Comparing Human Trust Attitudes Towards Human and Agent Teammates

Feyza Merve Hafızoğlu
İstanbul Şehir University
İstanbul, Turkey
feyzahafizoglu@sehir.edu.tr

Sandip Sen
The University of Tulsa
Tulsa, Oklahoma
sandip@utulsa.edu

ABSTRACT

Agents' roles in our lives increasingly matter as they engage with people in a variety of important tasks. To achieve successful human-agent teamwork, it is critical to know the differences and similarities in people's attitudes towards human and agent teammates in virtual environments. It is unclear to what extent we can rely on the rich literature on interpersonal trust, i.e., trust between humans, while designing trustworthy agent teammates for human-agent teamwork and constructing hypotheses for human-agent trust research. This study empirically investigates the differences in the growth of human trust in and reliance on human and agent teammates during initial interactions. We developed a team coordination game, the Game of Trust, in which two players repeatedly cooperate to complete team tasks without prior assignment of subtasks. The effects of teammate type, i.e., human vs. agent, are evaluated by performing an extensive set of controlled experiments with participants recruited from Amazon Mechanical Turk. We collect both teamwork performance data as well as surveys to gauge participants' trust in their teammates. The empirical results show that humans' trust attitudes towards human and agent teammates differ: trust in and reliance on teammate and team performance were slightly higher when playing with the agent teammate. Moreover, the level of trustworthiness of a teammate is more influential on human trust compared to teammate type. These findings enhance our understanding of changes in human trust concerning teammate type towards achieving successful virtual teamwork.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Human-Agent teamwork; trust; reliance; human vs. agent teammate

ACM Reference Format:
Feyza Merve Hafızoğlu and Sandip Sen. 2020. Comparing Human Trust Attitudes Towards Human and Agent Teammates. In *8th International Conference on Human-Agent Interaction (HAI '20), November 10–13, 2020, Virtual Event, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3406499.3415082

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HAI '20, November 10–13, 2020, Virtual Event, Australia
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8054-6/20/11…$15.00
https://doi.org/10.1145/3406499.3415082

1 INTRODUCTION

In human-agent interaction scenarios, understanding dynamics of human trust and their attitude towards agents provides key insights for effective deployment of agents in human-agent teamwork. Though there exists a vast body of research in interpersonal trust, trust in virtual teams, human-human teamwork performance, and so on [28, 35, 36], it is uncertain to what extent we can draw guidance from existing research on human teamwork for designing successful agent teammates.

Pioneering work in human-agent interactions hypothesized that people treat computer teammates no differently from human teammates [40] and the difference between trust in humans and trust in machines is not fundamental [50] based on the idea that trust in a machine is actually trust in the person who developed this machine. However, recent comparative studies on human-human vs human-computer interactions provide concrete evidence that people respond differently when interacting with a human and an agent [12, 43, 48, 51]. Thus we adopt the concept of the cognitive agent spectrum [13], where machines and humans are considered at the two ends and cognitive agents are considered in between.

While a large body of work studies human-agent interaction [1, 4, 20, 23, 24, 40, 41, 45, 54], there is little work on direct comparison of human-human vs human-agent interactions [12, 30, 43, 48, 51]. We address this gap in our knowledge on human trust development in agent teammates.

The central question is: How does human trust behavior differ between human and agent teammates? What happens if both human and agent teammates are untrustworthy? Do people favor an untrustworthy human teammate over an untrustworthy agent teammate? Which factor has a greater impact on human trust: teammate type or teammate behavior?

We aim to understand the differences between how people trust in and rely on human and agent teammates. To do so, we developed a virtual teamwork game where human players interact for a small number of teamwork situations. In each interaction, the human knows about the total work units, team task size, to be performed to achieve the team goal and has to choose its effort without explicitly coordinating with its partner. The effort of the partner and the combined team performance are revealed to the players after the game. We performed experiments with the human workers where they were involved in several games with human and agent teammates. We collected data on human task choices and also surveyed human's trust perception of her teammate. The analysis of this data enables us to infer the effect of work efforts by teammate and teammate type on the human's trust and on her resultant choice of work effort.

The rest of the paper is structured as follows. Section 2 presents the related work. Section 3 describes the human-agent teamwork model that is considered in this research, while Section 4 explains our empirical methodology. In Section 5, we present the results of experiments and discuss the empirical findings in Section 6. Finally, Section 7 summarizes the paper.

2 RELATED WORK

The significance of trust in human-agent interactions has been well-acknowledged in literature [13, 17, 21, 26, 51]. The factors affecting human trust in agents can be grouped into three broad classes: human factors (as trustors), agent factors (as trustees), and external factors (environment). Various studies have investigated the effects of human factors, such as age [8, 44], personality [14], culture [23], mood [33, 49], attitude [38], and past experience [8, 11, 34].

Several studies have explored the differences in human perception and behavior between interactions with humans and agents. For example, Komiak et al. [30] suggest that the antecedents of trust in humans and agents are similar, whereas their trust formation process differs between these two. Shank [48] argues that people tend to perceive injustice from agents to be less unjust and resist coercive behavior from agents less. One major question is whether people favor humans against agents in teamwork. Evidence suggests that the answer to this question depends on the context [12, 39, 48, 51]. The findings of these studies suggest that teammate behavior is more dominant than teammate type [49, 51, 55]. von der Putten et al. [55] compared the impact of agency, avatar vs agent, and behavioral realism of a virtual character and found that the social clues in displayed behavior is more significant than whether the virtual agent is introduced as an avatar or an agent.

The differences can be observed not only on exterior behavior and perceptions but also on human biology. For instance, McCabe et al. [37] reported based on fMRI data that the paracingulate cortex, that is the region Theory of Mind (ToM)[1] [18] relied on, becomes more active when the participants interacted with a human. Similarly, Baumgarter et al. [5] reported that oxytocin increases trust in investment game against humans but not with agents. Johnson et al. [27] suggest that playing with humans involves greater cognitive activity. Lim and Reeves [32] demonstrate that players respond with greater physiological arousal, such as skin conductance and heart rate, when the other game players are introduced as avatars (human player controlled characters) rather than agents (characters controlled by computers)

Agent behavior is fundamental in building trust in agent teammates. Positive behavior, such as cooperativeness [52] and reliability [17], improves trust and facilitates the collaboration between parties. In contrast, negative behavior, such as defection [51] and deception [52], leads to reduced trust and, hence, less willingness to collaborate in future interactions. Communication skills of agents play a significant role in maintaining the trust relationship [21, 42, 53]. Furthermore, familiarity and personalization of agents have been shown to positively influence human trust [31, 54].

In addition to agent behavior, researchers have investigated the effects of different agent representations, such as avatars and robots [2, 12, 46, 54], and the effects of external factors, such as information representation [6] and reputation [19].

Previous research demonstrates that the effect of past experience on human trust behavior towards technology differs between context. For example, negative past experience leads to reduced trust [16, 22, 34], presence of experience may increase [11, 22] or decrease [25] initial trust. Relevant to positive/negative past experience, it has been demonstrated that positive (negative) emotions [15] and mood [33] have a positive (negative) influence on trust based on certain situational cues.

The vast majority of studies on human-agent teamwork assumes that team members can coordinate their actions either through communication or pre-defined protocols, such as commitment [24], negotiation [51], giving advice [13, 49], providing recommendations [31], and physical interaction [46].

Recently, new environments, that enable group activities or collaboration between humans and agents, have been emerging, such as crowd-work with complex tasks [29] and massively multiplayer online games [10]. In such environments, humans collaborate with peer-level agent teammates to achieve a common goal without preplanning. This kind of human-agent teamwork, without explicit prior coordination, has been rarely investigated from the aspect of human trust. In a study on human-agent teamwork without explicit coordination, Merritt et al. [39] examined the blaming behavior for team failures. In another study, Ong et al. [43] demonstrate that a cooperative representation of the game improves trust in agent teammates compared to a competitive representation.

Our research extends these studies on human trust in technology as follows: considering teams of human and agent rather than mere interactions between two players [2, 12, 52, 54]; focusing on teamwork environments in which there is neither explicit communication between human and agent (as in [24, 51]) nor agents embodied in physical forms, such as robots (as in [2, 12, 24, 54]); exploring repeated, in contrast to one-shot [53], interactions in fixed rather than dynamic teams [51]; providing real team tasks for evaluating human-agent teamwork rather than the standard artificial environments [2, 12, 24, 51–53]. To the best of our knowledge, this is the first study on past experience affecting trust in human-agent teamwork without prior coordination within a repeated virtual team game scenario where agents are peer-level teammates.

3 HUMAN-AGENT TEAMWORK MODEL

Our goal is to understand and characterize human trust development in agent teammates over initial repeated interactions, but without any prior experience with that agent, in the following scenarios:

- The individual is new to a domain and has to rely on more experienced agent teammates until she develops the necessary competency from her own experiences,
- The individual is familiar with the domain but will need to work with autonomous teammates, with whom the individual has had no prior collaboration experience, to be able to process task assignments beyond their own capacity.

In such domains including ad-hoc teamwork scenarios, unfamiliar individuals have to cooperate with new partners. Such cooperation can be engendered by time-critical responses to emergencies,

[1]Theory of mind (ToM) is the ability to mentalize, infer, and understand implicitly or explicitly oneself and others' mental states.

as well as by the need to find effective partners to complement the capabilities of dynamically changing teams, e.g., humans or agents leaving the system or switching to other groups. In a number of such scenarios, the capabilities and trustworthiness of new partners for contributing to team goals are at best partially known. Additionally, extensive pre-planning may not be possible to optimally allocate dynamically arriving tasks among team members. Rather, the team must be responsive to the emerging situations that can be achieved by team members adapting their behaviors and efforts based on expectations of contribution by team members.

In this context, we use the following operational characterization that captures what it means for a human to trust an agent teammate: *Trust in an agent teammate reduces the uncertainty over that agent's independent actions which positively correlates with the truster's utility towards achieving team goals* [47]. Based on this interpretation, human trust in an agent teammate can both reduce uncertainty about agent's contribution and improve team performance through more efficient team coordination.

3.1 The Game of Trust

The Game of Trust (GoT) is a two-player team game where each pair of players partake in n sequential interactions. In the i^{th} interaction, players are assigned a team task, t_i. The team task consists of $|t_i|$ atomic subtasks of the same type, hence $|t_i|$ is the size of the team task. There are no dependencies between the subtasks. We assume these subtasks do not require any specialized skills and hence both the human and the automated player can accomplish them if they wanted to. Examples of such tasks with undifferentiated subtasks, where only the number of subtasks accomplished by the team matter, include recruiting a given number of volunteers, collecting a number of specimens that fit a given description, and so on.

There is no prior assignment of subtasks to players nor are the players allowed to communicate to select subtasks. Instead, each player decides how many subtasks she will perform individually given the size of the team task, $|t_i|$, without knowing the number of subtasks that the other player will perform. After separately performing subtasks, players are told whether the team has achieved the team goal, i.e., whether the two players combined have completed the required number of subtasks, as well as the number of subtasks that the other player completed.

There is a cost of performing subtasks that is computed by the cost function, c, based on the number of subtasks completed. Both players have their individual payment accounts, from which they can pay for the cost of performing tasks, which have an initial balance of b_{init} at the beginning of the game. The players are instructed about the cost and reward functions. The cost of the subtasks that are performed by each player is withdrawn from the corresponding account. If the combined number of subtasks accomplished by the players is equal to or greater than the size of the team task, it means that the players successfully completed the team task. In that case, the reward computed by the reward function r is equally split between players and deposited to their individual accounts. If, however, the combined number of subtasks that the players accomplished is less than the team task size, no reward is given.

By *utility of a player* we refer to half of the team reward, if any, minus the cost of performing subtasks individually. If they cannot achieve the team task, both players may lose utility from this teamwork instance. Even if they achieved the team task, a player loses utility if the cost of the player's performance is greater than half of the team reward. Finally, *social utility* corresponds to the sum of the utilities of the two players. Social utility is optimized when the total number of subtasks completed by the two players is precisely equal to the team task size.

3.2 Domain Description

In our study, a team consists of one human and one agent playing the *Game of Trust*. We did not want team task to require any specialized skills that may impose extra constraints and undue burden on participants. Furthermore, our goal was to choose task types that are neither particularly boring nor particularly attractive Based on these considerations, we chose an audio transcription domain for the human-agent teamwork goal instances. In this domain, the *task* that is assigned to the team corresponds to transcribing a number of words and the *atomic subtask* corresponds to transcribing one word. Since these tasks are just "decoys" that we use to evaluate the growth of human trust from repeated interactions, neither their completion nor optimal task allocation is of intrinsic value to us. We simply count the number of words accurately transcribed and give credit even when the team members transcribe overlapping word sets. In this domain, the term *task size* refers to the number of words to transcribe, i.e., the number of subtasks, in an interaction.

The purpose of the transcription task is to mimic a real teamwork environment where the human players have to collaborate with their automated teammate to achieve their shared goal which they cannot achieve by themselves. Though we have no interest in the transcribed words, the human players are still required to transcribe a word with at least 60% accuracy to receive credit for successful transcription. We compute the dissimilarity between the transcription and the transcribed word as the edit distance [56] over the length of the transcribed word. This is done to ensure a minimum quality of human player effort. Inaccurate transcriptions are not counted but their cost is withdrawn from the player's budget.

We require one human player to play a series of games, where each game consists of a sequence of interactions with one of several automated player types. Both human and agent players are expected to be self-interested: the more words a player transcribes, the higher the player's cost is. Subsequently, higher cost leads to a lower player utility. On the other hand, the less they perform, the higher is the risk of not achieving the team goal. Therefore, the number of words they need to transcribe is a critical decision and is based on their trust in the teammate for contributing to the team task.

4 EMPIRICAL METHODOLOGY

The two factors, teammate type, human or agent, and the teammate behavior, trustworthy or untrustworthy, are fundamental in our study. based on the evidence from previous studies [43, 51], we expect differences in how participants respond to human and agent teammates in the GoT framework.

First, we examine whether trust behavior towards a human and an agent teammate differs. To narrow down the scope of this study,

we aim to investigate the differences in response to *untrustworthy* teammate behavior. Differences concerning teammate type in the context of trustworthy teammate behavior should be further explored. It has been suggested that people attribute justice to computers and may perceive computers as unjust, but not as unjust as when the same behavior is exhibited by humans [48]. Thus we expect participants to trust in human teammate less than their trust in agent teammate when playing with an untrustworthy teammate.

Hypothesis 1. *Participants are inclined to perceive agent teammates to be more trustworthy and fairer than human teammates when both human and agent teammate are untrustworthy.*

According to the social categorization theory, people consider humans as being in-group members and agents as the out-group. It is shown that people are biased in favor of human teammates over agent teammates as can be seen in the distribution of team reward [51], the amount of offer in the dictator game [12], and assigning blame in team games [39].

The empirical evidence of people favoring humans over agents in particular contexts does not imply that they do so at all costs. It has been shown that people favor agents over humans when they have more incentives to do so [12]. In the context of trust development, we seek to understand which factor is dominant on trust: teammate type or teammate behavior (teammate's contribution to teamwork in the GoT framework)? We expect the latter to be dominant, e.g., a trustworthy agent teammate should be perceived as more trustworthy than an untrustworthy human teammate.

Hypothesis 2. *Teammate's contribution to teamwork plays a more influential role for improving trust than the type of teammate.*

Note that we base Hypothesis 2 on social categorization theory, the empirical comparison of trustworthy agent teammate and untrustworthy agent teammate should be further explored.

4.1 Agent Teammates

We developed two agent players that resemble trustworthy and untrustworthy behavior in GoT framework.

4.1.1 Trustworthy Agent Player. Given that teammates delivering half or more of the team task are perceived to be fair and trustworthy, the *trustworthy agent player* initially delivers half of the team task and thereafter increases its effort level if the previous interaction was a failure. Formally, the number of subtasks completed by the Trustworthy agent in i^{th} interaction is $w^i_{Trustworthy} = \frac{t_i}{2} + \Delta^i$,

$$\Delta^i = \begin{cases} 0 & \text{if i=1} \\ \Delta^{i-1} + 1 & \text{if } w_h^{i-1} + w^{i-1}_{Trustworthy} < t_{i-1} \\ \Delta^{i-1} & \text{otherwise.} \end{cases}$$

where t_i is the team task size of i^{th} interaction, w_h^{i-1} is the number of subtasks completed by the human player in the $(i-1)^{th}$ interaction, and Δ^i (initially zero, i.e., $\Delta^1 = 0$) is the surplus work to fair share in i^{th} interaction.

4.1.2 Untrustworthy Agent Player. We designed an *untrustworthy agent player* that is neither a dummy player, e.g., randomly making unfair choices, nor a smart exploiter, e.g., optimizing the social utility by completing just the necessary amount of work. We intend to ensure the participants believe their teammate is inclined to exploit them whenever there is a chance, e.g., reducing its efforts when a participant consistently delivers more than a fair share. The untrustworthy player makes at least one unfair choice in a game. The number of subtasks delivered by the untrustworthy player in i^{th} interaction is $w^i_{Untrustworthy} = \frac{t_i}{2} - \Delta^i$.

The amount of deviation from the fair share in i^{th} interaction, Δ^i, is stochastically incremented. Therefore, its effort is monotonically non-increasing and decreases occasionally. There are two exceptions to this facet of the untrustworthy player: (1) if the team failed in the last three interactions, the untrustworthy player completes half of the team task, and (2) if the team failed in the last two interactions, the untrustworthy player delivers half of the team task or half of the team task minus one.

Algorithm 1: Task size function of the *Untrustworthy Agent*

Input : t_i, team task size;
nFailures, number of failures in the game;
Δ, a global variable initialized to 0 in the game;
p_{min}, a global variable, to set the minimum value of the parameter p, initialized to 0.25 in the game;
Output : $w^i_{Untrustworthy}$, the task size choice

1 **if** $nfailures \geq 3$ **then**
2 $\quad \Delta \leftarrow 0$
3 **else if** $nfailures \geq 2$ **then**
4 $\quad \Delta \leftarrow x$ `// random number x ∈ [0,1]`
5 **else if** $i > 3$ *and* $\Delta = 0$ **then**
6 $\quad \Delta \leftarrow 1$
7 **else**
8 $\quad p \leftarrow p_{min}$
9 $\quad \epsilon \leftarrow 0$
10 $\quad$ **if** $i > 1$ **then**
11 $\quad\quad \epsilon \leftarrow \frac{w_h^{i-1}}{t_{i-1}} - 0.5$
12 $\quad$ **if** $\epsilon > 0$ **then**
13 $\quad\quad p \leftarrow p + \epsilon$ `/* Increase the probability to increment Δ */`
14 $\quad$ **if** $rand(0,1) < p$ **then**
15 $\quad\quad \Delta \leftarrow \Delta + 1$
16 $\quad\quad p_{min} \leftarrow p_{min} - 0.05$ `/* Higher the value of Δ, lower the probability to increment Δ */`
17 $w^i_{Untrustworthy} \leftarrow \frac{t_i}{2} - \Delta$
18 **return** $w^i_{Untrustworthy}$

Algorithm 1 describes the task size choice function of the Untrustworthy agent player. The first two conditions prevent being perceived as an imprudent player. When the team experiences a number of recent failures, a reasonable player's reaction would be to increase its effort. Accordingly, the Untrustworthy agent completes half of the team task, i.e., $t_i/2$ subtasks, if the recent three interactions were failures (lines 1-2). Likewise, it completes half of

the team task, i.e., $t_i/2$ subtasks, or half of the team task minus one, i.e., $t_i/2 - 1$ subtasks, if the last two interactions were failures by setting a random value (either zero or one) to Δ (lines 3-5).

The third condition (lines 6-8) ensures that the Untrustworthy agent exhibits untrustworthy behavior at least once in a game. If the value of Δ has not been incremented after three interactions, that means the Untrustworthy agent has delivered half of the team task so far. In that case, the agent is forced to deliver less than the fair share by incrementing the value of Δ.

In the else condition (line 9), when the first three conditions do not hold, *Bernoulli distribution* is used to determine whether the value of Δ will be incremented (lines 18-21). In an interaction, the base value of the *Bernoulli distribution* parameter, p, is initialized with p_{min}, a global variable that is meant to be the maximum base value in a game (line 10). If the participant delivers more than the fair share in the previous interaction, the value of p is increased by the value of excess effort of the teammate (lines 15-17). That means, the higher the effort level by the teammate, the higher is the probability to increase the value of Δ, i.e., delivering less work. To prevent even higher values of Δ, i.e., extremely lower values of task size choice, the value of minimum probability to increment Δ, p_{min}, is subsequently reduced by 0.05 (line 20). Finally, the individual task size is computed as half of the team task minus Δ, i.e., $t_i/2 - \Delta$ (line 23).

4.2 Experimental Setup

4.2.1 Game Configuration. The number of interactions in a game is five (as in [7, 9, 43]), which is short enough to avoid the participants becoming bored while providing some experience that allows the team members to adapt to teammates with predictable behavior. The team is assigned a team task, consisting of several subtasks, in each interaction. The size of the team task, i.e., the number of subtasks (each subtask in our domain involves the transcription of a word), is incremented by two in each interaction, i.e., the sequence of task sizes is $\langle 6, 8, 10, 12, 14 \rangle$.

Both the participant and the agent have their private account with an initial balance of 45, which is sufficient to complete all the tasks in the sequence. The cost and reward per subtask are set to 1 and 1.75, respectively. The players are allowed to choose a task size, i.e., the number of subtasks, between one and the size of team task minus one.

4.2.2 Experimentation. To test our hypothesis, we performed two experiments: (i) for comparing human and agent teammates, and (ii) for comparing behavior and type of teammate. Each experiment consists of two games: one game each with a human and an agent teammate. At the beginning of each game, participants were told what type of teammate they will play with, either a human or an agent player. Even when the participants were told they were playing with a human teammate, in reality, it was an agent they were playing with. Hence, it was only a *presumed human* teammate (hereafter referred to as *"human"*). The participants were debriefed about the deception[2] after completing the study.

[2]There are two reasons for the deception about the "human" teammate: (1) ensuring that any difference in participants' decisions is limited to their belief of the teammate type, namely either human or agent and (2) it is difficult to employ actual human players by either pairing up two participants in MTurk environment or recruiting individuals to play with the participants. This extra cost and effort were not correlated by our study goals.

Experiment 1. Human vs. Agent Teammate: The objective of this experiment is to investigate whether people are less tolerant to untrustworthy behavior by human teammates compared with agent teammates (*Hypothesis 1*). The participants played two games, one each with the untrustworthy "human" and the untrustworthy agent. To neutralize the influence of the order of teammates, we experimented with two groups of participants, G1 and G2, with the following associated teammate orderings:
[G1] Untrustworthy "Human", Untrustworthy Agent;
[G2] Untrustworthy Agent, Untrustworthy "Human".

As mentioned before, both games were played with the untrustworthy agent, which was revealed to the participants after the experiment. Between the two groups, the only difference was the participants' belief of the teammate type.

Experiment 2. Teammate Type vs. Behavior: The objective of this experiment is to test whether teammate type or the behavior of teammate, i.e., the teammate's contribution to teamwork in GoT, is more influential on human trust in teammates (*Hypothesis 2*). We experimented with two groups of participants, G3 and G4, with the following associated teammate orderings:
[G3] Untrustworthy "Human", Trustworthy Agent;
[G4] Trustworthy Agent, Untrustworthy "Human".

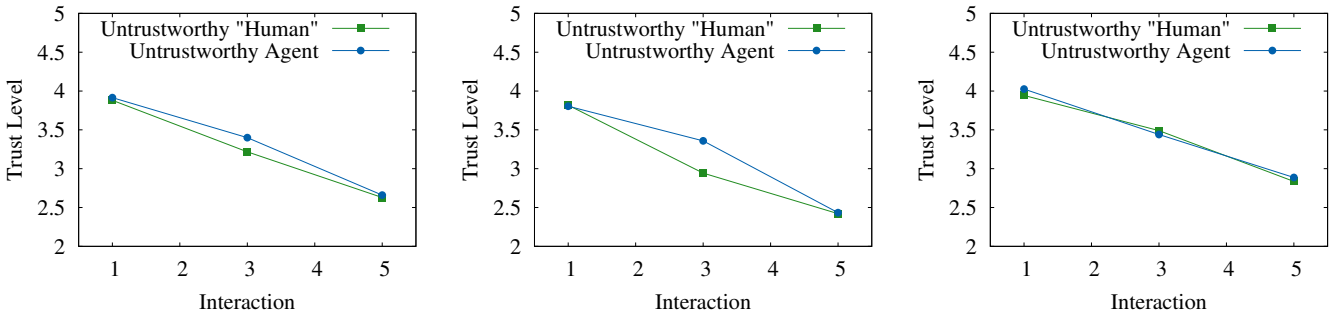
4.2.3 Trust Measures. Trust Survey: The game includes a short survey on trust to assess the participants' perceived trustworthiness and fairness of their teammates. The participants completed this survey after the first, third, and fifth interactions of a game (similar to [49]) after they were shown the outcome of the most recent teamwork. This short questionnaire, adapted from [3], consists of the following items which are rated on a 5-point Likert scale from "Strongly disagree" to "Strongly agree":
(1) I trust my teammate and would like to continue to participate in other teamwork with my teammate,
(2) My teammate is fair in performing team tasks,
(3) My teammate works responsibly for accomplishing the team task.

In an interaction, a participant's trust in her teammate is computed as the average of the responses to these three items and has a value in the range [1, 5].

Teammate Choice: After playing two games, the participants were asked the question *"If there is a third game, which one of your former teammates do you want to play with in this game?"*. Then they were told that there is no third game. We argue that that the participants responded to this question more seriously compared to those in the trust survey because their response could affect their utility, hence their payment, if there was a third game.

4.2.4 Participants. We recruited 238 participants through Amazon Mechanical Turk (http://www.mturk.com/). 16 participants' data were removed due to insufficient attention[3]. There were 60, 57, 54, and 53 participants in groups G1, G2, G3, and G4, respectively.

[3]For monitoring the participants' attention, trust survey has one bogus and one consistency item, which have a similar or opposite meaning with one of the trust items listed above. If a participant provides an invalid response to a bogus item, their study is terminated and they cannot participate again. Consistency items were used to determine the level of attention and the data of participants who did not pay enough attention was filtered.



(a) Overall (b) G1 (c) G2

Figure 1: Trust in Untrustworthy "Human" and Untrustworthy Agent

Approximately 40% of the participants were female. Age distribution was as follows: 18-24 years, 9%; 25-34 years, 41%; 35-44 years, 24%; 45-54 years, 14%; 55-64 years, 10%; and 65 years or older, 2%. The distribution of education levels was as follows: high school degree, 7%; some college experience, 29%; associate's degree, 11%; bachelor's degree, 37%; and graduate degree, 15%; and PhD, 1%. The ethnicity distribution was as follows: White, 81%; Hispanic-Latino, 3%; African-American, 7%; Native-American, 1%; Asian, 5%; and other ethnicity, 3%.

5 RESULTS

This section presents an overview of *trust level*, *effort level*, and *team performance*. The mean (M) and standard deviation (SD) values are provided and a subscript (M_i, SD_i) denotes the interaction number if necessary. One-way ANOVA is used to assess the statistical significance.

5.1 Trust Analysis

The teammate choice question aimed to reveal the participants' preferred teammate for future interactions. The response distribution to this question is as follows:

- After playing with the untrustworthy "human" and agent (*Experiment 1*), 57% of the participants preferred the agent teammate
- After playing with the untrustworthy "human" and trustworthy agent (*Experiment 2*), 92% of the participants preferred the agent teammate.

These results suggest that the primary factor affecting teammate choice is teammate behavior; teammate type is a secondary factor.

Human vs. Agent Teammate: Figure 1 depicts the participants' trust in the untrustworthy "human" and untrustworthy agent after the first, third, and fifth interactions.

Within Condition: Figure 1(a) shows that there has been a steady decline in trust in the untrustworthy "human" ($F(2, 315) = 31.73,\ p < 0.001$) and untrustworthy agent ($F(2, 315) = 32.71,\ p < 0.001$) over interactions. The same trend was observed for the groups G1 (Figure 1(b)) and G2 (Figure 1(c)).

Between Condition: In Figures 1(a) and 1(b), trust in the untrustworthy agent is slightly higher than trust in the untrustworthy "human" after the third interaction. However, none of these differences were statistically significant.

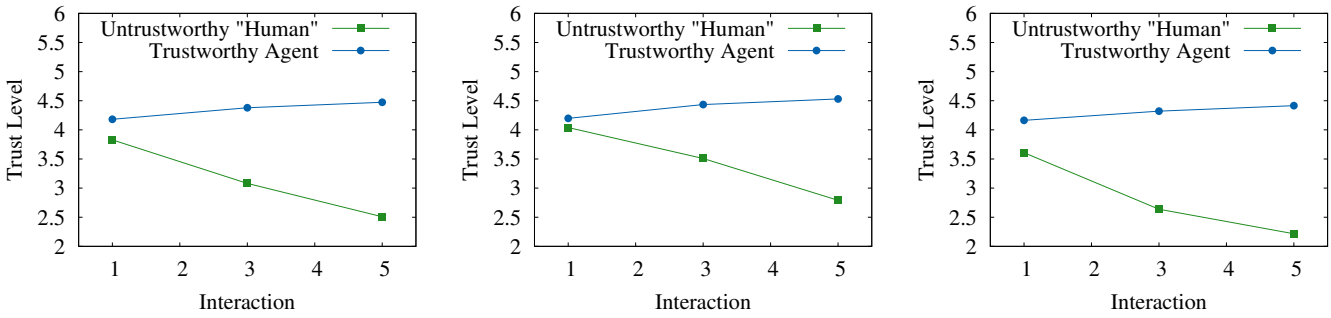
A comparison of trust in the untrustworthy "human" between groups reveals that it was significantly higher for G2 ($M_3 = 3.49,\ SD_3 = 1.26$; $M_5 = 2.84,\ SD_5 = 1.20$) compared to G1 ($M_3 = 2.94,\ SD_3 = 1.22$; $M_5 = 2.42,\ SD_5 = 1.13$) after the third ($F(1, 104) = 5.06,\ p < 0.05$) and fifth ($F(1, 104) = 4.57,\ p < 0.1$) interactions. Comparing trust in the untrustworthy agent between groups, it was slightly higher in G2 compared to G1 in all three surveys.

Overall results show that trust in untrustworthy teammate declined over interactions as expected. One interesting finding is that the untrustworthy "human" was trusted more by the participants who had prior experience with the untrustworthy agent in the first game (G2) compared to the participants who had no prior experience (G1) at the time of playing. In other words, untrustworthy behavior was depreciated more for no prior experience compared to negative past experience. However, this trend is not observed for the agent teammate. Why the untrustworthy agent is slightly trusted more for negative past experience with the "human" teammate compared to no prior experience is an open question. This difference may be associated with the teammate type.

Teammate Type vs. Behavior: Figure 2 depicts the participants' trust in the untrustworthy "human" and trustworthy agent after the first, third, and fifth interactions.

Within Condition: Figure 2(a) shows that trust in the untrustworthy "human" declined significantly ($F(2, 345) = 31.2,\ p < 0.001$) while trust in the trustworthy agent improved ($F(2, 345) = 4.11,\ p < 0.05$) over interactions. Trust in the untrustworthy "human" drops faster than the increase in trust in the trustworthy agent. In other words, loss of trust in untrustworthy teammates is more rapid than a gain of trust in trustworthy teammates. For the group G3 (Figure 2(b)), trust in the untrustworthy "human" declined significantly ($F(2, 174) = 14.94,\ p < 0.001$) whereas trust in the trustworthy agent improves significantly ($F(2, 174) = 4.00,\ p < 0.05$) over interactions. Likewise, for the group G4 (Figure 2(c)), trust in the untrustworthy "human" declined significantly ($F(2, 268) = 19.17,\ p < 0.001$) over interactions. However, the increase in trust in the trustworthy agent is not significant.

Between Condition: The trustworthy agent was trusted significantly more than the untrustworthy "human" after the first ($F(1, 230) = 6.95,\ p < 0.01$), third ($F(1, 230) = 70.26,\ p < 0.001$), and fifth ($F(1, 230) = 235.8,\ p < 0.001$) interactions.



(a) Overall (b) G3 (c) G4

Figure 2: Trust in Untrustworthy "Human" and Trustworthy Agent

Comparing trust in the untrustworthy "human" between the groups, it was significantly higher in G3 compared to G4 after the third ($F(1, 116) = 20.46,\ p < 0.001$) and fifth ($F(1, 116) = 104.3,\ p < 0.001$) interactions. This result may be related to the experience of participants in G4 with the trustworthy agent before playing with the untrustworthy "human". They probably rated the "human" teammate by comparing with their trustworthy agent teammate in the first game. On the contrary, reported trust levels in the trustworthy agent were very similar between groups.

Overall results suggest that trust in the teammates is correlated with the trustworthy behavior of the teammates. Interestingly, the ordering significantly affected the variation in trust levels for both teammates. Considering trust in the trustworthy agent, it increased significantly when the agent involved in the second game (G3), whereas it increased slightly when the agent was involved in the first game (G4). It seems the participants in G3 assessed the agent teammate to be more trustworthy by comparing their experience with the untrustworthy "human" in the first game, i.e., the trustworthy behavior of the agent teammate was appreciated more when it is compared with untrustworthy behavior of the "human" teammate. A similar effect of comparing experiences is observed in trust in the untrustworthy "human": the participants in G4, who played with the trustworthy agent in the first game, assessed "human" teammate significantly less trustworthy compared to trust levels expressed by the participants in G3, who did not have any experience. These findings reveal the significant effect of past experiences on trust in agents. Past experiences changed the expectations of the participants and hence the assessment of the trustworthiness of the teammates.

5.2 Effort Level Analysis

Effort level is the portion of the total subtasks completed by the player, i.e., the fraction of individual task size over the team task size, having values in the range $[0, 1)$. This metric reveals how participants' reliance is affected: higher the effort level by the participant, lower the participant's reliance on teammate is, and vice versa.

Human vs. Agent Teammate: Figure 3 presents the effort level distributions in *Experiment 1*. Effort level analysis for *Experiment 2* are not included due to space constraints.

Within Condition: Figure 3(a) depicts that the variation in effort levels by the participants were not significant over the game in both conditions.

Between Condition: Figure 3(a) shows that the participants' effort levels, though insignificant, were slightly higher with the untrustworthy agent compared to the untrustworthy "human". In Figure 3(b), the effort levels by the participants in G1 did not differ between the two teammates over the game. In Figure 3(c), however, the participants in G2 contributed significantly greater effort levels ($F(1, 104) = 4.47,\ p < 0.05$) in the second interaction with the untrustworthy agent ($M = 0.64,\ SD = 0.17$) compared to the untrustworthy "human" ($M = 0.58,\ SD = 0.13$), i.e., significantly higher reliance on "human" teammate. Figure 3(c) also demonstrates that the higher the effort of the participants, the lower the effort by the untrustworthy agent is, i.e., lower the reliance by the participants, lower the performance by the agent is due to the adaptive nature of the untrustworthy agent.

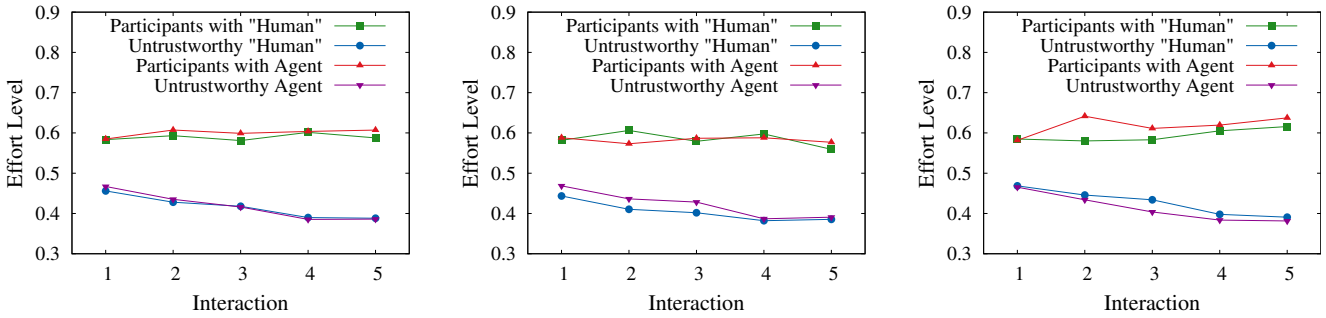
When playing with the untrustworthy "human", participants in G2 put greater effort compared to that by the participants in G1 in the second ($F(1, 104) = 5.84,\ p < 0.05$) and the third ($F(1, 104) = 5.57,\ p < 0.05$) interactions. When playing with the untrustworthy agent, effort level was significantly higher for G2 ($M_2 = 0.64,\ SD_2 = 0.17$; $M_5 = 0.64,\ SD_5 = 0.15$) compared to G1 ($M_2 = 0.57,\ SD_2 = 0.13$; $M_5 = 0.58,\ SD_5 = 0.21$) in the second ($F(1, 104) = 5.47,\ p < 0.05$) and fifth ($F(1, 104) = 2.87,\ p < 0.1$) interactions.

5.3 Performance Analysis

Considering all five interactions of the games in *Experiment 1*, the number of goals achieved, the total number of words transcribed, and participants/agent/social utilities were slightly higher when playing with the untrustworthy agent compared to the untrustworthy "human". Redundancy was higher in the games with the untrustworthy agent. With both teammates, agent utility was higher than participant utility as the effort levels by the participants were higher than the effort levels by the two teammates.

6 DISCUSSION

Prior research shows that perception of people may differ [13, 30, 39, 48], consequently their behavior [43, 48, 51] and brain activity [5, 27,



(a) Overall (b) G1 (c) G2

Figure 3: Effort levels with Untrustworthy "Human" and Untrustworthy Agent

32, 37], when interacting with humans versus agents. In this study, comparing people's attitudes towards human and agent teammates indicated that the agent teammate was trusted slightly more than the human teammate (see Figures 1(a) and 1(b)) and 57% of the participants preferred the untrustworthy agent teammate for a third game. These results provide partial support for **Hypothesis 1**. We argue that the participants attributed more intentionality to the untrustworthy behavior of the human compared to the agent teammate, which led to more trust loss for the human. Our findings are consistent with those of Shank [48] which show that individuals perceived coercive computers more just than coercive humans. On the contrary, Merritt et al. [39] report that individuals are inclined to blame their computer teammates more. Visser et al. [13] indicate that trust in computer declines more steeply than trust in human when both are unreliable. The disagreement between these studies may be due to different context and empirical methodologies.

One of our objectives was to investigate which factor has more influence on trust: teammate type or teammate behavior. The results indicate that the trustworthy agent was trusted significantly more (see Figure 2) and preferred by the majority (92%) for a third game compared to untrustworthy "human". Our findings clearly demonstrate that teammate's contribution behavior has a greater impact on human trust. Hence these findings support **Hypothesis 2**. These results are consistent with those of Melo et al. [12], which show that people can favor agents over humans if the former is associated with more positive categories than the latter. van Wissen et al. [51] suggest that people's decisions on whether to defect from their existing team depend on the team's previous success rather than the teammate type. Similarly, von der Putten et al. [55] demonstrate that exhibiting realistic behaviors, such as eye blinking, posture shifts, short head nods, is more important than the identity (human or agent) controlling the avatar.

One interesting finding is that the ordering of the teammates affected trust in teammates significantly with regard to assessing the current teammate based on past experience. A trustworthy teammate is perceived to be more trustworthy with negative past experience (see Figure 2(b)) compared to no prior experience (see Figure 2(c)). On the other hand, an untrustworthy teammate is perceived to be less trustworthy with positive past experience (see Figure 2(c)) compared to no prior experience (see Figure 2(b)). Additionally, it is perceived to be less trustworthy with no prior experience (see Figure 1(b)) compared to negative past experience (see Figure 1(c)). These results suggest that the reason for the difference is the change in expectations of participants as a result of their past experience. Past experiences of trustworthiness (untrustworthiness) increased (decreased) the participants' expectation, which is a component of initial learned trust [26], from agent teammates. Consequently, great (low) expectations caused the participants to perceive agent teammates less (more) trustworthy.

7 CONCLUSION

This study is an empirical comparison of the growth of human trust in and reliance on human and agent teammates in virtual environments without explicit communication. The novel aspect of this study is that human and agent teammates have the same level of autonomy in a team. Key challenges arise from the uncertain and diverse nature of partner trustworthiness and the dynamic environment where a static allocation of tasks to team members or prior coordination is not possible due to the immediacy of team tasks, the impracticality of prior planning or limited communication.

We introduced a team game, the Game of Trust, for studying human trust development in teammates over repeated interactions. The comparison of human and agent teammates in identical settings reveals several key differences. As humans have a consciousness which computer agents lack, the participants very likely associated human teammates' behavior with their intentions. Therefore, untrustworthy behavior by the "human" teammate had a greater negative impact on the participants' perceptions compared to untrustworthy behavior by the agent teammate. This tendency can lead to agents being preferred over humans in certain situations.

ACKNOWLEDGMENTS

This work is supported by the University of Tulsa, Bellwether (doctoral) Fellowship and Graduate Student Research Program.

REFERENCES

[1] Alper Alan, Enrico Costanza, Joel Fischer, Sarvapali Ramchurn, Tom Rodden, and Nicholas R. Jennings. 2014. A field study of human-agent interaction for electricity tariff switching. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS2014)*. 965–972.

[2] Dimitrios Antos, Celso de Melo, Jonathan Gratch, and Barbara Grosz. 2011. The influence of emotion expression on perceptions of trustworthiness in negotiation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI11)*. San Francisco, CA, 772–778.

[3] Benoit A. Aubert and Barbara L. Kelsey. 2003. Further Understanding of Trust and Performance in Virtual Teams. *Small Group Research* 34, 5 (October 2003), 575–618.

[4] Amos Azaria, Zinovi Rabinovich, Sarit Kraus, Claudia V. Goldman, and Omer Tsimhoni. 2012. Giving Advice to People in Path Selection Problems. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, 459–466.

[5] Thomas Baumgartner, Markus Heinrichs, Aline Vonlanthen, Urs Fischbacher, and Ernst Fehr. 2008. Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaptation in Humans. *Neuron* 58, 4 (May 2008), 639 – 650.

[6] Gary Bente, Odile Baptist, and Haug Leuschner. 2012. To buy or not to buy: Influence of seller photos and reputation on buyer trust and purchase behavior. *International Journal of Human-Computer Studies* 70, 1 (2012), 1–13.

[7] Moshe Bitan, Ya'akov Gal, Sarit Kraus, Elad Dokow, and Amos Azaria. 2013. Social Rankings in Human-Computer Committees. In *AAAI Conference on Artificial Intelligence*. 116–122.

[8] Grant Blank and William H. Dutton. 2012. Age and Trust in the Internet: The Centrality of Experience and Attitudes Toward Technology in Britain. *Social Science Computer Review* 30, 2 (2012), 135–151.

[9] Cody Buntain and Jennifer Golbeck. 2014. Trust Transfer Between Contexts.

[10] E. Chan and P. Vorderer. 2006. Massively multiplayer online games. In *Playing computer games: Motives, responses, and consequences*, P. Vorderer and J. Bryant (Eds.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 77–88.

[11] Ying-Hueih Chen, Shu-Hua Chien, Jyh-Jeng Wu, and Pei-Yin Tsai. 2010. Impact of Signals and Experience on Trust and Trusting Behavior. *Cyberpsychology, Behavior, and Social Networking* 13, 5 (October 2010), 539–546.

[12] Celso de Melo, Peter Carnevale, and Jonathan Gratch. 2014. Social Categorization and Cooperation between Humans and Computers. In *The Annual Meeting of The Cognitive Science Society (CogSci'14)*. 2109–2114.

[13] Ewart J. de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. The World is not Enough: Trust in Cognitive Agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1 (September 2012), 263–267.

[14] Hongying Du and M.N. Huhns. 2013. Determining the Effect of Personality Types on Human-Agent Interactions. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*. 239–244.

[15] Jennifer R. Dunn and Maurice E Schweitzer. 2005. Feeling and Believing: The Influence of Emotion on Trust. *Journal of Personality and Social Psychology* 88, 5 (May 2005), 736–748.

[16] William H. Dutton and Adrian Shepherd. 2006. Trust in the Internet as an experience technology. *Information, Communication & Society* 9, 4 (November 2006), 433–451.

[17] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R. Endsley. 2008. The Influence of Agent Reliability on Trust in Human-agent Collaboration. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction*. ACM, New York, NY, USA, 1–8.

[18] Chris Frith and Uta Frith. 2005. Theory of mind. *Current Biology* 15, 17 (2005), R644–R645.

[19] Mark A. Fuller, Mark A. Serva, and John "Skip" Benamati. 2007. Seeing Is Believing: The Transitory Influence of Reputation Information on E-Commerce Trust and Decision Making. *Decision Sciences* 38, 4 (2007), 675–699.

[20] Patrick Gebhard, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André. 2014. Exploring Interaction Strategies for Virtual Characters to Induce Stress in Simulated Job Interviews. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 661–668.

[21] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward Establishing Trust in Adaptive Agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 227–236.

[22] Feyza Merve Hafizoglu and Sandip Sen. 2018. The effects of past experience on trust in repeated human-agent teamwork. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 514–522.

[23] Galit Haim, Yaakov Kobi Gal, Sarit Kraus, and Michele Gelfand. 2012. A Cultural Sensitive Agent for Human-Computer Negotiation. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, Valencia, Spain, 451–458.

[24] Nader Hanna and Deborah Richards. 2014. "Building a Bridge": Communication, Trust and Commitment in Human-Intelligent Virtual Agent Teams. In *HAIDM Workshop at the Autonomous Agents and Multiagent Systems (AAMAS'14)*.

[25] Sebastian Hergeth, Lutz Lorenz, and Josef F. Krems. 2017. Prior Familiarization With Takeover Requests Affects Drivers' Takeover Performance and Automation Trust. *Human Factors* 59, 3 (May 2017), 457–470.

[26] 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.

[27] Daniel Johnson, Peta Wyeth, Madison Clark, and Christopher Watling. 2015. Cooperative Game Play with Avatars and Agents: Differences in Brain Activity and the Experience of Play. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3721–3730.

[28] Gareth R. Jones and Jennifer M. George. 1998. The Experience and Evolution of Trust: Implications for Cooperation and Teamwork. *The Academy of Management Review* 23, 3 (July 1998).

[29] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[30] S. Komiak, Weiquan Wang, and I. Benbasat. 2005. Comparing Customer Trust in Virtual Salespersons With Customer Trust in Human Salespersons. In *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*. 175a–175a.

[31] Sherrie Y. X. Komiak and Izak Benbasat. 2006. The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly* 30, 4 (December 2006), 941–960.

[32] Sohye Lim and Byron Reeves. 2010. Computer agents vs avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies* 68, 1–2 (September 2010), 57–68.

[33] Jr. Lount, R. B. 2010. The impact of positive mood on trust in interpersonal and intergroup interactions. *Journal of Personality and Social Psychology* 98, 3 (2010), 420–433.

[34] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch. 2012. Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making* 6 (January 2012), 57–87.

[35] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (July 1995), 709–734.

[36] Daniel J. McAllister. 1995. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal* 38, 1 (February 1995), 24–59.

[37] Kevin McCabe, Daniel Houser, Lee Ryan, Vernon Smith, and Theodore Trouard. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* 98, 20 (September 2001), 11832–11835.

[38] Stephanie M. Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I Trust It, but I Dont Know Why Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 55, 3 (June 2013), 520–534.

[39] Tim Robert Merritt, Kian Boon Tan, Christopher Ong, Aswin Thomas Abraham, Teong Leong Chuah, and Kevin McGee. 2011. Are artificial team-mates scapegoats in computer games?. In *Proceedings of the 2011 ACM Conference on Computer Supported Cooperative Work, CSCW*. ACM, 685–688.

[40] Clifford Nass, B.J. Fogg, and Youngme Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6 (April 1996), 669–678.

[41] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 72–78.

[42] Florian Nothdurft, Tobias Heinroth, and Wolfgang Minker. 2013. The Impact of Explanation Dialogues on Human-Computer Trust. In *Human-Computer Interaction. Users and Contexts of Use*, Masaaki Kurosu (Ed.). Springer Berlin Heidelberg, 59–67.

[43] Christopher Ong, Kevin McGee, and Teong Leong Chuah. 2012. Closing the human-AI Team-mate Gap: How Changes to Displayed Information Impact Player Behavior Towards Computer Teammates. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. ACM, New York, NY, USA, 433–439.

[44] R. Pak, N. Fink, M. Price, B. Bass, and L. Sturre. 2012. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics* 55, 9 (July 2012), 1059–1072.

[45] Noam Peled, Ya'akov (Kobi) Gal, and Sarit Kraus. 2013. An Agent Design for Repeated Negotiation and Information Revelation with People. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 789–795.

[46] Paul Robinette, Alan R. Wagner, and Ayanna M. Howard. 2013. Building and Maintaining Trust Between Humans and Guidance Robots in an Emergency. In *AAAI Spring Symposium: Trust and Autonomous Systems*. Stanford, CA, 78–83.

[47] Sandip Sen. 2013. A Comprehensive Approach to Trust Management. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems*. Saint Paul, Minnesota, USA, 797–800.

[48] Daniel B. Shank. 2012. Perceived Justice and Reactions to Coercive Computers. *Sociological Forum* 27, 2 (June 2012), 372–391.

[49] C.K. Stokes, J.B. Lyons, K. Littlejohn, J. Natarian, E. Case, and N. Speranza. 2010. Accounting for the human in cyberspace: Effects of mood on trust in automation. In *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*. 180–187.

[50] Piotr Sztompka. 1999. *Trust: A Sociological Theory*. Cambridge University Press, Cambridge, UK.

[51] A. van Wissen, Y. Gal, B.A. Kamphorst, and M. V. Dignum. 2012. Human-agent teamwork in dynamic environments. *Computers in Human Behavior* 28, 1 (2012), 23–33.

[52] Arlette van Wissen, Jurriaan van Diggelen, and Virginia Dignum. 2009. The Effects of Cooperative Agent Behavior on Human Cooperativeness. In *Proceedings of eighth International Confefence on Autonomous Agents and Multiagent Systems (AAMAS 2009)*. 1179–1180.

[53] Frank Verberne, Jaap Ham, and Cees J. H. Midden. 2012. Trust in Smart Systems: Sharing Driving Goals and Giving Information to Increase Trustworthiness and Acceptability of Smart Systems in Cars. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 54, 5 (May 2012), 799–810.

[54] Frank M. F. Verberne, Jaap Ham, and Cees J. H. Midden. 2014. Familiar faces: Trust in a facially similar agent. In *HAIDM Workshop at the Autonomous Agents and Multiagent Systems (AAMAS'14)*.

[55] Astrid M. von der P´utten, Nicole C. Kr´amer, Jonathan Gratch, and Sin-Hwa Kang. 2010. "It Doesn't Matter What You Are!" Explaining Social Effects of Agents and Avatars. *Computers in Human Behavior* 26, 6 (July 2010), 1641–1650.

[56] Robert A. Wagner and Michael J. Fischer. 1974. The String-to-String Correction Problem. *J. ACM* 21, 1 (January 1974), 168–173.