Using Trust to Determine User Decision Making & Task Outcome During a Human-Agent Collaborative Task

Sarita Herse*
University of Technology Sydney
Sydney, Australia
Sarita.Herse@student.uts.edu.au

Jonathan Vitale*
University of Technology Sydney
Sydney, Australia
Jonathan.Vitale@uts.edu.au

Benjamin Johnston
University of Technology Sydney
Sydney, Australia
Benjamin.Johnston@uts.edu.au

Mary-Anne Williams
University of New South Wales
Sydney, Australia
Mary-Anne.Williams@unsw.edu.au

ABSTRACT

Optimal performance of collaborative tasks requires consideration of the interactions between socially intelligent agents, such as social robots, and their human counterparts. The functionality and success of these systems lie in their ability to establish and maintain user trust; with too much or too little trust leading to over-reliance and under-utilisation, respectively. This problem highlights the need for an appropriate trust calibration methodology, with the work in this paper focusing on the first step: investigating user trust as a behavioural prior. Two pilot studies (Study 1 and 2) are presented, the results of which inform the design of Study 3. Study 3 investigates whether trust can determine user decision making and task outcome during a human-agent collaborative task. Results demonstrate that trust can be behaviourally assessed in this context using an adapted version of the Trust Game. Further, an initial behavioural measure of trust can significantly predict task outcome. Finally, assistance type and task difficulty interact to impact user performance. Notably, participants were able to improve their performance on the hard task when paired with correct assistance, with this improvement comparable to performance on the easy task with no assistance. Future work will focus on investigating factors that influence user trust during human-agent collaborative tasks and providing a domain-independent model of trust calibration.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *User studies*; HCI theory, concepts and models.

KEYWORDS

trust; decision making; signal detection theory; recommender system; human-agent collaboration; socially intelligent agent

*Both authors contributed equally to this research.

HRI '21, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8289-2/21/03…$15.00
https://doi.org/10.1145/3434073.3444673

ACM Reference Format:
Sarita Herse, Jonathan Vitale, Benjamin Johnston, and Mary-Anne Williams. 2021. Using Trust to Determine User Decision Making & Task Outcome During a Human-Agent Collaborative Task. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3434073.3444673

1 INTRODUCTION

A semi-autonomous Tesla vehicle crashed in March 2018. The car veered into a roadside barrier and caught fire, resulting in the death of the driver. Investigation found that the Autopilot function was engaged at the time of the accident and that both visual and audible safety warnings were displayed to the driver before the accident occurred [17]. In this scenario, the driver trusted the collaborative system too much - over-relying on the vehicle's ability to navigate varied road conditions using the Autopilot system. Further, the driver under-relied on the vehicle's inbuilt warning system - lacking trust to acknowledge the warning signals as a legitimate call to action. To avoid events like this, it is necessary to design socially intelligent systems that build appropriate levels of user trust.

Now, imagine purchasing a socially intelligent semi-autonomous car. Before your first drive, the car prompts you to complete a brief task to get to know you. This task offers the Autopilot system a prior - an idea of your current level of trust toward the system. The system then employs particular factors (e.g., presenting you with information in a certain way) to nudge your trust towards an optimum level, therefore minimising accidents caused by over-reliance and under-utilisation. The system continues to learn about your decision making and behaviours over time, adjusting the factors accordingly to keep your trust within an optimal range. This system identifies your initial prior of trust and employs a model of trust calibration to ensure optimal collaborative performance while driving.

Trust calibration is timely for human-robot interaction considering the adoption of robots within collaborative team environments in industries like education [42, 55, 71] and healthcare [30, 35, 45]. Notably, trust calibration models have been employed in human-robot interaction contexts to improve team performance [12, 13, 67, 68], as well as agent-agent [58] and human-agent interaction [2, 52]. However, the addition of a prior to these models is novel and significant given the importance of initial user trust,

especially in the case of new technologies [44]. Taken together, if trust can determine user decision making and task accuracy, then a methodology is required that:

(1) Establishes a relationship between user trust and collaborative task performance.
(2) Identifies factors that influence user trust during collaborative human-agent interaction.
(3) Provides a domain-independent model of trust calibration.

While investigation of steps 2 and 3 is beyond the scope of this paper, the current work will provide evidence of the first step in this methodology. This is done by addressing the following aims:

(I) Provide a domain-independent methodology that can be used to behaviourally assess trust.
(II) Demonstrate that user trust is crucial during a human-agent collaborative task, with different levels of user trust leading to differences in task outcome.

Secondary aims are as follows:

(III) Investigate the impact of correct and incorrect assistance on task outcome.
(IV) Investigate the impact of task difficulty on task outcome.

2 BACKGROUND

Trust is imperative in establishing successful relationships between users and robots, with this being especially true for mixed initiative teams and relationships requiring collaboration [11, 21]. However, the development and maintenance of trust between users and intelligent agents is both complicated and nuanced. The functionality and success of collaborative systems lies in their ability to forge trusting relationships with users [25]. In this context, trust refers to an attitude which includes the belief that an agent will perform as expected and can be relied on to achieve its design goal [21, 40, 46]. Designing for trust in the context of collaboration requires the consideration of many factors central to the user, machine, and task [56]. Failure to account for these factors has the potential to harm user perception, adoption, and the ultimate success of these technologies.

2.1 Calibration

When a user trusts a system too much it can lead to misuse and over-reliance on a system. During collaboration, this occurs when users trust the system more than their own or their teammates' capabilities and is more likely to occur when the system is highly reliable [69]. Dickie and Boyle [18] highlight this complacency with owners of semi-automatic cars, reporting that large groups of users had incorrect knowledge about the boundaries of the adaptive cruise control system of their vehicle. This resulted in drivers relying on the system even in situations when it could not work, such as with tight turns and in stop-and-go traffic. Conversely, too little trust in a system can lead to users ignoring or turning off features inappropriately, under-utilisation, or even complete lack of use [51]. This distrust can lead to an operator over-monitoring the system, potentially making more errors due to neglecting supervision of other systems [50]. As such, it is crucial to employ specific practices or factors to ensure the appropriate calibration of user trust to a system [39].

Trust can be properly calibrated given an accurate understanding of the outputs and likely failures of a robotic system, as well as the factors that may lead to those outcomes [50]. A higher level of user trust of an intelligent system will lead to higher compliance with the system's recommendations [33]. However, Akash et al. [2] confirm that it is not always beneficial to increase trust, stating that we should instead prioritise the optimisation of context-specific performance when designing intelligent decision-aid systems that influence user trust and behaviour.

Thus, adjustment of machine features to calibrate user trust should not be the focus. Ososky and colleagues argue that it is not the features and behaviours of the system that are mediating trust - rather, it's the user's perception of these features and behaviours [50]. Human perception is sensitive and dynamic compared to discrete and stable system features. As such, no matter how well you initially design for user trust of a socially intelligent system, interaction over time and user-specific events will continue to alter user perception of the system over time [56].

Take the example of over-reliance on the Autopilot feature of a semi-autonomous vehicle. An Autopilot system designed to calibrate user trust would detect over-reliance given the behaviour of the driver, such as late recognition of warning signals. It would then engage in actions to ensure faster recognition of these signals and calibrate this factor over time to ensure the safety of the user. Since user trust is always shifting, it is far more efficient to develop a calibration model that adjusts the warning signals to the driver's perceived trust level compared to a total overhaul and re-design of the vehicle's warning signals - which may work perfectly for a driver with a different level of trust of the system. Taken together, calibration of user perception via trust is plausibly the most efficient, dynamic, and long term solution to optimising performance outcomes on collaborative tasks.

2.2 Assessment of Trust

In order to calibrate user trust of a system, it is imperative that we have the ability to appropriately assess trust across varied environments and domains. It is not practical to collect human self-reported behaviour for use with real-time feedback algorithms [2]. An alternative way of approaching this problem is by utilising indirect measures [1]. In comparison to direct measures such as self report instruments, an indirect measure would assess trust using a disguised method of unobtrusive behavioural observation [4], such as implicit inference of user trust via compliance [21, 68]. Indirect measures of attitudes have been demonstrated as more effective than direct measures [23] and are useful when responses collected from direct measures may be biased [4].

Behavioural economics and psychology utilise experimental games as a way of providing an indirect, quantifiable measure of the underlying state of trust [19]. These games are designed to represent certain critical features of real-life situations [19], where the outcomes that occur are determined by the choices of the players [15]. Simplified investment games are popular and require two parties, a Trustor and Trustee. Decisions in the game are discrete choices [43]: the Trustor has a choice between trust and mistrust. There is a potential benefit for choosing trust, but the decision to trust requires the acceptance of risk as the outcome of trust depends

solely upon the behaviour of the Trustee. This decision should be perceived by users as having some risk. Notably, the likelihood of trust decreases as the risk increases, with people less likely to trust when there is little to gain and a lot to lose [19]. As such, these games offer a sound approach for indirectly assessing trust and are malleable enough to be altered and applied to different scenarios.

2.2.1 The Trust Game. The Trust Game [9] has been successfully employed in human-agent research [49] and human-robot interaction studies [28]. Further, the Trust Game has been adapted to match particular contexts and experiment constraints. However, it is important to note that methodological differences may impact behaviour during the game [36]. Taken together, an adapted version of the Trust Game could be used to assess the trust prior in the context of trust calibration.

The standard procedure for the Trust Game, which only has one round, is as follows: Both Trustor and Trustee are given X amount of dollars by the experimenter and are aware they are each given an equal amount. The following instructions are then relayed to the Trustor and Trustee together: Trustor, you now have the opportunity to give the Trustee some amount of money (Y). This can be anywhere from nothing to the full amount you currently have ($0 \leq Y \leq X$). I will triple this amount before giving it to the Trustee ($3Y$). So the Trustee will now have their original amount plus what you have given them which has been tripled ($X + 3Y$). The Trustee will then have the opportunity to decide how much of their total amount they would like to give back to the Trustor (Z), which can be anywhere from nothing to the full amount they currently have ($0 \leq Z \leq X + 3Y$).

2.3 Recommender Systems & Signal Detection Theory

Signal detection theory (SDT) can be applied to binary decision-making (e.g. yes-no tasks) as long as participant responses can be compared to the presence or absence of the target stimulus. Thus, tasks involving recommender systems can be interpreted using SDT [59, 70]. Recommender systems offer a simple real world application that is easy to understand and harbours risk through decision making [33], with previous work utilising this applied use case to investigate decision making [14, 16, 26, 48] and trust [8, 54, 61]. In human-human interaction, people are more likely to follow the recommendations of the person they trust [22]. This effect of trust generalises to intelligent agents, with research demonstrating that higher trust of a social robot recommender system results in more recommendations being followed [54]. In this context, trust is viewed as a tool that recommender systems can use to increase their chance of convincing a user to select the recommended option [54].

SDT can be used in this context to provide a unitless measure of sensitivity in decision making. Decisions are made against a background of uncertainty, where the participants' aim is to tease out the decision *signal* from background *noise* (Figure 1). The basic premise behind SDT is that both signal and noise are represented probabilistically within each participant, and that the extent to which those representations overlap can be estimated based on the participants' responses and whether or not the signal is present. The participant bases their decision relative to their own internal criterion β, where a signal will be reported present when the internal signal is stronger than β and absent when the internal signal is weaker than β [3]. Importantly, for every individual, an optimal operating point $\beta*$ exists where performance on a discriminatory task is maximised [59].

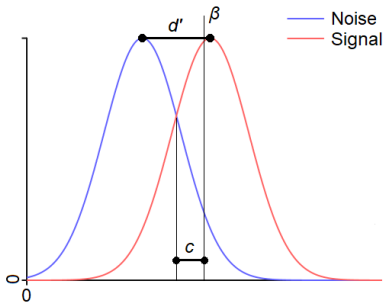

Figure 1: SDT demonstrating signal and noise distributions as well as sensitivity d', response bias c, and the criterion β.

Sensitivity d' measures the distance between the signal and noise means in standard deviation units [62]. In other words, d' is the degree of overlap between the signal and noise distribution. A small d' would reflect substantial overlap between the signal and noise distributions, meaning the discrimination task would be harder given there is a higher chance that noise would distract from the signal. Further, d' is independent of where β is placed, thus d' is a measure of performance that is independent of subject bias [3].

Response bias is the general tendency to respond *yes* or *no* as determined by the location of the criterion. It is estimated from the difference between the placement of β by the participant and the placement of β by an unbiased participant [3]. Sheridan [59] highlights the link between the criterion β and the human agent's level of trust, suggesting that the criterion β indicates how the subject calibrates trust during interactions with an automated system over a set of repeated trials.

While β has historically been the most popular measure of response bias [62], c offers an analogue of β, unaffected by changes in d'. As such, c will be used in this paper to assess response bias. Response bias, c, is negative when the participant is more likely to report the signal is present (i.e. when β is placed further left along the distribution - a liberal criterion). The absolute value of c provides an indication of the strength of the subject's bias. Taken together, c can be modelled as a function, say g, of the individual's trust level τ, namely $c = g(\tau, \theta)$, with θ being a set of model parameters. By calibrating user trust of a system, it is possible to adapt c to influence task outcomes.

3 PILOT STUDIES

The goal of the research herein is to provide evidence for the inclusion of an initial behavioural measure of trust within trust calibration methodologies. To do this, we first need an appropriate behavioural measure of trust that could be applied across different platforms (Study 1) and a task that could be completed by participants online (Study 2). The outcomes of Study 1 and 2 were

necessary to design Study 3, an experiment that allowed investigation into the relationship between an initial behavioural measure of trust, user decision making, and task outcome. Study 1 and 2 received ethics approval from the University of Technology Sydney Human Research Ethics Committee.

3.1 Study 1: Adapted Trust Game

An exploratory pilot study was conducted in order to identify whether an adapted Trust Game could be used as an appropriate behavioural measure of trust across different platforms. Here, domain-independence refers to methods that are independent of application, context, and that could be applied in a wide range of scenarios. The embodiment of a system is one factor necessary to consider when designing domain-independent methods, with previous work demonstrating the impact of agent embodiment on user behaviour [32, 33, 64, 65].

3.1.1 Design. A controlled, between-subjects experiment was conducted in The Innovation and Enterprise Research Laboratory ("The Magic Lab") of the University of Technology Sydney. The independent variable was task partner, with three levels: social robot (Softbank Pepper Robot), disembodied virtual assistant, and human (confederate). The dependent variable was participant trust behaviour towards the task partner, operationalised via score on an adapted Trust Game.

3.1.2 Participants. A social media roll out was used to recruit 62 participants (*Male* = 33, *Female* = 28, *Rather not say* = 1; M_{age} 26.45 years, SD_{age} = 8.20). Participants were informed that the experiment would take 30 minutes to complete and that participation would place them in the draw to win one of five AUD$50 grocery vouchers. The inclusion criteria required participants to be proficient in English and at least 18 years old. All participants provided informed consent in accordance with human research ethical standards prior to experimentation.

3.1.3 Materials. Instead of money, the Trust Game was adjusted with participants playing to increase their chance of winning one of the five grocery vouchers. This adjustment was made in order to create a sense of risk to the participant that suited the context of the experiment.

3.1.4 Procedure. Participants were given instructions on the adapted Trust Game and were verbally tested by the experimenter to confirm understanding of the task. Participants were taken to a private testing room by the experimenter and briefly introduced to their task partner, with partner type randomised across participants. The experimenter left the room and participants completed the adapted Trust Game with the task partner. Finally, the experimenter debriefed participants.

3.1.5 Results. A one-way ANOVA on the effect of trust behaviour on partner type yielded no evidence to suggest a significant difference between the human, robot, and virtual assistant conditions $F(2,59) = 0.26$, $p = .77$. This finding is surprising as it counters established work that suggests there are differences between virtual agents and social robots [41, 57]. Considering the exploratory nature and aims of this work, we place little weight on this finding. Rather, this study is of interest as it provided an in-person experimental use case for an adapted Trust Game across different agent platforms. Further investigation into the use of an adapted Trust Game as a behavioural measure of trust is discussed in the results of Study 3 (section 5.1). This is done by checking that trust behaviour is a predictor of response bias *c*, an analogue to trust [59].

3.2 Study 2: Radiology Diagnostics Task

An exploratory pilot study was conducted to investigate whether detection of viral pneumonia within X-ray images could be applied online as a real world application of a yes-no task. In psychophysics, a yes-no task is a signal detection task where participants undergo a series of trials in which they must judge the presence (yes) or absence (no) of a signal [6], which in this case is viral pneumonia.

3.2.1 Participants. Amazon Mechanical Turk was used to recruit 48 participants (*Male* = 30, *Female* = 17, *Non-Binary* = 1; M_{age} = 39.83, SD_{age} = 13.26; *Age distribution:* 18-24 = 4, 25-34 = 18, 35-44 = 8, 45-54 = 12, 55-64 = 3, 65-74 = 3, 75+ = 0) from Australia, Canada, the United Kingdom, and the United States. Participants were informed the experiment would take 15 minutes and were reimbursed USD$1 for their time. The inclusion criteria required participants to be proficient in English and at least 18 years old. All participants provided informed consent in accordance with human research ethical standards prior to experimentation.

3.2.2 Materials. A radiology diagnostics task was used given that SDT can be applied to diagnostic accuracy [47]. X-ray images of a pair of lungs with or without viral pneumonia were sourced from an Open Source data set [37, 38]. These images were pre-classified as *viral pneumonia present* and *viral pneumonia absent*.

3.2.3 Procedure. Participants were presented 64 X-ray images (32 with viral pneumonia present and 32 with viral pneumonia absent) for either 2, 4, 6, or 8 seconds, where they were able to move on after 2 seconds. Then, on a new screen, participants answered Yes/No to whether they thought the X-ray image showed signs of viral pneumonia.

3.2.4 Results. The X-ray images were further classified to determine their difficulty. This was done using a T-test to compare overall participant accuracy against chance rate. Images were defined as *significant correct* ($p < .05$, positive mean difference), *significant incorrect* ($p < .05$, negative mean difference), and *non-significant chance* ($p > .05$). Four seconds was selected for the timing of stimulus presentation in order to maximise the number of trials presented while not impairing performance.

4 STUDY 3: INVESTIGATING USER TRUST

The aim of Study 3 was to assess the use of trust as a prior for modelling trust calibration. The relationship between user trust and collaborative task performance was investigated. A radiology diagnostics task was developed from the outcomes of Study 2. Participants first completed this task alone. They were then introduced to the virtual assistant (Trustee) and asked to play against them as the Trustor in an adapted version of the Trust Game, as developed from the outcomes of Study 1. Finally, participants completed the same radiology diagnostics task with the help of the virtual

Figure 2: Example of virtual assistant banner.

assistant. Study 3 received ethics approval from the University of Technology Sydney Human Research Ethics Committee.

4.1 Design

A 2x3 within subjects design was executed. The first independent variable was agent assistance with task. There were three levels: no assistance (control), agent giving correct assistance, and agent giving incorrect assistance. The second independent variable was difficulty of the radiology diagnostics task. There were two levels: easy and hard, as determined by Study 2. The dependent variables were sensitivity and response bias, operationalised using SDT with d' and c, respectively [62].

4.2 Participants

An a priori power analysis was conducted using G*Power3 [20] to test the difference between assistance group means taken from pilot data. A two-tailed test with $\alpha = .05$ was run for the smallest effect size ($d = .41$). Results showed that a total sample of 64 participants were required to achieve a power of .90.

Amazon Mechanical Turk was used to recruit 68 participants from Australia, Canada, the United Kingdom, and the United States. Participants were informed the experiment would take 15-20 minutes and were reimbursed USD$1.50 for their time. The inclusion criteria required participants to be proficient in English and at least 18 years old. All participants provided informed consent in accordance with human research ethical standards prior to experimentation. The data of 2 participants were excluded from analysis as they did not complete the task seriously; therefore $N = 66$ (*Male* = 41, *Female* = 25, *Non-Binary* = 0; $M_{age} = 40.33$, $SD_{age} = 11.82$, *Age distribution*: 18-24 = 3, 25-34 = 22, 35-44 = 20, 45-54 = 10, 55-64 = 10, 65-74 = 1, 75+ = 0).

4.3 Materials

4.3.1 Virtual Assistant. Assisto was used to represent the virtual assistant recommender system (see Figure 2). Participants were instructed that Assisto was *An AI system that has been specifically designed to help complete this task with you* and that it would provide a systems report with a recommended decision when working collaboratively with the user on the radiology diagnostics task.

4.3.2 Adapted Trust Game. In order to indirectly measure trust, participants completed a brief trading task with Assisto. The task is an adjusted version of the Trust Game. Adjustment was made in order to create a sense of risk to the participant that suited the context of the experiment.

The task was adjusted to make participants believe they were playing for a chance to reduce the number of trials they had to complete in the final block with Assisto. Participants were informed the final block had 60 trials and were then given the following instructions: *You begin the trading task with 10 points. Each point deducts 1 image from the total number of images you will have to judge with Assisto in the next part of the study. i.e. if you have 10 points, you will only need to complete 50 images with Assisto, not the original 60. You now have the opportunity to trade your points with Assisto. Assisto is also currently on 10 points. The number of points you give to Assisto will be tripled. For example, if you give 0 points, Assisto receives 0 points - therefore, Assisto has a total of 10 points; if you give 5 points, Assisto receives 15 points - therefore, Assisto has a total of 25 points; if you give 10 points, Assisto receives 30 points - therefore, Assisto has a total of 40 points. Assisto then has the opportunity to return points to you. It can give you anywhere between 0 points and its total amount (which depends on how many points you initially gave). This will determine your total point score and the final number of X-rays you will view with Assisto in the next task.*

Before completing the task, three multiple choice questions were used to test participants on the rules of the adapted Trust Game to confirm understanding. Participants were informed that each incorrect answer would incur a 10 second "wait" time penalty before moving on. This penalty was included to motivate participants to read the instructions and complete the task seriously.

After making a decision during the task participants were told: *Assisto has decided how many points to return to you. The final number of X-rays you are to complete with Assisto has been adjusted accordingly.* They were not told how many points were returned or the final amount of X-rays they would have to complete to avoid positively or negatively biasing participants to Assisto. All participants were made to complete 40 trials in the second block.

4.3.3 Radiology Diagnostics Task. 20 X-ray images were used for the test, with 50% of trials using *significant correct* images (easy task difficulty) and 50% using *non-significant chance* images (hard task difficulty) as defined by Study 2 (3.2.4). This breakdown of task difficulty was selected to ensure participants' ability to complete the task above chance level without assistance. For each difficulty condition, half of the trials had the virus present and half had the virus absent.

For the control condition, the 20 images were randomly presented. Participants were prompted with a screen with instructions: *Press the button below when you are ready to view the X-ray.* Participants were presented the X-ray image for 4 seconds and then automatically displayed a new screen where they answered Yes/No to the following question *Did the X-ray show signs of Viral Pneumonia?* For the assistance conditions, the same 20 images were presented twice more - once for each assistance condition. These trials were presented together in randomised order in one 40 trial block, simulating a collaborative virtual assistant with a 50% accuracy rate. A banner with an image of the virtual assistant (similar to Figure 2) was presented above the Yes/No question. Depending on the condition of the trial, the banner either displayed *System Report: YES - Virus Present* or *System Report: NO - Virus Absent.* There was no penalty for incorrect response during the radiology diagnostics task.

4.4 Procedure

The experiment contained six main stages: (1) Demographics questions. (2) Information on the radiology diagnostics task with example images and 8 practice trials. (3) Radiology diagnostics task with no assistance. 20 trials were presented. (4) Information on the virtual assistant. (5) Adapted trust game with virtual assistant. (6) Radiology diagnostics task with virtual assistant. 20 correct and 20 incorrect assistance trials, randomly presented in one 40 trial block. Participants were blind to the assistance condition.

5 RESULTS

All analyses were conducted using SPSS Statistical Package [34]. SDT was employed to interpret the data. Data from the radiology diagnostics task were organised into four cells (*Hits*, *Misses*, *False Alarms* and *Correct Rejects*) for each participant, for every condition. To avoid any biasing effect of extreme proportions, 0.5 was added to each cell [10, 29]. Sensitivity d' and response bias c were then calculated using methods outlined by Stanislaw & Todorov [62].

5.1 Prediction

Linear regressions were run to understand the effect of trust score on response bias c for both the correct and incorrect assistance conditions. To assess linearity, scatterplots of c for correct and incorrect assistance against trust score with superimposed regression line were plotted. Visual inspection indicated a linear relationship between the variables. There was independence of residuals, as assessed by a Durbin-Watson statistic of 1.55 and 1.90 for the correct and incorrect assistance conditions, respectively. There was homoscedasticity for both conditions, as assessed by visual inspection of plots of standardized residuals versus standardized predicted values. Residuals were normally distributed as assessed by visual inspection of a normal probability plot and there were no outliers.

The prediction equation for c with correct assistance = -0.97 + 0.09 × (trust score). Trust score statistically significantly predicted c with correct assistance, $F(1, 64) = 7.56$, $p < .01$, accounting for 10.6% of the variation in c with correct assistance with adjusted $R^2 = 9.2\%$. The prediction equation was: c for incorrect assistance = -0.72 + 0.07 × (trust score). Trust score statistically significantly predicted c for incorrect assistance, $F(1, 64) = 4.00$, $p = .05$, accounting for 5.9% of the variation in c for incorrect assistance with adjusted R^2 = 4.4% . Both linear regressions are depicted in Figure 3.

5.2 Group Differences

A two-way repeated measures ANOVA was run to assess the effect of assistance and task difficulty on sensitivity d'. No outliers were found from examination of the studentized residuals for values greater than ±3. Sensitivity was not normally distributed ($p < .05$) as assessed by Shapiro-Wilk's test of normality on the studentized residuals; however, no adjustments were made [24, 63]. Mauchly's test of sphericity indicated that the assumption of sphericity was met for the two-way interaction, $\chi^2(2) = 3.26$, $p = .20$.

There was a statistically significant two-way interaction between assistance type and task difficulty for sensitivity d', $F(2, 130) = 4.28$, $p = .02$. Therefore, simple main effects were run (summarised in Figure 4). Mean sensitivity d' was significantly higher for the easy task ($M = 0.93$, $SD = 1.12$) compared to the hard task ($M = 0.00$, $SD = 0.72$) for the control condition, $F(1, 65) = 28.87$, $p < .01$, a mean difference of 0.94, 95% CI [0.59 to 1.29]. Mean sensitivity was significantly higher for the easy task ($M = 1.46$, $SD = 1.27$) compared to the hard task ($M = 0.94$, $SD = 1.20$) when participants were provided with correct assistance, $F(1, 65) = 13.46$, $p < .01$, a mean difference of 0.52, 95% CI [0.24 to 0.80]. Mean sensitivity was significantly higher for the easy task (M = -0.43, $SD = 1.34$) compared to the hard task (M = -0.93, $SD = 1.33$) when participants were provided with incorrect assistance, $F(1, 65) = 14.31$, $p < .01$, a mean difference of 0.51, 95% CI [0.24 to 0.77].

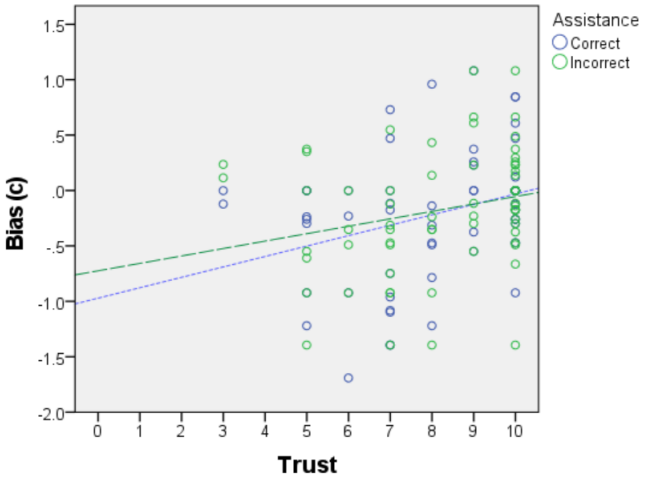

Figure 3: Scatterplot of response bias (c) against adapted Trust Game score for the correct and incorrect assistance conditions with regression lines superimposed.

The assumption of sphericity was violated for both simple main effects across task difficulty ($p < .05$), thus the Greenhouse-Geisser adjustment was used [63]. Mean sensitivity was significantly different across assistance conditions for the easy task, $F(1.44, 93.66) = 43.72$, $p < .01$, with participants able to complete the task with no assistance ($M = 0.93$, $SD = 1.12$) and increasing their sensitivity when teamed with correct assistance ($M = 1.46$, $SD = 1.27$). However, there was a large decrease in performance to below chance level when paired with incorrect assistance (M = -0.43, $SD = 1.34$). A similar effect was also seen for the hard task, with a statistically significant difference of mean sensitivity found across assistance conditions, $F(1.34, 87.05) = 36.23$, $p < .01$. While participants did not demonstrate an aptitude for the hard task without assistance ($M = 0.00$, $SD = 0.72$), they were able to perform significantly better when paired with correct assistance ($M = 0.94$, $SD = 1.20$). However, performance was also lowest when participants were paired with incorrect assistance (M = -0.93, $SD = 1.33$).

6 DISCUSSION

The authors' overall aim is to develop a calibration methodology that leverages user trust to optimise performance outcomes during human-agent collaboration. This work investigated the first of three research questions with the introduction of a trust prior: *Is there a relationship between user trust and collaborative task performance?*

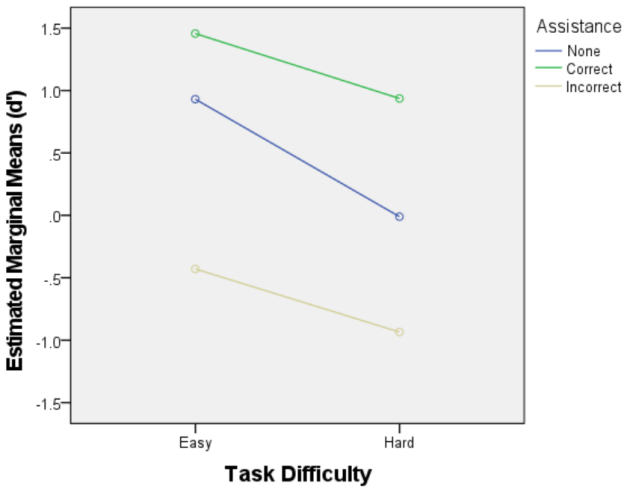

Figure 4: Plot of the estimated marginal means of Assistance Type and Task Difficulty for sensitivity (d').

User trust can be behaviourally assessed. Use of an adapted Trust Game as a behavioural measure of trust was successful in this context. This was evidenced through the correlational relationship between trust behaviour and response bias c. In this context, c was interpreted as an analogue to the criterion β, where β indicates how a subject calibrates trust during interactions with an automated system over a set of repeated trials [59]. While this is promising evidence, more work is needed to confirm that behavioural measures, such as an adapted Trust Game, can be used as a measure for trust that is both context and domain-independent. Future work should employ variants of contextually-relevant investment games alongside different socially intelligent agents like social robots.

User trust is a predictor of task outcome for human-agent collaborative tasks. Correlational evidence demonstrated that the initial behavioural measure of user trust significantly predicted response bias towards a collaborative assistant, regardless of whether the system provided correct or incorrect recommendations. Across both conditions, participants with low levels of initial trust demonstrated more response bias, i.e. they were more liberal with reporting the signal as present. Additionally, participants with higher levels of initial trust saw their response bias tend to zero, suggesting that the more trust they had of the assistant, the less biased they were with the task.

This offers promising evidence for the employment of trust as a prior when modelling for trust calibration. If we are able to estimate initial user trust before engaging in human-agent interaction, we can calibrate trust more efficiently, reducing the risk of negative user perceptions and interactions. This has implications for social robots, highlighting the importance of considering user trust throughout the design and development stages, as well as for initial interactions. While an optimistic result, it is imperative that these findings are bolstered with causational evidence of a relationship between a trust prior and collaborative task outcome.

Assistance type and task difficulty impact the outcome of human-agent collaborative tasks. Participants were able to successfully complete the easy task, i.e. identify true presentations and absences of the signal, without any assistance. When paired with correct assistance, participants were able to further improve their performance on the easy task, achieving the highest sensitivity of all the conditions. This improvement was also seen for the difficult task, where participants' sensitivity without assistance was approximately zero - suggesting participants were only able to complete the task at chance rate (an expected outcome given the design of the hard task). However, when collaborating with an assistant giving correct recommendations, performance improved dramatically. Participants were able to recover sensitivity to a level comparable to completion of the easy task with no assistance.

Performance on the task suffered significantly when participants collaborated with an assistant giving incorrect recommendations. Participants were no longer able to successfully complete the task, regardless of task difficulty. This reinforces the idea that people blindly trust (over-rely) collaborative assistants, regardless of their own ability. Notably, performance was worse for the hard task compared to the easy task when participants were paired with incorrect assistance. This suggests that, when placed in a position of difficulty or uncertainty, participants are more willing to follow the direction of a collaborative agent - with this finding supporting previous work on human-robot trust [33].

6.1 Limitations

Low trust levels. The most significant limitation in this work is the lack of data available for the low levels of initial trust behaviour. This is an unfortunate consequence of working with data that employs user characteristics to assign participants to independent groups. It is unlikely that there are no individuals that would have lower levels of trust behaviour. While random selection of participants aims to provide a normalised distribution of data, this is not always the case. As such, it is likely that our selection of participants was not a true representation of the population. While an acknowledged limitation, issues with sampling bias are widespread across human studies. For example, the sampling bias present when running experiments on university students [27, 31]. Overcoming this limitation requires replication with varied samples to ensure generalisability and external validity.

Practice effects. Practice effects are any change or improvement that results from the practice or repetition of a task and are of particular concern with within-subjects designs [5]. In the current work, the lack of an additional between-subjects condition to counterbalance the presentation order of the assistant/no assistant conditions is a limitation as we are unable to rule out the possibility of practice effects during the radiology diagnostics task. However, similar to real-world collaborative tasks, no feedback is provided to participants at any point during the radiology diagnostics task. This means it would be difficult for participants to improve on the task without assistance, regardless of cumulative experience with the X-ray stimulus. Further, prior experience with the assistant would likely bias future decision making, potentially compromising the baseline assessment of task performance. Thus, while practice

effects cannot be ruled out, the authors believe the present work grants appropriate investigation of our research goals.

Incoherent behaviour of the assistant. Since participants were blind to the correct/incorrect assistance condition, the authors decided to present these conditions together in a randomised order across the 40 trials. This simulated participants collaborating on the radiology diagnostics task with an assistant who performed at chance level. Though identifying X-ray images is difficult, participants may have recognised the repeated X-ray images and become aware of the assistant's incoherent behaviour. However, the results demonstrate bias towards following behaviour and a relationship between this behaviour and initial trust. This suggests participants did not regard the agent's behaviour as incoherent or even trusted the agent enough to follow its recommendations regardless. Future work would benefit from investigating various accuracy levels and through the inclusion of post-test questions about the X-ray stimuli.

Non-expert sample. Although real radiology diagnostics tasks are typically completed by domain experts, the task in this study is merely a yes-no psychophysics task. However, since participants perceived this as a legitimate task, were blind to the correct/incorrect assistance conditions, and did not receive feedback on the task, it is possible that bias towards agent recommendation occurred due to participants feeling under-qualified. Further, the positive framing of the agent's introduction paired with user uncertainty on performance could impact user expectations, priming them to follow system recommendations.

This narrative is contrary to the relationship found between initial trust and user behaviour. We would expect participants to follow the recommendations if they felt under-qualified, regardless of their trust level towards the agent. However, this was not the case given that initial trust predicted decision making and task outcome. Future work should consider the inclusion of participant and agent performance information in-test as well as varying participant priming when the agent is first introduced.

Ecological validity. A common limitation of human-agent interaction experiments in simulated environments is low ecological validity [60]. Ecological validity is a particular form of external validity, referring to the extent that research findings generalise to settings typical of everyday life [7]. Research illustrates that behaviour towards artificial agents within simulated environments differs from real life decision making [53]. We suggest that this is due to an absence of perceived risk attached to participant decisions made in-test. We attempted to resolve this in the adapted Trust Game by creating the perception that there would be real world consequences to the in-test decisions made. Instead of money, participants were incentivised with the potential of reducing the number of trials in the final radiology diagnostics block. This allowed participants to believe they could save time by shortening the overall experiment. Risk was created as participants understood that engaging in trading during the task could also increase the number of trials they had to complete. By introducing perceived risk to the decision making process, we attempted to mediate the limitation of using a simulated environment and in turn strengthen the ecological validity of the study.

Cultural context. Attitudes towards machines can be subject to cultural influences and impact human-robot collaboration [66]. While participants in Study 1 were all from Australia, participants in Study 2 and 3 came from a handful of different countries. Although investigation into the impact of cultural context was beyond the scope of this work, it is important to flag its importance - especially when considering practical, real-world applications such as with autonomous vehicles. Future work should be sensitive to the impact of cultural context when collecting and interpreting participant data, with this being particularly noteworthy for online data collection methods like Amazon Mechanical Turk.

6.2 Future work

This work is the first step in the larger context of optimising human-agent collaborative task outcomes through trust calibration. The introduction of a user to a new assistive agent was simulated. No prior experience with the agent allowed for assessment on whether initial trust could predict decision making and task outcome. A relationship between trust and decision making was demonstrated, highlighting the possibility of inputting behavioural measures of trust within trust calibration models to refine initial behaviour.

While long term interactions are necessary for the dynamic adjustments seen in trust calibration, investigation of this was beyond the scope of this paper. Rather, future work will focus on investigating factors that influence user trust during collaborative human-agent interaction. Finally, a domain-independent model of trust calibration with a trust prior will be presented. This will be developed with reference to current work on trust calibration during human-agent collaborative tasks using partially observable Markov decision process models [2, 52].

7 CONCLUSION

Promising results have been provided to support the inclusion of a behavioural trust prior within trust calibration methodologies. User trust can be behaviourally assessed using an adapted version of the Trust Game. However, future work is needed to confirm this methodology across varied environments and domains, including with social robots. An initial behavioural measure of user trust can significantly predict task outcome during a human-agent collaborative task. While an exciting result, causational evidence is necessary to bolster this finding. Finally, assistance type and task difficulty interact to impact user performance and outcome during a human-agent collaborative task. Task performance improved significantly when the agent provided correct assistance. This improvement was greatest for the hard task, where participants were able to recover performance to a level comparable to completion of the easy task with no assistance. Further, performance on the task suffered significantly when the agent provided incorrect assistance. This was particularly so for the hard condition, suggesting that participants are more willing to follow the recommendation of a collaborative agent when placed in a position of difficulty or uncertainty.

ACKNOWLEDGMENTS

This research is supported by an Australian Government Research Training Program Scholarship. The authors thank Suwen Leong, Bethany Lu, & The Duc Vu for acting as the confederate in Study 1.

REFERENCES

[1] McLeod S. A. 2018. *Attitude measurement.* Retrieved June 6, 2019 from https://www.simplypsychology.org/attitude-measurement.html

[2] Kumar Akash, Tahira Reid, and Neera Jain. 2019. Improving Human-Machine Collaboration Through Transparency-based Feedback–Part II: Control Design and Synthesis. *IFAC-PapersOnLine* 51, 34 (2019), 322–328.

[3] Nicole D. Anderson. 2015. Teaching signal detection theory with pseudoscience. *Frontiers in psychology* 6 (2015), 762.

[4] Richard F. Antonak and Hanoch Livneh. 1995. Direct and indirect methods to measure attitudes toward persons with disabilities, with an exegesis of the error-choice test method. *Rehabilitation Psychology* 40, 1 (1995), 3.

[5] American Psychological Association. 2020. *APA Dictionary of Psychology: practice effect.* Retrieved January 2, 2021 from https://dictionary.apa.org/practice-effect

[6] American Psychological Association. 2020. *APA Dictionary of Psychology: yes-no task.* Retrieved January 2, 2021 from https://dictionary.apa.org/yes-no-task

[7] Roy F. Baumeister and Kathleen D. Vohs. 2007. *Encyclopedia of social psychology.* Vol. 1. Sage.

[8] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[9] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142.

[10] Glenn S. Brown and K. Geoffrey White. 2005. The optimal correction for estimating extreme discriminability. *Behavior research methods* 37, 3 (2005), 436–449.

[11] Hua Cai and Yingzi Lin. 2010. Tuning trust using cognitive cues for better human-machine collaboration. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 2437–2441.

[12] 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*. 307–315.

[13] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2020. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 2 (2020), 1–23.

[14] Jaewon Choi, Hong Joo Lee, and Yong Cheol Kim. 2011. The influence of social presence on customer intention to reuse online recommender systems: the roles of personalization and product type. *International Journal of Electronic Commerce* 16, 1 (2011), 129–154.

[15] Michael G. Collins, Ion Juvina, and Kevin A. Gluck. 2016. Cognitive Model of Trust Dynamics Predicts Human Behavior within and between Two Games of Strategic Interaction with Computerized Confederate Agents. *Frontiers in Psychology* 7 (2016), 49. https://doi.org/10.3389/fpsyg.2016.00049

[16] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. 2012. Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: An Empirical Study. *ACM Trans. Interact. Intell. Syst.* 2, 2, Article 11 (June 2012), 41 pages. https://doi.org/10.1145/2209310.2209314

[17] Steve Dent. 2017. *Tesla driver in fatal Autopilot crash ignored safety warnings.* Retrieved September 30, 2020 from https://www.engadget.com/2017/06/20/tesla-driver-in-fatal-autopilot-crash-ignored-safety-warnings/?guccounter=1

[18] David A. Dickie and Linda N. Boyle. 2009. Drivers' understanding of adaptive cruise control limitations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. SAGE Publications Sage CA: Los Angeles, CA, 1806–1810.

[19] Anthony M. Evans and Joachim I. Krueger. 2009. The psychology (and economics) of trust. *Social and Personality Psychology Compass* 3, 6 (2009), 1003–1017.

[20] 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 (2007), 175–191.

[21] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *2007 International Symposium on Collaborative Technologies and Systems*. IEEE, 106–114.

[22] David Gefen. 2000. E-commerce: the role of familiarity and trust. *Omega* 28, 6 (2000), 725–737.

[23] Edward L. Glaeser, David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter. 2000. Measuring trust. *The quarterly journal of economics* 115, 3 (2000), 811–846.

[24] Gene V. Glass, Percy D. Peckham, and James R. Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research* 42, 3 (1972), 237–288.

[25] Victoria Groom and Clifford Nass. 2007. Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies* 8, 3 (2007), 483–500.

[26] Jaap Ham, Raymond H. Cuijpers, and John-John Cabibihan. 2015. Combining Robotic Persuasive Strategies: The Persuasive Power of a Storytelling Robot that Uses Gazing and Gestures. *International Journal of Social Robotics* 7, 4 (01 Aug 2015), 479–487. https://doi.org/10.1007/s12369-015-0280-4

[27] Paul H.P. Hanel and Katia C. Vione. 2016. Do student samples provide an accurate estimate of the general public? *PloS one* 11, 12 (2016), e0168354.

[28] Kerstin S. Haring, Yoshio Matsumoto, and Katsumi Watanabe. 2013. How do people perceive and trust a lifelike robot. In *Proceedings of the world congress on engineering and computer science*, Vol. 1.

[29] Michael J. Hautus. 1995. Corrections for extreme proportions and their biasing effects on estimated values ofd'. *Behavior Research Methods, Instruments, & Computers* 27, 1 (1995), 46–51.

[30] Denise Hebesberger, Tobias Koertner, Christoph Gisinger, and Jürgen Pripfl. 2017. A long-term autonomous robot at a care hospital: A mixed methods study on social acceptance and experiences of staff and older adults. *International Journal of Social Robotics* 9, 3 (2017), 417–429.

[31] Patrick J. Henry. 2008. Student sampling as a theoretical problem. *Psychological Inquiry* 19, 2 (2008), 114–126.

[32] Sarita Herse, Jonathan Vitale, Daniel Ebrahimian, Meg Tonkin, Suman Ojha, Sidra Sidra, Benjamin Johnston, Sophie Phillips, Siva Leela Krishna Chand Gudi, Jesse Clark, et al. 2018. Bon appetit! robot persuasion for food recommendation. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 125–126.

[33] Sarita Herse, Jonathan Vitale, Meg Tonkin, Daniel Ebrahimian, Suman Ojha, Benjamin Johnston, William Judge, and Mary-Anne Williams. 2018. Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 7–14.

[34] Industrial Business Machines [IBM]. 2016. IBM SPSS Statistics for Windows, Version 24.

[35] Sooyeon Jeong, Deirdre E. Logan, Matthew S. Goodwin, Suzanne Graca, Brianna O'Connell, Honey Goodenough, Laurel Anderson, Nicole Stenquist, Katie Fitzpatrick, Miriam Zisook, et al. 2015. A social robot to mitigate stress, anxiety, and pain in hospital pediatric care. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 103–104.

[36] Noel D. Johnson and Alexandra A. Mislin. 2011. Trust games: A meta-analysis. *Journal of Economic Psychology* 32, 5 (2011), 865–889.

[37] Daniel Kermany, Kang Zhang, and Michael Goldbaum. 2018. Labeled optical coherence tomography (OCT) and Chest X-Ray images for classification. *Mendeley data* 2 (2018).

[38] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 5 (2018), 1122–1131.

[39] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[40] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The role of trust in human-robot interaction. In *Foundations of Trusted Autonomy*. Springer, Cham, 135–159.

[41] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.

[42] Rose Luckin, Wayne Holmes, Mark Griffiths, and Laurie B Forcier. 2016. Intelligence unleashed: An argument for AI in education. (2016).

[43] Kevin A. McCabe and Vernon L. Smith. 2000. A comparison of naive and sophisticated subject behavior with game theoretic predictions. *Proceedings of the National Academy of Sciences* 97, 7 (2000), 3777–3781.

[44] Harrison D. McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.

[45] Ali Meghdari, Azadeh Shariati, Minoo Alemi, Gholamreza R. Vossoughi, Abdollah Eydi, Ehsan Ahmadi, Behrad Mozafari, Ali Amoozandeh Nobaveh, and Reza Tahami. 2018. Arash: A social robot buddy to support children with cancer in a hospital environment. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 232, 6 (2018), 605–618.

[46] Neville Moray and T. Inagaki. 1999. Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control* 21, 4-5 (1999), 203–211.

[47] Nancy A. Obuchowski. 2003. Receiver operating characteristic curves and their use in radiology. *Radiology* 229, 1 (2003), 3–8.

[48] Kohei Ogawa, Christoph Bartneck, Daisuke Sakamoto, Takayuki Kanda, Tetsuo Ono, and Hiroshi Ishiguro. 2009. Can an android persuade you?. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE, 516–521.

[49] Emrah Onal, James Schaffer, John O'Donovan, Laura Marusich, S. Yu Michael, Cleotilde Gonzalez, and Tobias Höllerer. 2014. Decision-making in abstract trust games: A user interface perspective. In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE, 21–27.

[50] Scott Ososky, Tracy Sanders, Florian Jentsch, Peter Hancock, and Jessie YC Chen. 2014. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Unmanned Systems Technology XVI*, Vol. 9084. International Society for Optics and Photonics, 90840E.

[51] Raja Parasuraman and Christopher A. Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (2004), 51–55.

[52] David V. Pynadath, Ning Wang, and Sreekar Kamireddy. 2019. A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 171–178.

[53] Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems* 25, 4 (2009), 145–182.

[54] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

[55] Mohammad Nasser Saadatzi, Robert C. Pennington, Karla C. Welch, and James H. Graham. 2018. Effects of a Robot Peer on the Acquisition and Observational Learning of Sight Words in Young Adults With Autism Spectrum Disorder. *Journal of Special Education Technology* 33, 4 (2018), 284–296.

[56] Tracy Sanders, Kristin E. Oleson, Deborah R. Billings, Jessie Y.C. Chen, and Peter A. Hancock. 2011. A model of human-robot trust: Theoretical model development. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 55. SAGE Publications Sage CA: Los Angeles, CA, 1432–1436.

[57] Sebastian Schneider and Franz Kummert. 2018. Comparing the effects of social robots and virtual agents on exercising motivation. In *International Conference on Social Robotics*. Springer, 451–461.

[58] Richard Seymour and Gilbert L. Peterson. 2009. A trust-based multiagent system. In *2009 International Conference on Computational Science and Engineering*, Vol. 3. IEEE, 109–116.

[59] Thomas B. Sheridan. 2019. Extending three existing models to analysis of trust in automation: signal detection, statistical parameter estimation, and model-based control. *Human factors* 61, 7 (2019), 1162–1170.

[60] Rashmi R. Sinha and Kirsten Swearingen. 2001. Comparing recommendations made by online systems and friends.. In *DELOS workshop: personalisation and recommender systems in digital libraries*, Vol. 106.

[61] Mariacarla Staffa and Silvia Rossi. 2016. Recommender Interfaces: The More Human-Like, the More Humans Like. In *Social Robotics*, Arvin Agah, John-John Cabibihan, Ayanna M. Howard, Miguel A. Salichs, and Hongsheng He (Eds.). Springer International Publishing, Cham, 200–210.

[62] Harold Stanislaw and Natasha Todorov. 1999. Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers* 31, 1 (1999), 137–149.

[63] Laerd Statistics. 2015. *Two-way repeated measures ANOVA using SPSS Statistics*. Retrieved September 30, 2020 from https://statistics.laerd.com/

[64] Meg Tonkin, Jonathan Vitale, Suman Ojha, Jesse Clark, Sammy Pfeiffer, William Judge, Xun Wang, and Mary-Anne Williams. 2017. Embodiment, privacy and social robots: May i remember you?. In *International Conference on Social Robotics*. Springer, 506–515.

[65] Jonathan Vitale, Meg Tonkin, Sarita Herse, Suman Ojha, Jesse Clark, Mary-Anne Williams, Xun Wang, and William Judge. 2018. Be more transparent and users will like you: A robot privacy and user experience design experiment. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 379–387.

[66] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. 2010. When in Rome: the role of culture & context in adherence to robot recommendations. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 359–366. https://doi.org/10.1109/HRI.2010.5453165

[67] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams.. In *AAMAS*. 997–1005.

[68] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 109–116.

[69] Philipp Wintersberger, Anna-Katharina Frison, Andreas Riener, and Linda Ng Boyle. 2016. Towards a personalized trust model for highly automated driving. *Mensch und Computer 2016–Workshopband* (2016).

[70] Michelle Yeh and Christopher D. Wickens. 2001. Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors* 43, 3 (2001), 355–365.

[71] Tatjana Zorcec, Ben Robins, and Kerstin Dautenhahn. 2018. Getting Engaged: Assisted Play with a Humanoid Robot Kaspar for Children with Severe Autism. In *International Conference on Telecommunications*. Springer, 198–207.