Difficulties in Perceiving and Understanding Simulated Robot Reliability Changes in a Sequential Binary Task

Hiroshi Furuya
hiroshi.furuya@ucf.edu
University of Central Florida
Orlando, Florida, USA

Laura Battistel
laura.battistel@eurac.edu
University of Trento, Eurac Research
Trento, Italy

Zubin Choudhary
zubin.choudhary@ucf.edu
University of Central Florida
Orlando, Florida, USA

Matt Gottsacker
mattg@ucf.edu
University of Central Florida
Orlando, Florida, USA

Gerd Bruder
bruder@ucf.edu
University of Central Florida
Orlando, Florida, USA

Gregory F. Welch
welch@ucf.edu
University of Central Florida
Orlando, Florida, USA

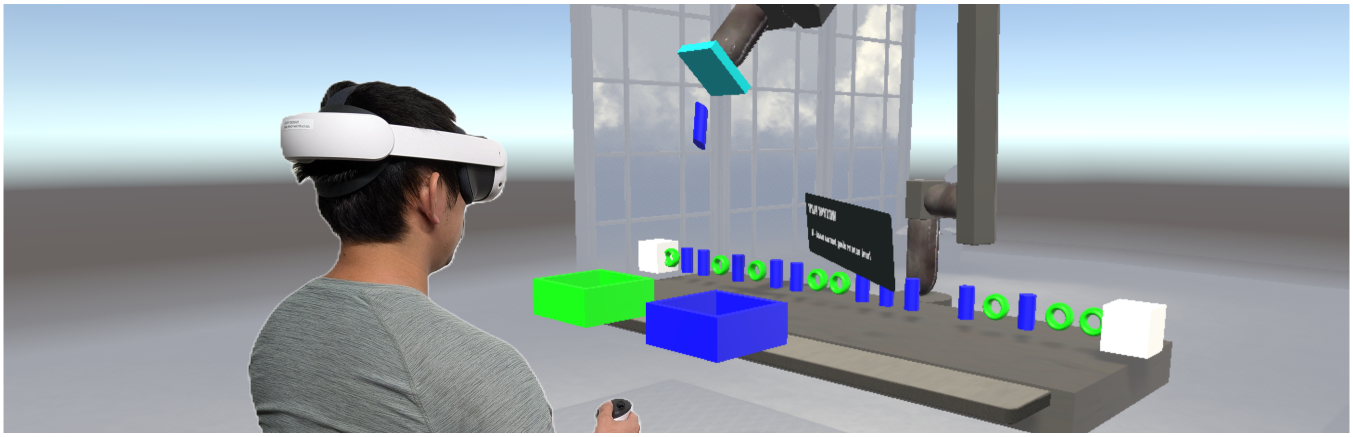

Figure 1: Composite image of participant monitoring a simulated robotic arm in the experimental virtual reality environment.

ABSTRACT

Human-robot teams push the boundaries of what both humans and robots can accomplish. In order for the team to function well, the human must accurately assess the robot's capabilities to calibrate the trust between the human and robot. In this paper, we use virtual reality (VR), a widely accepted tool in studying human-robot interaction (HRI), to study human behaviors affecting their detection and understanding of changes in a simulated robot's reliability. We present a human-subject study to see how different reliability change factors may affect this process. Our results demonstrate that participants make judgements about robot reliability before they have accumulated sufficient evidence to make objectively high-confidence inferences about robot reliability. We show that this reliability change observation behavior diverges from behavior expectations based on the probability distribution functions used to describe observation outcomes.

CCS CONCEPTS

• **Computing methodologies** → **Virtual reality**; • **Human-centered computing** → **User studies**; **Virtual reality**; • **Computer systems organization** → **Robotics**.

KEYWORDS

Human-robot interaction, reliability, trust, virtual reality

ACM Reference Format:
Hiroshi Furuya, Laura Battistel, Zubin Choudhary, Matt Gottsacker, Gerd Bruder, and Gregory F. Welch. 2024. Difficulties in Perceiving and Understanding Simulated Robot Reliability Changes in a Sequential Binary Task. In *ACM Symposium on Spatial User Interaction (SUI '24), October 7–8, 2024, Trier, Germany.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3677386.3682083

1 INTRODUCTION

Human-robot teams are often considered the future of human operations in space exploration, defense, search-and-rescue, and many other important applications [4, 10, 40, 43]. In these and other domains, trust, commonly defined as "the attitude that an agent will

ACM SUI '24, October 07–08, 2024, Trier, Germany

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/3677386.3682083

help achieve an individual's goals in a situation characterized by uncertainty and vulnerability," is one of the most important determinants of the human-robot team's success [38]. Human-robot trust is in turn pivotally affected by the human's perception of the robot's performance [25]. Incorrect perception of the robot's performance will negatively affect the calibration between trust and actual robot capabilities, which can jeopardize the team [49].

Investigating factors governing HRI such as perceived performance can be difficult to perform with physical robots due to high technical overhead and safety considerations [16]. In response, the HRI community has widely adopted the use of simulated robots in VR environments to enable growth in HRI research while avoiding issues with the use of real-world robots [15, 20, 55].

In this paper, we investigate aspects of human perception and understanding of changes in reliability of a simulated robot in VR. As the notion of performance and reliability in a robot can be complex [11], we utilized a binary sorting task to abstract performance and reliability to an easy to understand quantity: How accurately does the robot determine to which of two bins an object belongs? We used this task in a user study to investigate how quickly and how accurately humans make judgments about changes in a robot's reliability level.

The main contributions of this work are:

(1) We describe a gap in current HRI literature on the trust and human-robot teaming related to perceived performance.
(2) We present a human-subject experiment (N=20) demonstrating participant behaviors leading to poor performance in detecting robot reliability changes.
(3) We discuss our results and their implications for future work in HRI.

The remainder of this paper is structured as follows: first we discuss concepts and related work that form the background for our paper in Section 2. Then we describe our experimental design in Section 3. We then present our results in Section 4. In Section 5 we provide a general discussion of these results. Finally, we conclude the paper in Section 6.

2 BACKGROUND

In this section, we discuss concepts and prior work that inform our investigation of human detection of changes in robot reliability. These include efforts using VR for HRI research, the role of perceived performance in trust models in HRI, and methods for capturing human perception and understanding of change.

2.1 Human-Robot Interaction in Virtual Reality

VR and HRI research have converged to explore novel methods for robotic design, programming, force feedback in virtual environments, and human operation of robots [8]. VR offers significant advantages in studying HRI due to its ability to provide a controlled, repeatable, and safe experimental environment. Tang and Yamada [56] demonstrated that VR can be used to provide safer and more effective ways to operate construction robots. Additionally, the depth perception afforded by stereoscopic VR has been shown to improve human observer performance in understanding the motion of a robot compared to a non-stereo display [42]. The high level of control available in programming virtual environments makes VR a useful tool for investigating HRI questions related to human factors [54]. For instance, it is possible to programmatically manipulate virtual robot performance to investigate the effect of changes in robot reliability on human factors, such as trust [22]. Robinette et al. performed an experiment using simulated robots in a virtual emergency evacuation task to demonstrate that poor robot performance led to lower levels of trust and less frequent choices to depend on the robot [51]. Additionally, VR allows evaluating human-robot interactions that would be prohibitive to conduct in the real world due to material and safety considerations [19]. For example, Mara et al. implemented a large-scale industrial-style virtual environment where user studies can be conducted in VR for investigating HRI topics [44]. VR and simulated robots have also been successfully used to train humans in HRI, including in medical [39, 46] and industrial contexts [50].

2.2 Perception of Robot Performance in Human-Robot Interaction

HRI researchers have modeled human perceptual and cognitive processes when interacting with robots. For instance, Boos et al. presented an information processing model linking human perception, comprehension, and action in response to cues from a robot [7]. Honig et al. developed a model for human understanding of failures in HRI, including separate steps for perceiving and comprehending failures [28]. Observation of a robot's performance is a critical step in the formation of human trust in robotic systems [38]. Hancock's influential meta-analysis of trust factors in HRI [25], supported with a follow-up meta-analysis [26], highlights the central role of perceived robot reliability in human-robot trust. Khavas et al.'s survey also found that robot performance is an important factor in determining the quality of HRI, and that performance-based models are often used as a feedback source for improving subsequent human-robot interactions [32]. Moreover, researchers have demonstrated that operator trust in robots changes as the reliability of the robot changes [9, 29].

However, a complicating factor for these models is the uncertainty regarding the amount of evidence required for a human to achieve an accurate comprehension of reliability and whether human perception of performance is reasonably aligned with the robot's actual performance. Prior work has demonstrated fatigue resulting from extended vigilance results in decreased trust in autonomous systems [24]. Perceived automation reliability has also been shown to vary inversely with human monitoring performance [48], meaning that incorrect monitoring can lead to perceptions of better performance. The kind of evidence humans observe also has been shown to affect human trust. For example, Yang et al. [65] demonstrated that events diminishing trust have greater impacts on trust than events that would repair it. To more fully understand the human factors involved in forming trust and reliability judgments, it is necessary to examine how human make reliability judgments and how those judgments vary over time. In HRI, inaccurate perceptions of robot reliability lead to poor team outcomes.

2.3 Temporal Dynamics of Trust in Human-Robot Interaction

It is well-established that trust is a time-dependent construct, but the underlying temporal dynamics are not well understood, and so new experimental methodologies are required [61]. Lee and Moray proposed modeling trust using time series representations [37]. Xu and Dudek [63] modeled real-time changes in trust using robot performance relying on episodic trust measurements, where the instrument is administered after observation of an event taking place over thirty to sixty seconds [63]. Trust has also been modeled and measured as a result of a series of binary outcomes [23], similar to the sorting outcomes in our work. There is a growing awareness and importance placed on the idea that HRI unfolds over various scales of time [57], such as over the course of a single task [14]. Other works further focus on developing the notion of "trust dynamics," time-dependent changes in trust, over micro-time scales over the course of individual interactions in HRI. For example, Li et al. [41] used conversation analysis to measure predictors of trust during an interaction. Bhat et al. demonstrated the use of a slider to record self-reported trust on a moment-to-moment basis over the course of an experiment [6]. Guo and Yang presented a similar experiment employing a high number of trials with frequent trust reporting to capture granular changes in trust [21]. Kintz et al. presented a human-subject study exploring ways to estimate trust in real time based on a task-dependent model of human actions [33].

Among the numerous definitions and models of trust, one common feature is a feedback loop where observations and interactions with the robot informs subsequent human trust evaluations of the robot [34]. Of the numerous factors in these feedback loops, the most influential is typically considered to be the performance of the robot [25]. It is reasonable to assume that perceiving and understanding the robot's performance takes time, with additional time added for cognitive processes to resolve and update the human's trust in the robot. While substantial literature exists on developing our understanding of the potential effects of various robot performance outcomes [12], it is typically assumed that users have enough time to understand these events and have correctly perceived them. Our study aims to shed light on this assumption by taking time and accuracy measurements of participants' observation behaviors (i.e., how long it took them to detect and identify changes to robot reliability, and whether their observations were correct).

3 EXPERIMENT

In this section, we describe the experiment we conducted using a simulated robot in VR to study participants' performance in monitoring for changes in the robot's reliability. The study protocol was approved by the institutional review board of our university under protocol number: SBE-17-13446.

3.1 Participants

Estimating sample sizes for Generalized Linear Mixed Models (GLMMs), the use of which we will discuss in Sec. 3.5, is notoriously difficult [35]. This is further complicated by the fact that we did not decide on a particular model apriori to use for non-analytical power analysis techniques such as simulation. To get a rough estimate for an appropriate sample size, we instead performed an apriori power analysis in G*Power [17] for a repeated measures, within-subjects ANOVA. We utilized a large effect size, $\eta_p^2 = 0.25$, based on prior work in HRI that demonstrates that differences in robot performance have large effects on human trust [59]. Using standard values of $\alpha = 0.05$ and power of $1 - \beta = 0.8$, we calculated a required sample size of 18 participants. Furthermore, recent work in our community involving HRI (e.g., [64]) utilizes similar sample sizes. Based on these factors, we chose to recruit 20 participants (14 male, 6 female), all students from our university community. All of the participants had normal or corrected-to-normal vision. None of the participants reported any visual or vestibular disorders, such as color or night blindness, dyschromatopsia, or a displacement of balance. 18 participants had used a VR head-mounted display (HMD) before. Participants received monetary compensation for their participation. The experimental task and the questionnaires took participants approximately 60 minutes to complete.

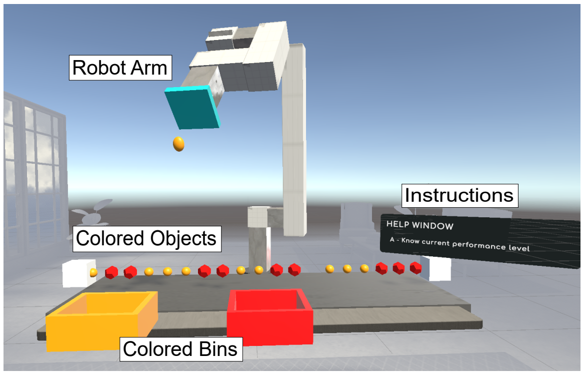

Figure 2: Virtual environment setup consisting of 6DOF robot arm, colored objects, two colored bins and a billboard text window for the instructions. Objects flowed from the left side of the conveyor belt to the right.

3.2 Materials

3.2.1 Experimental Setup. Participants were seated and immersed in a virtual environment via the Meta Quest 3, which provides a horizontal field of view of approximately 110 degrees and a vertical field of view of approximately 96 degrees, as well as a native resolution of 2064 × 2208 per eye at a refresh rate of 120 Hz. Participants were provided with the right controller only. We used a separate computer for participants to answer the questionnaires. The virtual environment application was deployed to the HMD. Development for the virtual environment application was performed using Unity version 2022.3.

3.2.2 Virtual Environment. The virtual environment consisted of a background environment, a conveyor belt asset, a cube situated at each end of the conveyor belt, a 6 degree-of-freedom (6DOF) robot arm, two colored bins, and a billboard text window used to deliver instructions. The background environment was adapted from sample assets provided as part of the Meta XR All-in-One SDK [1]. The conveyor belt asset was created using a long box and a black rubber texture. The cubes at each end of the conveyor belt functioned as the start and end points for objects that travel down the conveyor belt. The 6DOF robot arm was created using Unity primitive shapes and utilized a 6DOF robot arm inverse kinematics solver and linear joint interpolation to simulate realistic robot arm motion. The two colored bins served as sorting targets for the robot to sort objects into. The billboard text window provided instructions for participants to follow. Figure 2 depicts this environment.

3.2.3 Simulated Robot Motion, Appearance, and Sound Effects. Each sorting repetition, hereon referred to as a 'pick,' begins with the robot at a neutral rest position. The robot then moves the end effector down to pick up an object. After picking up an object, it returns to the rest position. It then moves the end effector to place the object into one of the bins. Finally, it returns back to the rest position to finish the pick animation. This process took four seconds in total, referred to as the pick cycle time. Importantly, while the pick cycle time is four seconds, it may take less than four seconds for an observer to perceive a pick, as they may be able to predict the path of the robot after observing part of the motion. We selected four seconds as the pick cycle time as pilot testing revealed that shorter cycle times made it more difficult for participants to maintain attention and observe the results of each action. This is in line with prior work that demonstrates that decreasing robot work pace eases perceived cognitive and temporal demands of humans tasked with observing the robot's performance [60]. Therefore, we chose to use a slower pick cycle time to mitigate problems that may arise from extended periods of high perceived workload.

Pilot testing also revealed a risk that participants would not correctly perceive each trial (see Section 3.3.2) as independent if the robot looked and behaved the same way each trial. To mitigate this risk, we implemented several different shapes and materials for the robot arm links, robot sound effects, shapes and colors of objects to be sorted, and paths the robot would take between its rest position and picking up a new object. In each trial, a random combination of the above elements would be assigned, resulting in a robot that looked, sounded, and moved differently than robots in other trials. To ensure that these changes would not affect participant perception, the actual length and size of the robot, the pick cycle time, approximate object size, and placement of objects and destination bins were kept constant. Furthermore, all shapes, materials, sounds, and motion paths for the robot were similar to each other in realism.

3.3 Methods

3.3.1 Study Design. We ran a within-subjects design study for our experiment with the following two factors:

- **Reliability Change Magnitude (3 levels)** — Zero, Small (25%), Large (50%)
- **Reliability Change Direction (3 levels)** — None, Up (+), Down (-)

These factors were implemented using a 3×3 square of initial and final reliability values: 25% (Low), 50% (Medium), and 75% (High). This resulted in 9 different reliability conditions. These conditions were mapped to the levels for each reliability change factor by the direction of change between the initial and final value (i.e., increase or decrease) and the magnitude of that change (i.e., 0% difference is "zero," 25% is "small," and 50% is "large"). The mapping between these conditions and the two reliability change factors can be seen in Table 1. This mapping results in 3 conditions for each reliability change direction level, 3 conditions for zero change magnitude, 4 conditions for small change magnitude, and 2 conditions for large change magnitude. For example, the 50%-75% condition is labeled in Table 1 as (Small, Up). During the trials for this condition, the robot would initially sort items with 50% accuracy, then increase (see Section 3.3.2) to 75% accuracy for the experimental manipulation.

Each condition was repeated 3 times, resulting in 27 experimental trials. Results from pilot testing indicated a significant risk of participants learning the 3 different reliability values (i.e., Low, Medium, and High), so we added an additional 10 distractor trials of random initial and final reliability values to mask the experimental values. Trial order was simply randomized by adding all trials, including distractor trials, into a single list and randomly shuffling the order of list elements. This shuffling was performed separately for each participant.

3.3.2 Procedure. Upon arrival, participants read through the consent form, and were asked to give their verbal consent to participate in the experiment. We assigned them a participant ID, and asked them to complete a demographics survey and the Simulator Sickness Questionnaire [31]. Participants were then briefed about the experiment and then donned the headset for a tutorial session. The tutorial explained the experiment and took participants through three practice trials demonstrating no reliability change, an increase in reliability, and a decrease in reliability, respectively. These trials were not counted as part of the experimental data set and served to familiarize and "warm up" participants with the setting and the task. After the tutorial, the experimenter debriefed and asked the participant to explain the task in their own words. The experimenter issued corrections or clarifications as needed. Then participants completed all trials, taking breaks between trials as needed. In each trial, participants observed the robot sorting items at an initial reliability level. Participants pressed the 'A' button on the controller when they felt like they understood this initial reliability level well. After pressing the button, the robot could change its reliability level, depending on the particular condition the trial was assigned to. Participants continued to observe the robot and pressed the "Trigger" button as soon as they noticed a change in reliability (this recorded *Detection Time*, see Section 3.3.3). Participants were briefed that noticing a change may occur without fully understanding the new reliability level. Participants then continued to observed the robot until they felt confident about the robot's new reliability level. At this point, participants pressed the "Trigger" button (this recorded *Identification Time*, see Section 3.3.3). Finally, participants used the joystick and the 'A' button to respond to a prompt asking them to report whether the perceived change in reliability was an increase or a decrease (see *Correctness* in Section 3.3.3). For all three inputs recording DT, IT, and direction of perceived change, respectively,

Table 1: Mapping from reliability values to reliability change variable levels for: (Magnitude, Direction)

		Initial Reliability		
		25%	50%	75%
Final Reliability	25%	(Zero, None)	(Small, Down)	(Large, Down)
	50%	(Small, Up)	(Zero, None)	(Small, Down)
	75%	(Large, Up)	(Small, Up)	(Zero, None)

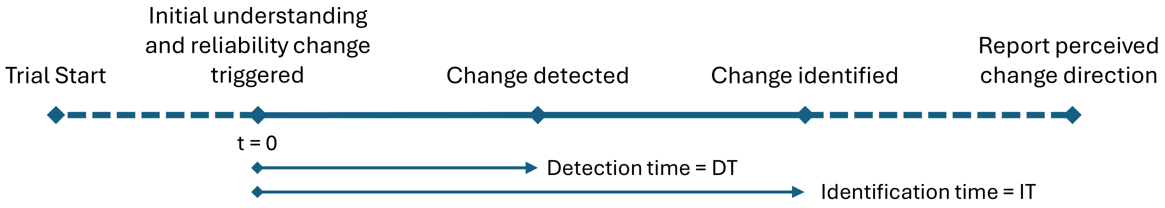

Figure 3: Timeline illustrating each trial and the measures DT and IT.

the participant could also press the 'B' button to indicate that they perceived no change. After these inputs were recorded, the trial ended. Participants were prompted to press the 'A' button when they felt ready to begin the next trial, allowing them to take a short break between trials. After completing all the trials, participants answered the post-experience Simulator Sickness Questionnaire. Participants then received monetary compensation.

3.3.3 Measures. To explore human time-dependent behavior in perceiving and understanding performance changes of a robot, we measured the following:

- **Detection time (DT)** — The time taken by participants to detect that the robot's reliability has changed. This is the time between the robot changing its reliability level and participants first noticing a change in reliability.
- **Identification time (IT)** — The time taken by participants to identify the robot's new reliability level. This is the time between the robot changing its reliability level and participants feeling like they confidently understood the robot's new reliability level.
- **Correctness** — If the reported direction of change in reliability (increase, decrease, or none) was correct. In aggregate, this is the percent of correct reports of direction of change in reliability.

Both DT and IT are calculated relative to the moment when the reliability of the robot changed, which was in turn triggered by participants indicating that they had understood the robot's initial reliability level as discussed in Section 3.3.2. This moment was instantaneous, implemented as a change in a parameter used to simulate the robot. In Section 3.5 we further interpret DT and IT not as times, but the number of observations until the recording of DT and IT. Figure 3 illustrates how DT and IT are calculated. It is of note that the reference time for calculating DT and IT is itself some time after the start of trial. This time between the start of the trial and triggering the change in reliability was not measured, as the task of understanding the robot's initial reliability is not the focus of this study. We were instead interested in measuring the time for participants to perceive and understand a change in reliability from this initial state. In addition, from DT and IT we can calculate the time between the two reports, hereon referred to as BT for 'between time,' as follows:

$$BT = IT - DT \tag{1}$$

3.4 Hypotheses

The following hypotheses were formulated based on expectations that we used to develop our experimental design:

H1 The difference between DT and IT is greater than zero.
H2 DT decreases with increasing magnitude of change ($Large < Small < Zero$).
H3 DT is lower for decreases in reliability than for increases ($Down < Up$).
H4 IT is the same for all conditions ($Large = Small = Zero$; $Down = Up = None$).
H5 Correctness increases with increasing reliability change magnitude.
H6 Correctness has a direct relationship with IT.

Hypothesis **H1** is informed by information process models in HRI, where perceiving and understanding robot cues are separate steps (see Section 2) and therefore could be expected to each take a detectable amount of time. In this experiment, DT would correspond to the time taken to perceive the cue (i.e., change in reliability) and IT corresponds to the time taken to understand the cue.

Hypothesis **H2** is informed by signal detection concepts, where greater magnitude signals are easier to perceive.

Hypothesis **H3** is informed by trust literature that demonstrates a stronger effect of trust diminishing events on trust than trust building events [65]. We are interested in observing if this effect extends to perception of reliability, which is an antecedent to trust [25].

Hypothesis **H4** is informed by the expectation that strategies for estimating binomial probability rates do not differ by change characteristics, i.e. a naive strategy is to simply observe the robot for a set number of repetitions and make a decision based on the number of successes observed. Such a strategy is not affected by the direction or magnitude of the reliability change.

Hypothesis **H5** is related to H2, where greater stimuli are easier to detect.

Hypothesis **H6** is based on the expectation that longer observation allows for greater accumulation of observations to inform more accurate assessments of reliability change.

3.5 Analysis

For DT, IT, and Correctness results, GLMMs [18, 53] were used to assess the relationship between reliability change factors and DT, IT, and Correctness. We used GLMMs because they allow modeling of data that do not follow a normal distribution through the selection of probability distributions that are more suitable for the data's distribution.

Table 2: Descriptive statistics for DT, IT, and Correctness for all reliability pairs

		DT (s)			IT (s)			Correctness
		Q1	Median	Q3	Q1	Median	Q3	Mean
Reliability Pair (%)	25–25	5.8	17.6	26.9	11.1	21.7	32.9	0.450
	25–50	6.7	16.0	25.8	12.7	26.0	37.7	0.407
	25–75	7.8	15.9	21.2	14.1	21.2	33.1	0.633
	50–25	4.9	14.7	23.5	10.4	21.1	33.9	0.467
	50–50	7.6	14.9	22.4	12.5	22.7	36.8	0.233
	50–75	8.1	16.7	27.8	11.8	22.2	35.6	0.500
	75–25	7.8	12.4	20.1	12.6	19.7	27.2	0.700
	75–50	8.9	15.5	22.8	12.3	24.5	33.3	0.542
	75–75	8.7	16.2	24.0	13.0	25.1	36.1	0.400

For each measure, we first examined density plots to determine the appropriate distribution family to use. Then, we compared different models using different independent variables as fixed effects and independent variables and participant IDs as random effects. Not all possible models converged and many resulted in singular fits, including most models with multiple fixed effects or random slopes. We removed these from consideration, and compared the remaining ones using AIC [2] and BIC [52] values to choose the best fitting model, i.e., the model with lowest AIC and BIC values, then fit the model by maximum likelihood estimation. We evaluated significance using the Wald Chi-Square test. For models with significant Wald Chi-Square results, we performed pairwise t-tests with Tukey HSD correction as necessary. Selected models were evaluated using goodness-of-fit measures, including Marginal R^2 and Conditional R^2, variance explained by fixed effects and variance explained by both fixed and random effects, respectively [47].

The experiment in this paper was designed to help us better understand the nature of the relationship between robot reliability factors and participant perception behavior, a task for which R-statistics are well-suited as the summary statistic [45]. Thus, the above Marginal and Conditional R^2 values serve as the effect size indices based on variance explained. These R^2 values are analogous to the η^2 summary statistic used to describe effect sizes in ANOVA analyses [30, 47]. Analysis was performed using R 4.4.0 using the lme4 package for fitting models, emmeans package for evaluating significance, and MuMIn package for evaluating goodness of fit.

We also explored implications of DT and IT by dividing them by 4 seconds, the amount of time required for the robot to perform one repetition of its task. In doing so we always round up the result, as it may be possible for participants to correctly predict the result of any task repetition by closely observing the robot's motion, i.e., observing to see which bin the robot is moving the currently held object towards. In such a case, it would take less than the prescribed 4 seconds for a participant to observe a new outcome in the robot's sorting task.

4 RESULTS

In this section we present the results from our experiment and analysis. We primarily report DT and IT using median and interquartile range due to their better suitability compared to mean and standard deviation for describing distributions similar to the gamma distribution [62]. Table 2 displays descriptive statistics for DT, IT, and Correctness, aggregated by experimental reliability pair.

In the course of interpreting the data, we also present DT and IT in terms of the number of picks, based on the pick cycle time described in Section 3.2.3. Because participants could potentially observe or predict a pick before the full pick cycle time elapses, the formula for computing picks using the ceiling function, which rounds numbers up to the next whole number, is as follows:

$$Picks_{DT,\ IT} = \left\lceil \frac{Time_{DT,\ IT}}{Time_{Cycle}} \right\rceil \tag{2}$$

4.1 Analysis of Detection Time

Table 3 shows descriptive statistics for DT by reliability change direction. We chose to aggregate by reliability change direction due to the following GLMM analysis.

Following the process described in Section 3.5, we found that the data distribution for the continuous time variable was clearly

Table 3: Descriptive statistics for DT by Direction of Reliability Change (None, Up, Down)

	Direction					
	None		Up		Down	
	Time (s)	Picks	Time (s)	Picks	Time (s)	Picks
Q1	10.1	3	9.7	3	8.4	3
Median	16.3	5	16.4	5	15.4	4
Mean	17.8	5	18.0	5	16.1	5
Q3	24.7	7	24.7	7	20.0	6

Table 4: GLMM employed to analyse the effect of direction of reliability change on DT.

Model Declaration	AIC (%)	BIC (%)
DT ~ Condition + (1 | PID)	11235 (2.3%)	11283 (<0.1%)
DT ~ Direction + (1 | PID)	**11229 (61.2%)**	**11250 (79.4%)**
DT ~ InitialReliability + (1 | PID)	11233 (9.0%)	11283 (11.7%)
DT ~ InitialReliability + Direction + (1 | PID)	11232 (14.6%)	11262 (0.3%)
DT ~ InitialReliability + Direction + (Direction | PID)	11233 (6.3%)	11285 (<0.1%)
DT ~ Magnitude + (1 | PID)	11233 (6.6%)	11255 (8.6%)

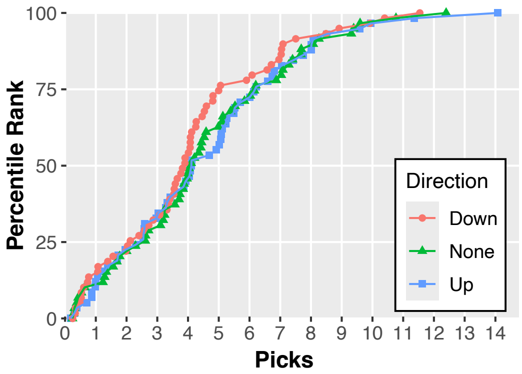

Figure 4: Percentile rank for DT as number of picks, grouped by reliability change direction

mounded and right-skewed, leading to the selection of the Gamma distribution family with log-link function. AIC (weighted 61.2%) and BIC (79.4%) agreed on selecting a model with reliability change direction as fixed effect and participant ID (PID) as random effect, as shown in Equation 3.

$$DT \sim Direction + (1|PID) \tag{3}$$

Table 4 shows the other models that were compared to this one with corresponding AIC and BIC values. This result **does not support H2**, the hypothesis that DT is significantly affected by different magnitudes of change in reliability; instead we found that a model with change direction as fixed effect is best. Furthermore, using a Wald Chi-Square test, we do not find a significant effect of reliability change direction on DT; $\chi^2(2)$ = 4.66, p = 0.097. This **does not support H3**, the hypothesis that DT is significantly affected by reliability change direction. Goodness of fit was also low (marginal R^2 = 0.3%, conditional R^2 = 33% using the trigamma function).

For the same reasons discussed in Sec. 4, our data is well-suited for aggregation by percentile ranks, or the percentage of samples at or less than the value at that rank. For example, a rank of 25% indicates that 25% of points are at or below the given value [62]. Q1 corresponds to the 25% percentile rank, median to the 50% percentile rank, and Q3 to the 75% percentile rank. Figure 4 shows a plot of percentile rank of DT by number of picks.

4.2 Analysis of Identification Time

Table 5 shows descriptive statistics for IT by reliability change direction. We chose to aggregate by reliability change direction due to the following GLMM analysis.

Following the process described in Section 3.5, we found that the data distribution for the continuous time variable was clearly mounded and right-skewed, leading to the selection of the Gamma distribution family with log-link. Just as for IT, AIC (50.0%) and BIC (56.9%) agreed on selecting a model with reliability change direction as fixed effect and participant ID as random effect, as shown in Equation 4.

$$IT \sim Direction + (1|PID) \tag{4}$$

Table 5: Descriptive statistics for IT by Direction of Reliability Change (None, Up, Down)

	Direction					
	None		Up		Down	
	Time (s)	Picks	Time (s)	Picks	Time (s)	Picks
Q1	16.3	5	14.3	4	15.5	4
Median	26.7	7	26.1	7	21.7	5
Mean	26.4	7	25.6	7	25.0	6
Q3	31.5	8	33.6	9	31.2	8

Table 6: GLMM employed to analyse the effect of direction of reliability change on IT. Akaike weight percentages are included with the raw values.

Model Declaration	AIC (%)	BIC (%)
IT ~ Condition + (1 | PID)	11569 (0.7%)	11616 (<0.1%)
IT ~ Direction + (1 | PID)	**11560 (50.0%)**	**11582 (56.9%)**
IT ~ Direction + (Direction | PID)	11568 (1.2%)	11611 (<0.1%)
IT ~ InitialReliability + (1 | PID)	11563 (14.5%)	11584.1 (16.5%)
IT ~ InitialReliability + Direction + (1 | PID)	11563 (10.4%)	11593 (0.2%)
IT ~ Magnitude + (1 | PID)	11562 (23.2%)	11583 (26.4%)

Table 6 shows the other models that were compared to this one with corresponding AIC and BIC values. Using a Wald Chi-Square test, we do not find a significant effect of reliability change direction on IT; $\chi^2(2)$ = 2.65, p = 0.266. Goodness of fit was also low (marginal R^2 = 0.2%, conditional R^2 = 35% using the trigamma function). These results **support H4**, that IT is not significantly affected by direction and magnitude characteristics of the change in reliability.

4.3 Analysis of Difference Between Detection and Identification Time

We expected BT to be non-zero as a reflection of the time required to proceed from one stage in information processing to another (see Section 3.4). To evaluate this, we first observed the data distribution, which appeared mounded and right-skewed just like DT and IT, matching a gamma distribution. Descriptive statistics for BT are as follows: Q1 = 0.93s, Median = 3.63s, Mean = 8.17s, Q3 = 11.76s. These

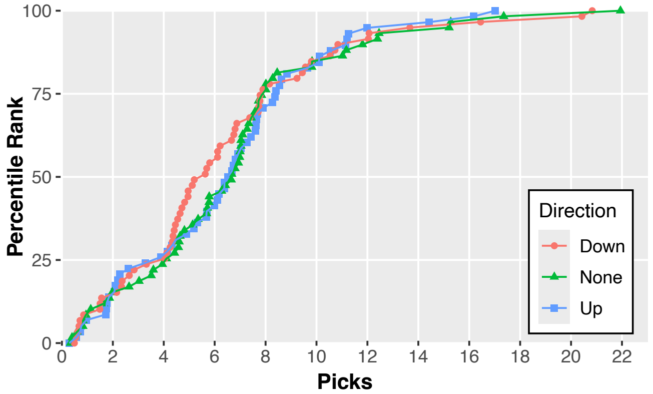

Figure 5: Percentile rank for IT as number of picks, grouped by reliability change direction

figures are clearly longer than the constant time required to simply press the button on the controller, which would be expected if there were no time spent on additional cognitive processes between information processing stages. To further assess the hypothesis that the mean of the data set was non-zero, we ran a one-sample t-test that confirmed that the true mean of BT is non-zero; t(537) = 16.57, **p < 0.001**. This **supports H1**, the hypothesis that there is a non-zero time period between perceiving and understanding a change in a robot's reliability.

4.4 Analysis of Correctness

As the correctness measure is a repeated measure with binary outcomes, we chose the binomial family with probit link function. AIC (69.6%) and BIC (99.4%) agreed on selecting a model with reliability change magnitude as fixed effect and participant ID as random effect, as shown in Equation 5.

$$Correctness \sim Magnitude + (1|PID) \tag{5}$$

Table 7 shows the other models that were compared to this one with corresponding AIC and BIC values. Using a Wald Chi-Square test, we found a significant effect of reliability change magnitude on Correctness in reporting the direction of reliability change; $\chi^2(2)$ = 27.05, **p < 0.001**. Pairwise comparison using the Tukey HSD test indicated significant differences between Zero and Small magnitudes (β = -0.31, SE = 0.13, **p = 0.039**), Zero and Large magnitudes (β = -0.80, SE = 0.15, **p < 0.001**), and Small and Large magnitudes (β = -0.49, SE = 0.15, **p = 0.002**). Goodness of fit was low (marginal R^2 = 5.1%, conditional R^2 = 6.9% using the delta method). This **supports H5**, that increasing magnitude characteristics of reliability change increases the rate of correct identification of the change in reliability.

Mean correctness was 36.1% for the Zero condition, 47.9% for the Small condition, and 66.7% for the Large condition.

We also performed a linear regression using the lm function in R to see if Correctness and IT have a linear relationship, i.e. if longer identification times, and therefore observation times, correlate with rates of correctly identifying changes in reliability. Results demonstrated a significant regression, F(1,18) = 15.71, **p < 0.001** with adjusted R^2 = 0.435. This result **supports H6**, the hypothesis that Correctness and IT correlate.

4.5 Other Exploratory Analyses

The low marginal R^2 values estimated for the models for all three measures motivated some exploratory analyses investigating other factors that may provide more explanatory power. We first explored the effect of trial number, an indication of how far through the experiment the participant was, on DT and IT:

$$\begin{aligned} DT &\sim TrialNumber + (1|PID) \\ IT &\sim TrialNmuber + (1|PID) \end{aligned} \tag{6}$$

These models were favored over previously assessed models both in AIC and BIC (DT: 11221 and 11238, respectively; IT: 11539 and 11556, respectively). A Wald Chi-Square test found a significant effect of trial number on DT; $\chi^2(1)$ = 10.96, **p < 0.001**; marginal R^2 = 0.08%, conditional R^2 = 36.6% using the trigamma function. A Wald Chi-Square test further found a significant effect of trial number on IT; $\chi^2(1)$ = 23.92, **p < 0.001**; marginal R^2 = 0.17%, conditional R^2 = 36.4% using the trigamma function.

Table 7: GLMM employed to analyse the effect of direction of reliability change on Correctness.

Model Declaration	AIC (%)	BIC (%)
Correctness ~ Condition + (1 | PID)	726 (27.5%)	769 (<0.1%)
Correctness ~ Direction + (1 | PID)	735 (0.4%)	752 (0.5%)
Correctness ~ InitialReliability + (1 | PID)	744 (<0.1%)	761 (<0.1%)
Correctness ~ InitialReliability + Direction + (1 | PID)	731 (2.5%)	756 (<0.1%)
Correctness ~ Magnitude + (1 | PID)	**724 (69.6%)**	**741.2 (99.4%)**

5 DISCUSSION

5.1 Differences Between DT and IT

Our results on BT align with information processing models of trust in HRI, as we observed a time difference between self-reported perception of a change (DT) and perceived understanding of the nature of the change (IT). We can interpret this difference as the time and evidence required for participants to transition from mere perception of a reliability change cue to comprehending the characteristics of the cue, i.e., the direction of reliability change. In the real world, for more complicated tasks in situations with greater uncertainty, we may expect longer time intervals and greater amounts of evidence required for humans to transition between perception of an HRI cue, understanding it, and experiencing subsequent changes in their mental model of the robot, to include trust. It will be important to further explore factors that affect these processing intervals to better understand how unpredictable events may unexpectedly shape HRI due to interrupted or incomplete processing of changes in robot state.

5.2 Participants Make Reliability Judgments with High Error Risk

Our results also shed light on not just the timing of perceiving and understanding cues, but also on the character of how participants completed the task of understanding changes in robot reliability as a whole. Participants declared understanding the change in robot reliability after a median of 5 to 7 observations (19.7 to 26.0 seconds, see Section 4 for discussion on the computation of Picks from DT and IT), depending on condition. Due to the simplicity of the sorting task, we were able to model theoretical outcomes of participant observations and compare them with participant behavior using the cumulative binomial probability function. The binomial probability distribution describes the probability of different binomial outcomes, such as the probability of observing 5 successes out of 7 picks. The cumulative binomial probability function describes the probability of observing less than or equal to a certain number of successes, such as the probability of seeing 5 or less successes out of 7 picks [13]. We can use this to estimate the likelihood of actually observing sequences of picks whose outcomes match the true reliability change, as seen in Table 8. We used the median number of picks at IT as the number of trials over which to evaluate our cumulative probabilities to see that there are only two conditions for which participant behavior could yield a

Table 8: Probability of observing outcomes matching final reliability

Reliability Pair	Median Observations	Successes Required	Probability of Observation
25%–25%	6	= 3	35.6%
25%–50%	7	≥ 2	77.3%
25%–75%	6	≥ 2	99.5%
50%–25%	6	≤ 2	83.1%
50%–50%	6	= 3	31.3%
50%–75%	6	≥ 4	83.1%
75%–25%	5	≤ 3	98.4%
75%–50%	7	≤ 5	93.8%
75%–75%	7	= 5	31.1%

theoretical change detection accuracy at or above the traditional 95% threshold, which represents the chance that an observation is within two standard deviations of the mean. It appears that participants frequently chose to make their reliability change judgments before accumulating enough observations to reach a high theoretical level of accuracy and thus assumed a higher level of risk. On the other hand, participants were instructed to declare their perceived changes in performance after reaching a "confident" level of understanding, potentially indicating that participants did not perceive this risk when completing the trials. In HRI and human-robot trust, the risk of inaccurate perception of robot reliability can directly lead to the risk of poor team outcomes and team failure. It is important to understand that in the real world, humans cannot always be monitored and forced to await further observations before making a decision. It is important to investigate the factors behind this behavior to better understand how to design HRI around risks that humans may take in assessing robot reliability.

5.3 Participants Perform Poorly Compared to Basic Probability Models

If a human were to use the cumulative binomial probability function described above to guide reliability change perception and understanding, they would simply observe a set number of task repetitions and to report perceived direction of reliability change based solely on the observed success rate. In such a case, we would expect outcomes of this strategy to match values from the cumulative binomial probability function, which evaluates the probability of observing a certain number of successes out of a series of independent binary events. For example, in the case of the 50% to 75% condition, we would expect participants to correctly detect the increase in performance if they observed at least 4 successes out of 6 observations (6 being the median number of picks observed for identification in that condition). The cumulative binomial probability of such observations occurring is 83.1%, implying that a robotic strategy of counting successes over repetitions of 6 observations would yield average correctness of 83.1%. Instead we see from Table 2 that participants' mean correctness for the 50%-75% reliability pair condition was approximately 50%. From the Tables 2 and 8 we can clearly see that participant performance does not match the probability distribution, further indicating that participant performance is poorer than would be expected. Strangely, despite this apparent degradation in performance compared to theoretical probability, we found a strong correlation between DT and IT, indicating that accumulation of more observations still does improve accuracy in identifying robot reliability changes.

From these results and the results discussed in Section 5.2, it appears that the process by which participants are making reliability change judgments is fundamentally different than theoretically expected. Participants do not appear to wait for a set of observations before committing to a decision on whether and in what direction reliability changed. Instead, participants appear to be driven by a few observations towards some unobserved decision threshold well below what would be expected if they used a binomial probability distribution model to make their decision.

5.4 Limitations and Future work

The serial observations of a binary sorting task we used in our study may not fully describe the way humans observe changes in robot reliability in the real-world. While it is clear that in the real world humans must perceive evidence in a time-sequenced manner, there may be numerous factors that may influence how reliability of the robot is understood. For example, in the real world, successive visual observations may be chunked and perceived as a single event [3], whereas in our experiment each sorting outcome is significantly temporally separated from the next in a way that may not reflect real world conditions. Future work can help clarify the extent to which real-world observations of robot performance may be modeled as sequential samples of binary outcomes.

It would be valuable to perform follow up work analyzing the effects of particular sequences of pick outcomes to better understand the effects of these micro-scale events on reliability change perception and understanding. Prior work also identified clusters of individual differences in trust dynamics [6, 65], suggesting that we may find individual differences that explain more of the variance in the data than what our analyses supported.

The gender distribution of our participants was skewed (14 male, 6 female), potentially limiting the generalization of our results. Although prior work has not found evidence to suggest a significant gender effect on trust in HRI [27, 36], the literature does represent evidence that when gender-relevant features are introduced to the robot, such as gendered anthropomorphism, differences based on human gender emerge [58]. Future work on perception of changes in performance, especially if they involve domains that carry strong gender associations or stereotypes [5], would benefit from better participant gender representation.

6 CONCLUSION

In this paper, we presented a user study (N=20), in which we investigated aspects of human perception and understanding of changes in reliability of a simulated robot in VR as one of the most important determinants for trust in robots. Specifically, we measured the time it took participants to detect a change and identify the direction of the change, as well as the correctness of this identified change in reliability. Our results show that participant behavior in perceiving and understanding robot reliability incurs high levels of risk in incorrectly perceiving the true robot reliability level and differs in character from the binomial cumulative probability distribution used to describe the actual outcomes. Future work may explore new methods for describing and modeling human performance in observing changes in robot reliability.

ACKNOWLEDGMENTS

This material includes work supported in part by the Office of Naval Research under Award Numbers N00014-21-1-2578 and N00014-21-1-2882 (Dr. Peter Squire, Code 34), and the AdventHealth Endowed Chair in Healthcare Simulation (Prof. Welch).

REFERENCES

[1] [n. d.]. Meta XR All-in-One SDK | Integration | Unity Asset Store. https://assetstore.unity.com/packages/tools/integration/meta-xr-all-in-one-sdk-269657

[2] Hirotugu Akaike. 1973. Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models. *Biometrika* 60, 2 (1973), 255–265. https://doi.org/10.2307/2334537 Publisher: [Oxford University Press, Biometrika Trust].

[3] Elkan G. Akyürek, Nils Kappelmann, Marc Volkert, and Hedderik van Rijn. 2017. What You See Is What You Remember: Visual Chunking by Temporal Integration Enhances Working Memory. *Journal of Cognitive Neuroscience* 29, 12 (Dec. 2017), 2025–2036. https://doi.org/10.1162/jocn_a_01175

[4] Michael J. Barnes, Jessie Y. C. Chen, Florian Jentsch, and Elizabeth S. Redden. 2011. Designing Effective Soldier-Robot Teams in Complex Environments: Training, Interfaces, and Individual Differences. In *Engineering Psychology and Cognitive Ergonomics*, Don Harris (Ed.). Springer, Berlin, Heidelberg, 484–493. https://doi.org/10.1007/978-3-642-21741-8_51

[5] Brock Bass, Meghan Goodwin, Kayla Brennan, Richard Pak, and Anne McLaughlin. 2013. Effects of Age and Gender Stereotypes on Trust in an Anthropomorphic Decision Aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1 (Sept. 2013), 1575–1579. https://doi.org/10.1177/1541931213571351

[6] Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. 2022. Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task. *IEEE Robotics and Automation Letters* 7, 4 (Oct. 2022), 8815–8822. https://doi.org/10.1109/LRA.2022.3188902

[7] Annika Boos, Olivia Herzog, Jakob Reinhardt, Klaus Bengler, and Markus Zimmermann. 2022. A Compliance–Reactance Framework for Evaluating Human-Robot Interaction. *Frontiers in Robotics and AI* 9 (May 2022). https://doi.org/10.3389/frobt.2022.733504 Publisher: Frontiers.

[8] Grigore C. Burdea. 1999. Invited review: the synergy between virtual reality and robotics. *IEEE Transactions on Robotics and Automation* 15, 3 (June 1999), 400–410. https://doi.org/10.1109/70.768174

[9] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with Trust for Human-Robot Collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery, New York, NY, USA, 307–315. https://doi.org/10.1145/3171221.3171264

[10] Committee on Human-System Integration Research Topics for the 711th Human Performance Wing of the Air Force Research Laboratory, Board on Human-Systems Integration, Division of Behavioral and Social Sciences and Education, and National Academies of Sciences, Engineering, and Medicine. 2022. *Human-AI Teaming: State-of-the-Art and Research Needs.* National Academies Press, Washington, D.C. https://doi.org/10.17226/26355

[11] Enrique Coronado, Takuya Kiyokawa, Gustavo A. Garcia Ricardez, Ixchel G. Ramirez-Alpizar, Gentiane Venture, and Natsuki Yamanobe. 2022. Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. *Journal of Manufacturing Systems* 63 (April 2022), 392–410. https://doi.org/10.1016/j.jmsy.2022.04.007

[12] Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12, 2 (May 2020), 459–478. https://doi.org/10.1007/s12369-019-00596-x

[13] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. 2020. *Mathematics for Machine Learning.* Cambridge University Press.

[14] Mustafa Demir, Nathan J. McNeese, Jaime C. Gorman, Nancy J. Cooke, Christopher W. Myers, and David A. Grimm. 2021. Exploration of Teammate Trust and Interaction Dynamics in Human-Autonomy Teaming. *IEEE Transactions on Human-Machine Systems* 51, 6 (Dec. 2021), 696–705. https://doi.org/10.1109/THMS.2021.3115058

[15] Morteza Dianatfar, Jyrki Latokartano, and Minna Lanz. 2021. Review on existing VR/AR solutions in human–robot collaboration. *Procedia CIRP* 97 (Jan. 2021), 407–411. https://doi.org/10.1016/j.procir.2020.05.259

[16] Mihai Duguleana, Florin Grigorie Barbuceanu, and Gheorghe Mogan. 2011. Evaluating Human-Robot Interaction during a Manipulation Experiment Conducted in Immersive Virtual Reality. In *Virtual and Mixed Reality - New Trends*, Randall Shumaker (Ed.). Springer, Berlin, Heidelberg, 164–173. https://doi.org/10.1007/978-3-642-22021-0_19

[17] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. https://doi.org/10.3758/BF03193146

[18] John Fox. 2015. *Applied regression analysis and generalized linear models.* Sage publications.

[19] Piotr Fratczak, Yee Mey Goh, Peter Kinnell, Andrea Soltoggio, and Laura Justham. 2019. Understanding Human Behaviour in Industrial Human-Robot Interaction by Means of Virtual Reality. In *Proceedings of the Halfway to the Future Symposium 2019 (HTTF 2019)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3363384.3363403

[20] Scott A. Green, Mark Billinghurst, XiaoQi Chen, and J. Geoffrey Chase. 2008. Human-Robot Collaboration: A Literature Review and Augmented Reality Approach in Design. *International Journal of Advanced Robotic Systems* 5, 1 (March 2008), 1. https://doi.org/10.5772/5664

[21] Yaohui Guo and X. Jessie Yang. 2021. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics* 13, 8 (Dec. 2021), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

[22] Kasper Hald, Katharina Weitz, Elisabeth André, and Matthias Rehm. 2021. "An Error Occurred!" - Trust Repair With Virtual Robot Using Levels of Mistake Explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 218–226. https://doi.org/10.1145/3472307.3484170

[23] Robert J. Hall. 1996. Trusting your assistant. In *Proceedings of the 11th Knowledge-Based Software Engineering Conference.* IEEE Comput. Soc. Press, Syracuse, NY, USA, 42–51. https://doi.org/10.1109/KBSE.1996.552822

[24] Peter A. Hancock. 2017. On the Nature of Vigilance. *Human Factors* 59, 1 (Feb. 2017), 35–43. https://doi.org/10.1177/0018720816655240

[25] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* 53, 5 (Oct. 2011), 517–527. https://doi.org/10.1177/0018720811417254

[26] Peter A. Hancock, Theresa T. Kessler, Alexandra D. Kaplan, John C. Brill, and James L. Szalma. 2021. Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors* 63, 7 (Nov. 2021), 1196–1229. https://doi.org/10.1177/0018720820922080

[27] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (May 2015), 407–434. https://doi.org/10.1177/0018720814547570

[28] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology* 9 (June 2018). https://doi.org/10.3389/fpsyg.2018.00861 Publisher: Frontiers.

[29] Sarah Hopko, Jingkun Wang, and Ranjana Mehta. 2022. Human Factors Considerations and Metrics in Shared Space Human-Robot Collaboration: A Systematic Review. *Frontiers in Robotics and AI* 9 (Feb. 2022). https://doi.org/10.3389/frobt.2022.799522 Publisher: Frontiers.

[30] Dawn Iacobucci, Deidre L. Popovich, Sangkil Moon, and Sergio Román. 2023. How to calculate, use, and report variance explained effect size indices and not die trying. *Journal of Consumer Psychology* 33, 1 (2023), 45–61. https://doi.org/10.1002/jcpy.1292

[31] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.

[32] Zahra Rezaei Khavas, S. Reza Ahmadzadeh, and Paul Robinette. 2020. Modeling Trust in Human-Robot Interaction: A Survey. In *Social Robotics*, Alan R. Wagner, David Feil-Seifer, Kerstin S. Haring, Silvia Rossi, Thomas Williams, Hongsheng He, and Shuzhi Sam Ge (Eds.). Springer International Publishing, Cham, 529–541. https://doi.org/10.1007/978-3-030-62056-1_44

[33] Jacob R. Kintz, Neil T. Banerjee, Johnny Y. Zhang, Allison P. Anderson, and Torin K. Clark. 2023. Estimation of Subjectively Reported Trust, Mental Workload, and Situation Awareness Using Unobtrusive Measures. *Human Factors* 65, 6 (Sept. 2023), 1142–1160. https://doi.org/10.1177/00187208221129371

[34] Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology* 12 (2021).

[35] Levi Kumle, Melissa L.-H. Võ, and Dejan Draschkow. 2021. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods* 53, 6 (Dec. 2021), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

[36] Jieun Lee, Genya Abe, Kenji Sato, Makoto Itoh, University of Tsukuba 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan, and Japanese Automobile Research Institute 2530 Karima, Tsukuba, Ibaraki 305-0822, Japan. 2020. Effects of Demographic Characteristics on Trust in Driving Automation. *Journal of Robotics and Mechatronics* 32, 3 (June 2020), 605–612. https://doi.org/10.20965/jrm.2020.p0605

[37] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (Oct. 1992), 1243–1270. https://doi.org/10.1080/00140139208967392

[38] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[39] Jason Y. Lee, Phillip Mucksavage, David C. Kerbl, Victor B. Huynh, Mohamed Etafy, and Elspeth M. McDougall. 2012. Validation Study of a Virtual Reality Robotic Simulator—Role as an Assessment Tool? *The Journal of Urology* 187, 3 (March 2012), 998–1002. https://doi.org/10.1016/j.juro.2011.10.160

[40] Mengyao Li, Isabel M Erickson, Ernest V Cross, and John D Lee. 2023. It's Not Only What You Say, But Also How You Say It: Machine Learning Approach to Estimate Trust from Conversation. *Human Factors* (April 2023), 00187208231166624. https://doi.org/10.1177/00187208231166624

[41] Mengyao Li, Amudha V. Kamaraj, and John D. Lee. 2023. Modeling Trust Dimensions and Dynamics in Human-Agent Conversation: A Trajectory Epistemic Network Analysis Approach. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–12. https://doi.org/10.1080/10447318.2023.2201555

[42] Oliver Liu, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. 2017. Understanding human-robot interaction in virtual reality. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 751–757. https://doi.org/10.1109/ROMAN.2017.8172387 ISSN: 1944-9437.

[43] Yugang Liu and Goldie Nejat. 2013. Robotic Urban Search and Rescue: A Survey from the Control Perspective. *Journal of Intelligent & Robotic Systems* 72, 2 (Nov. 2013), 147–165. https://doi.org/10.1007/s10846-013-9822-x

[44] Martina Mara, Kathrin Meyer, Michael Heiml, Horst Pichler, Roland Haring, Brigitte Krenn, Stephanie Gross, Bernhard Reiterer, and Thomas Layer-Wagner. 2021. CoBot Studio VR: A Virtual Reality Game Environment for Transdisciplinary Research on Interpretability and Trust in Human-Robot Collaboration.

[45] Robert McGrath and Gregory Meyer. 2006. When Effect Sizes Disagree: The Case of r and d. *Psychological methods* 11 (Dec. 2006), 386–401. https://doi.org/10.1037/1082-989X.11.4.386

[46] Andrea Moglia, Vincenzo Ferrari, Luca Morelli, Mauro Ferrari, Franco Mosca, and Alfred Cuschieri. 2016. A Systematic Review of Virtual Reality Simulators for Robot-assisted Surgery. *European Urology* 69, 6 (June 2016), 1065–1080. https://doi.org/10.1016/j.eururo.2015.09.021

[47] Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4, 2 (2013), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

[48] Brian Oakley, Mustapha Mouloua, and Peter Hancock. 2003. Effects of Automation Reliability on Human Monitoring Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47, 1 (Oct. 2003), 188–190. https://doi.org/10.1177/154193120304700139

[49] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2 (June 1997), 230–253. https://doi.org/10.1518/001872097778543886

[50] Luis Pérez, Eduardo Diez, Rubén Usamentiaga, and Daniel F. García. 2019. Industrial robot control and operator training using virtual reality interfaces. *Computers in Industry* 109 (Aug. 2019), 114–120. https://doi.org/10.1016/j.compind.2019.05.001

[51] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. 2017. Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (Aug. 2017), 425–436. https://doi.org/10.1109/THMS.2017.2648849

[52] Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464.

[53] Walter W Stroup. 2012. *Generalized linear mixed models: modern concepts, methods and applications.* CRC press.

[54] Ningyuan Sun and Jean Botev. 2021. Intelligent autonomous agents and trust in virtual reality. *Computers in Human Behavior Reports* 4 (Aug. 2021), 100146. https://doi.org/10.1016/j.chbr.2021.100146

[55] Daniel Szafir. 2019. Mediating Human-Robot Interactions with Virtual, Augmented, and Mixed Reality. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, Jessie Y.C. Chen and Gino Fragomeni (Eds.). Springer International Publishing, Cham, 124–149. https://doi.org/10.1007/978-3-030-21565-1_9

[56] Xinxing Tang and Hironao Yamada. 2011. Tele-operation Construction Robot Control System with Virtual Reality Technology. *Procedia Engineering* 15 (Jan. 2011), 1071–1076. https://doi.org/10.1016/j.proeng.2011.08.198

[57] Nathan Tenhundfeld, Mustafa Demir, and Ewart de Visser. 2022. Assessment of Trust in Automation in the "Real World": Requirements for New Trust in Automation Measurement Techniques for Use by Practitioners. *Journal of Cognitive Engineering and Decision Making* 16, 2 (June 2022), 101–118. https://doi.org/10.1177/15553434221096261

[58] Fang-Wu Tung. 2011. Influence of Gender and Age on the Attitudes of Children towards Humanoid Robots. In *Human-Computer Interaction. Users and Applications*, Julie A. Jacko (Ed.). Springer, Berlin, Heidelberg, 637–646. https://doi.org/10.1007/978-3-642-21619-0_76

[59] Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, and Pim Haselager. 2014. Do Robot Performance and Behavioral Style affect Human Trust? *International Journal of Social Robotics* 6, 4 (Nov. 2014), 519–531. https://doi.org/10.1007/s12369-014-0231-5

[60] Wietse van Dijk, Saskia J. Baltrusch, Ezra Dessers, and Michiel P. de Looze. 2023. The effect of human autonomy and robot work pace on perceived workload in human-robot collaborative assembly work. *Frontiers in Robotics and AI* 10 (Nov. 2023). https://doi.org/10.3389/frobt.2023.1244656

[61] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–39. https://doi.org/10.1145/3476068

[62] Robert Whelan. 2008. Effective Analysis of Reaction Time Data. *The Psychological Record* 58, 3 (July 2008), 475–482. https://doi.org/10.1007/BF03395630

[63] Anqi Xu and Gregory Dudek. 2016. Towards Modeling Real-Time Trust in Asymmetric Human–Robot Collaborations. In *Robotics Research: The 16th International Symposium ISRR*, Masayuki Inaba and Peter Corke (Eds.). Springer International Publishing, 113–129. https://doi.org/10.1007/978-3-319-28872-7_7

[64] Xiliu Yang, Aimée Sousa Calepso, Felix Amtsberg, Achim Menges, and Michael Sedlmair. 2023. Usability Evaluation of an Augmented Reality System for Collaborative Fabrication between Multiple Humans and Industrial Robots. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction (SUI '23)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3607822.3614528

[65] X. Jessie Yang, Christopher Schemanske, and Christine Searle. 2023. Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation. *Human Factors* 65, 5 (2023), 862–878. https://doi.org/10.1177/00187208211034716