Structure of Interaction Study

Fig. 3 depicts the different phases in our interaction study, beginning with the survey questionnaire. A set of tutorial slides then provided explanations of the interaction and task context, the interface, as well as the robot's capabilities (i.e. configurable to track various types of boundaries) and limitations (i.e. incapability of changing tracked targets deliberately). The tutorial also explicitly asked users to assume that the autonomous robot was well-intentioned, and to therefore provide trust assessments based solely on the robot's performance and competence.

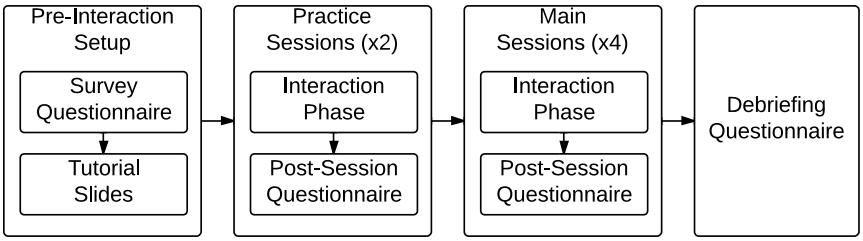

Fig. 3 The structure of our user study is designed around multiple controlled human-robot interaction sessions, which are supplemented by a number of questionnaires.

Next, participants interacted with boundary tracking robots during 6 distinct sessions, corresponding to 2 initial practice sessions, and followed by the 4 event scenarios in a random, counterbalanced order. Each session began by asking users to provide their initial trust assessments in the yet unseen robot, followed by the actual interaction phase, and ending with a post-session questionnaire.

In the first practice session, participants were instructed to get acquainted with the visual interface and robot controls during free roam, while tracking arbitrary terrain boundaries. The autonomous robot agent was toggled into its low-reliability state for short periods of time on several occasions, both to demonstrate robot failures as well as to implicitly prompt users to practice intervening and taking manual control of the vehicle. The second practice session consisted of a variant of the `Baseline` scenario, and acquainted participants with the road-following task objective by providing a demonstration of a typical interaction experience.

Following the 6 interaction sessions, the study concluded with a debriefing questionnaire that collected assessments of the aggregated interaction experience, as well as general feedback. This entire interaction study was completely automated, including the event triggers and data logging components.

A key aspect of this study design is that the user's trust assessments are affected by a full-fledged autonomous agent, operating within a simulated environment. This setup therefore provides conditions similar to a real world setting, while also enabling experimental control to ensure consistent and repeatable experiences.

4 Interaction Study Results and Analyses

We recruited 30 participants from the School of Computer Science at McGill University to participate in our interaction study. The user population is comprised of 24 males and 6 females, and included 11 undergraduate students, 13 graduate students, 2 professors, and 4 university personnel. Participants had varying degrees of experience operating and programming robots, although no user had prior interaction experience with our boundary tracking system.

The resulting dataset encompassed 60 practice session entries and 120 non-practice session entries. We carried out statistical analyses to investigate several key aspects of this human-robot trust dataset, including order effects resulting from our crossover session design, the effect of event scenarios on the amount of change in users' trust assessments, and the relative significance of various factors previously identified in the literature on the influence of real-time human-robot trust.

4.1 Session Order Effects and Properties of Pre-Session Trust

In order to mitigate learning effects of the crossover study design from introducing unwanted biases, we explicitly emphasized, both during the tutorial and during pre-session trust elicitations, that the experience from each session should be assessed *independently* from previous ones, and that each session may encompass robots with different reliability levels, differing task objectives, as well as distinct environments.

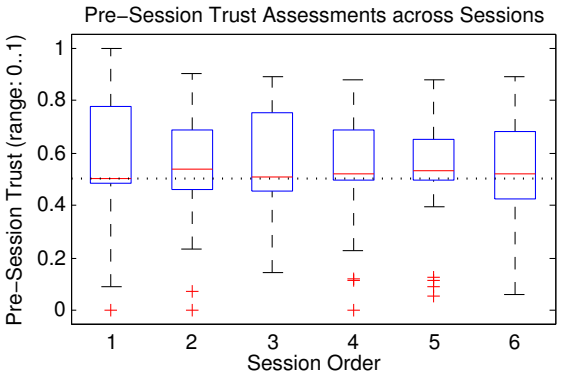

Fig. 4 Pre-session trust assessments from users are consistent across sessions and do not show any significant effects of the session ordering. These results from our study population also demonstrate slight positive bias in initial robot trust assessments, which are consistent with prior findings in the literature [5,6].

Fig. 4 shows the degrees of trust participants felt towards the boundary tracking robot prior to the start of each session. A repeated measures analysis of variance (rmANOVA) revealed no significant relationship between the mean pre-session trust values to the session ordering, $F(5,145) = 0.30$ $(p = 0.91)$, although there was a strongly significant effect from the different users, $F(29,145) = 24.18$ $(p < 1e^{-16})$. We thus conclude that although trust assessments at the beginning of sessions varied among individual users, these measures were not noticeably affected by the session ordering.

Fig. 4 also indicates that the experiment's population demonstrated slight positive bias for pre-session trust level that is above the uninformed prior response of 0.5. A one-way two-tailed Student's *t*-test revealed that the mean of the pre-session trust assessments across all users and all sessions (including practice sessions) was significantly different from 0.5 ($p < 0.05$). This positivity bias is consistent with similar findings in other human-robot studies [5,6].

4.2 Effects of Event Scenarios

Fig. 5 depicts changes between pre- and post-session trust assessments, in response to different events during the 4 main sessions. Repeated measures ANOVA revealed significant effects on the mean amount of trust changes both due to events, $F(3,87) = 16.61$ $(p < 1e^{-16})$, as well as due to users, $F(29,87) = 2.35$ $(p = 1.22e^{-03})$. Post-hoc pairwise comparisons using the Tukey range test at the $p < 0.05$ level showed *non-significant differences* in trust changes between `Baseline` & `Limitation`, `PoorStart` & `RobotFault`, and `PoorStart` & `Limitation`.

Looking at the average user responses, the gain in trust in reaction to the high-performance boundary tracker settings of `Baseline` is logically expected, and similarly so are the losses in trust due to failures during the `PoorStart` and `RobotFault` scenarios. Although the difference in the amount of trust lost between `PoorStart` and `RobotFault` was not significant, Fig. 5 suggests that users reacted more leniently when the robot in `PoorStart` started with poor performance but then soon showed improvements, as opposed to when the initially-reliable robot in `RobotFault` unexpectedly failed to track a straight road.

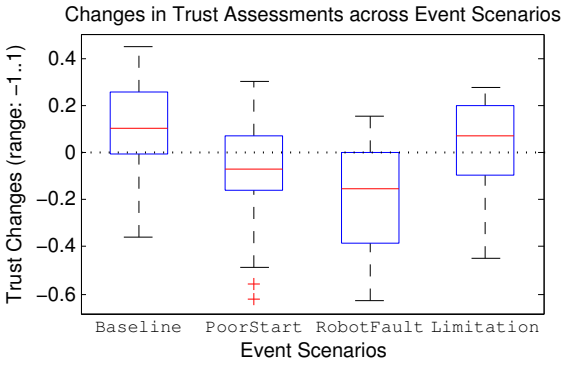

Fig. 5 Trust changes in response to different events are consistent with rational reactions based on causal attribution. In particular, the slight increase in trust in the `Limitation` scenario suggest that most users deliberately did not blame the robot for failing to execute a change in task objective, given known limitations in its programming.

In contrast, the typical response of *increase in trust* for the `Limitation` scenario may appear surprising at first, since the robot was programmatically switched into the low-reliability mode when it reached the intersection. Nevertheless, we believe that users deliberately did not penalize these failures because they were actively conscientious that the autonomous boundary tracker lacked the capabilities of carrying out changes in the task objective. Therefore, the robot's momentary drop in reliability following the intersection can be interpreted in analogy to a switch from the `Baseline` to the `PoorStart` conditions, which is consistent with our post-hoc pairwise comparative analysis. In summary, these results indicate that overall, participants evaluated trust responses in a manner consistent to a rational mindset that both considered causal event assessments, and carried out deliberations with the robot's capabilities and limitations in mind.

4.3 Descriptive Analysis of Factors Influencing Trust Change

A backwards stepwise linear regression analysis was carried out to identify the most significant relationships between factors enumerated in Sect. 3.2, and changes to the user's trust level in response to different events. Starting from a full linear model involving all considered trust factors, the session order, and event scenarios, we applied stepwise regression using the Sum Squared Error (SSE) criterion, which iteratively removed insignificant factors (when $p > 0.1$) and re-introduced relevant factors (when $p < 0.05$). Interactions and high-order terms were disallowed to preclude spurious associations between factors at different time scales. Experience-based factors reflecting the immediate post-event reactions were computed over a 10-second window, corresponding to the duration of the pre-determined lapse into the low-reliability mode. Also, since no failure events occurred during the `Baseline` scenarios, event windows were chosen at random for the corresponding sessions.

Table 1 provides a summary of the stepwise regression results. Starting with 52 trust factors at different time scales, only 22 factors remained in the final model. Significant factors at the experiment-level time scale included demographic entries (e.g. age, occupation), prior expertise and attitudes (e.g. driving and robot control experi-

Table 1 Descriptive Analysis Summary of Human-Robot Trust Factors using Stepwise Regression

Factor Categories (Initial Factor Count)	Final Factor Count	DF	Σ MS	Min. p	Avg. p	Max. p
Survey & debriefing assessments (24)	13	15	4.25	$< 1e^{-10}$	< 0.01	0.02
Post-session assessments (4)	2	2	1.27	$< 1e^{-16}$	< 0.01	0.02
Experience during session (12)	4	4	0.30	$< 1e^{-2}$	0.02	0.04
Experience within 10 sec. window (12)	3	3	0.21	$< 1e^{-3}$	0.27	0.79
Residual		84	0.01			

DF: degrees of freedom Σ MS: combined mean sum of squares

ence, willingness to use a self-driving car), as well as post-experiment assessments (e.g. measures of mental demand and performance). Session-level factors in the final model incorporated post-session user assessments of the robot's performance characteristics, as well as session-wide internal failure metrics. Finally, significant window-level factors corresponded to measures of short-term external task errors. Both the session order ($p = 0.60$) and event scenarios ($p = 0.20$) were excluded by the regression process. The final model showed excellent fit to the data, with a Root Mean Squared Error of $RMSE = 0.11$ and a goodness-of-fit of $R^2 = 0.83$.

We hypothesize that because the event scenario was categorical, it lacked metric precision in correlating to the quantitative amount of trust change, and was thus dropped in favor of other non-discrete window-level and session-wide metrics that characterized the interaction experience. More importantly, these experience-based factors were dwarfed in statistical significance compared to users' assessments, especially at the experiment-level time scale, as reflected by the aggregated Mean Sum of squares (MS) and average p-value statistics in Table 1. We therefore conclude that the evolution of real-time human-robot trust is dependent on each user's *personality* (e.g. expertise, beliefs, tendencies, and perceptions) more significantly than the *actual experiences* and their causalities during interaction.

5 Predictive Real-Time Trust Modeling

We now develop an initial model for predicting trust changes in real-time human-robot interactions, based on our statistical analyses above. This requires a critical change in methodology from our previous descriptive characterizations, which quantified the relationships between trust and its related factors using *all of the collected dataset*. In contrast, the main objective of predictive real-time trust modeling is to estimate event-centric trust responses during interactions with *potentially new users*, while having minimal or no prior knowledge about these users. This is achieved using the standard machine learning technique of Maximum Likelihood model parameter learning through cross-validation.

Sect. 4.3 previously revealed the strong dominance of personality-based factors over experience-based factors on event-centric human-robot trust. Unfortunately, such personalized information may not be available when the robot is interacting with a new user. Therefore, we excluded factors at the experiment-level time scale in this initial work towards predicting human-robot trust.

5.1 Parametric Event-Centric Trust Model

Our model for predicting changes in trust in reaction to interaction events is derived from the same stepwise regression approach previously used in our descriptive analysis. Specifically, event-based trust change $\Delta\mathbb{T}^{session} \in [-1..1]$ is quantified as a weighted linear sum (with weights ω) of both experience-based metrics at the post-event reactionary windowed time scale ($\mathbb{E}_i^{post\text{-}event}$) and at the event-centric session level ($\mathbb{E}_j^{session}$), as well as provided user assessments following each session ($\mathbb{A}_k^{session}$). This model is trained using supervised learning by computing the difference between the user's trust assessment before and after each session:

$$\begin{aligned}\Delta\mathbb{T}^{session}(W,Q) &= \frac{1}{Q}\left[round\left(Q\cdot\mathbb{T}^{post\text{-}session}\right) - round\left(Q\cdot\mathbb{T}^{pre\text{-}session}\right)\right] \\ &= \sum_{\forall i,j,k}\left(\omega_0 + \omega_i\,\mathbb{E}_i^{post\text{-}event}(W) + \omega_j\,\mathbb{E}_j^{session} + \omega_k\,\mathbb{A}_k^{session}\right) \quad (1)\end{aligned}$$

This trust model has two parameters: the post-event window duration W, and the quantization level Q for both the elicited pre-session and post-session trust assessments. The window duration parameter W is related to the temporal sensitivity in the predicted trust change, and allows our model to generalize to other HRI settings potentially, such as turn-based episodic human-robot interaction settings. In addition, within our specific domain of visual robot navigation tasks, W can be adjusted to optimize the predictability on trust changes from post-event experience-based metrics, such as internal robot failures and the rate of user interventions.

The trust response quantization parameter Q addresses a separate concern, namely that reported assessments within questionnaires may contain diverse sources of bias [3]. Some of these biases were addressed in our study design by using the Visual Analogue Scale (VAS) answer format, which exhibits desirable metric properties [17]. In conjunction, the Q parameter quantizes the user's trust responses in order to eliminate noise resulting from the variability in the exact pixel placement of selections on the VAS answer scale. Quantized trust responses are also re-normalized to facilitate comparisons between different quantization levels.

5.2 Parameter Learning

The predictive power of our trust model in Eq. 1 depends on the values of its parameters, namely the post-event window duration W, and the trust response quantization level Q. We used a Maximum Likelihood (ML) supervised learning and validation approach to determine optimal parameter settings. Concretely, we constructed a test set by isolating data from a random 20% of users, and carried out parameter fitting using 6-fold cross-validation on the remaining 80% of user data. We iterated over 11 window duration values of $W = [0.5, 1, 2, 3, 4, 6, 8, 10, 12, 15, 20]$ seconds, and

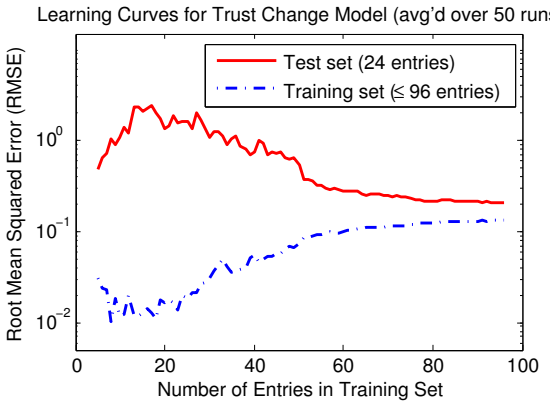

Fig. 6 Learning curves for our optimized trust model ($W = 2$, $Q = 31$), comparing the predictive error of different-sized training sets (up to 80% of the entire dataset) and of a complementary test set, averaged over 50 independent runs. The asymptotic gap between the two error curves indicates that our learning algorithm exhibits high variance, which suggests that larger-sized datasets could further improve the final model's quality.

across 13 trust quantization values, $Q = [3,5,7,9,11,15,21,31,51,71,101,201,501]$. Each stepwise regression model was built using the same procedure as described in Sect. 4.3. The accuracy of each model was then evaluated using the Root Mean Squared Error (RMSE) of the predicted trust changes in the cross-validation set.

In addition, we quantified the generalizability of the chosen model parameters by training and evaluating trust models on a variable-sized subset of data entries, after randomly excluding 20% of users into a separate test set. Training-set and test-set RMSE values (i.e. $RMSE^{train}$ & $RMSE^{test}$) from multiple runs were aggregated to reflect the typical level of generalizability for our trust model.

5.3 Results and Discussion

Our parameter fitting procedure trained over 800 stepwise regression trust models. Among these, the model with the smallest $RMSE^{test}$ revealed optimal parameter values of $W = 2$ seconds and $Q = 31$ levels. Thus, within our boundary tracking task domain, metrics reflecting the robot's behaviors and the user's interventions within a 2-second post-event window was found to be most useful at predicting trust changes. In addition, the large magnitude of the selected trust quantization level $Q = 31$ indicates that this trust model has the potential to provide fine-scaled predictions of the quantitative change in the users' trust assessments.

Fig. 6 depicts the progressions of the training-set and test-set errors averaged over 50 independent runs. The prediction errors of models built using the full training set were $\overline{RMSE}^{train}(96) = 0.13$ $(S.D. = 0.01)$ and $\overline{RMSE}^{test}(24) = 0.19$ $(S.D. = 0.05)$. The smooth asymptotic convergence in the training-set and test-set errors suggest that a significant portion of our training set was required to allow for generalization and avoid over-fitting. In addition, the small magnitude of $\overline{RMSE}^{train}$ and the gap after convergence between the training-set and test-set errors together indicate the presence of high variance in our learning technique. This implies that our (non-

regressed) trust model structure has sufficient expressibility, and we would thus expect to gain further predictive power and generalizability by expanding the size of our training dataset.

Finally, by interpreting $\overline{RMSE}^{test}$ as the standard deviation in the typical prediction error in trust responses for novel users, we deduce that 95% of times the predicted values differ from actual trust changes by ± 0.37 (recall that trust change values lie in $[-1..1]$). We acknowledge that the quantitative performance of our trained and regressed trust models in this work reflect only moderate levels of predictive power, especially compared to the numerical precision of the chosen trust quantization level, $1/Q \approx 0.03$. We suspect that a major source of the predictive error lies in the variability among different users, which has been consistently shown in our analyses to affect trust assessment, and therefore should be incorporated into our trust model in the future to further improve its predictive power.

6 Conclusions

In this work, we characterized key aspects of real-time trust in supervisor-worker human-robot teams, and illustrated these aspects concretely within a collaborative tele-robotics setting. We also carried out a controlled user study to collect both interaction experience and user assessment data, and whose results quantified different degrees of trust changes in reaction to various events during interaction. In particular, we found empirical support for the hypothesis that users in our interaction study typically behaved *rationally* and attributed trust changes based on the cause of each failure event. We further determined that the progression of human-robot trust was predominantly shaped by each user's *personality*, in comparison to the influence from the *actual experiences* in the interaction. Finally, we developed an initial, parametric model for predicting event-reactive trust changes within real-time continuous human-robot interactions, and empirically characterized its performance and generalizability.

We are currently working towards an extended version of our interaction study, which will target a wider user audience and a more elaborate set of event scenarios. We are also investigating ways to integrate personality-based factors into our real-time trust model, in order to further improve its predictive power. Separately, we are actively pursuing pragmatic applications of our real-time trust model to enable robots to intelligently adapt its behaviors based on different types of human interventions [20]. We anticipate that this work and its extensions will culminate into a robust real-time predictive trust model, which can then be applied to streamline the efficiency of human-robot collaborations.

Acknowledgements We would like to acknowledge the NSERC Canadian Field Robotics Network (NCFRN) for its funding support. We would also like to thank all of the participants who contributed to our user study.

References

1. Arkin, R.C., Ulam, P., Wagner, A.R.: Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. Proc. of the IEEE **100**(3), 571–589 (2012)
2. Bisantz, A.M., Seong, Y.: Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. Industrial Ergonomics **28**(2), 85–97 (2001)
3. Choi, B.C., Pak, A.W.: A catalog of biases in questionnaires. Preventing Chronic Disease **2**(1) (2005)
4. Desai, M.: Modeling trust to improve human-robot interaction. Ph.D. thesis, Computer Science Department, U. Massachusetts Lowell (2012)
5. Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., Yanco, H.: Effects of changing reliability on trust of robot systems. In: Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI'12), pp. 73–80 (2012)
6. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Human-Comp. Studies **58**(6), 697–718 (2003)
7. Fong, T., Thorpe, C., Baur, C.: Collaboration, dialogue, and human-robot interaction. In: Proc. of the Int. Sym. on Robotics Research (ISRR'01), pp. 255–266 (2002)
8. Freedy, E., DeVisser, E., Weltman, G., Coeyman, N.: Measurement of trust in human-robot collaboration. In: Int. Symposium on Collaborative Technologies and Systems (CTS'07), pp. 106–114 (2007)
9. Gao, F., Clare, A., Macbeth, J., Cummings, M.: Modeling the impact of operator trust on performance in multiple robot control. In: AAAI Spring Symposium: Trust and Autonomous Systems (2013)
10. Hall, R.J.: Trusting your assistant. In: Proc. of the 11th Knowledge-Based Software Engineering Conference, pp. 42–51 (1996)
11. Hancock, P., Billings, D., Schaefer, K., Chen, J., De Visser, E., Parasuraman, R.: 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)
12. Hart, S.G.: NASA-Task Load Index (NASA-TLX); 20 years later. In: Proc. of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 904–908 (2006)
13. Jian, J.Y., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. Int. Journal of Cognitive Ergonomics **4**(1), 53–71 (2000)
14. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10), 1243–1270 (1992)
15. McKnight, D.H., Chervany, N.L.: The meanings of trust. Tech. rep., U. Minnesota (1996)
16. Muir, B.M.: Operators trust in and use of automatic controllers in a supervisory process control task. Ph.D. thesis, U. Toronto (1989)
17. Reips, U.D., Funke, F.: Interval-level measurement with visual analogue scales in internet-based research: Vas generator. Behavior Research Methods **40**(3), 699–704 (2008)
18. de Vries, P., Midden, C., Bouwhuis, D.: The effects of errors on system trust, self-confidence, and the allocation of control in route planning. Human-Comp. Studies **58**(6), 719–735 (2003)
19. Xu, A., Dudek, G.: A vision-based boundary following framework for aerial vehicles. In: Proc. of the IEEE/RSJ Int. Conf. on Int. Robots and Systems (IROS'10), pp. 81–86 (2010)
20. Xu, A., Dudek, G.: Trust-driven interactive visual navigation for autonomous robots. In: Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA'12), pp. 3922–3929 (2012)
21. Yagoda, R.E., Gillan, D.J.: You want me to trust a ROBOT? the development of a humanrobot interaction trust scale. Social Robotics **4**(3), 235–248 (2012)