# Reputation Based Trust in Human-Agent Teamwork Without Explicit Coordination

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

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

## ABSTRACT

Interacting with strangers and agents through computer networks has become a routine aspect of our daily lives. In such environments, reputation plays a critical role in determining our future interactions and satisfaction derived from them. This paper empirically investigates the effects of agent reputation on human *trust in* and *behavior towards* "peer" level agent teammates over repeated interactions. We developed a team coordination game, the Game of Trust, in which a human player and an agent player repeatedly cooperate to complete team tasks without prior assignment of subtasks. Before the game begins, the agent player is introduced with either positive or negative reputation to the human player. The effects of agent reputation are evaluated by performing an extensive set of controlled experiments with participants recruited from Amazon Mechanical Turk, a crowdsourcing marketplace. We collect both teamwork performance data as well as surveys to gauge participants' trust in their agent teammates. The empirical results show that positive (negative) agent reputation led to greater (lower) human trust in agent teammates. Moreover, the interplay between the game expertise and expectation from agent teammate significantly affected the influence of reputation. These findings enhance our understanding of changes in human trust with respect to agent reputation towards achieving successful human-agent teamwork.

## CCS CONCEPTS

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

## KEYWORDS

Human-agent teamwork, reputation, trust

**ACM Reference Format:**
Feyza Merve Hafızoğlu and Sandip Sen. 2018. Reputation Based Trust in Human-Agent Teamwork Without Explicit Coordination. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3284432.3284454

## 1 INTRODUCTION

The importance of reputation-based decision-making has been increasing with the prominence of web-based connectivity among people and agents in diverse locations without clearly defined organizational or social structure. Reputation not only plays a key role in making decisions [2, 12, 32] but the presence of a reputation mechanism also encourages and sustains cooperative behavior [5, 6, 12, 25]. Under the circumstances, where numerous agents are deployed to perform tasks with people, e.g., e-commerce, crowdworking systems, reputation has become an invaluable source of information for choosing the right agent partner.

The role of trust is not limited to human interactions: trust also shapes the way people engage with technology. Therefore, establishing people's trust is a key cornerstone of fluid interactions between humans and agents. Reputation has been an important mechanism to build trust between people and organizations. Previous studies [21] demonstrate that trust can be biased by introducing reputation of a system. People may benefit from the reputation provided by third parties for making decisions of future agent partners. For utilizing the reputation mechanisms towards achieving successful human-agent teamwork, we must have a clear understanding how human trust in agent teammates is influenced by agent reputation.

This study aims to better understand the impact of agent reputation on human trust in agent teammates by developing and experimenting with a virtual repeated human-agent teamwork framework. By *virtual* human-agent teamwork we refer to domains where autonomous agents and humans work over a network without any physical embodiment of the agents, either in the form of robots or avatars. Besides agent reputation, we consider the human trust behavior based only on the agents' task performance or contribution towards achieving team goals over repeated interactions. 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.

In such virtual human-agent teamwork domains, human trust attitudes will be influenced by a variety of factors, including agent reliability, prior experience(s) of the humans, and agent reputation. In this study, we investigate the effect of *agent reputation* to address the gap in reputation-based human trust in agent teammates. Whether reputation-based trust can contribute to trust building in agent teammates is a critical question for designers of agent technology. On the whole, how can people be biased when they are provided agent reputation? Can positive reputation promote human trust in agents? In contrast, can negative reputation hinder trust development in human-agent teamwork?

*HAI '18, December 15–18, 2018, Southampton, United Kingdom*
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284454

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 concludes the paper with a summary and the directions for future research.

## 2 RELATED WORK

The importance of trust in human-agent interactions has been well-acknowledged [9, 11, 14, 33]. 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). Several studies have documented the effects of human factors, such as age [4], personality [10], culture [16], mood [31], attitude [22], and past experience [4, 21].

Among the agent factors, agent behavior is fundamental in building trust in agent teammates. Positive behavior, such as cooperativeness [34] and reliability [11], improves trust and facilitates the collaboration between parties. In contrast, negative behavior, such as defection [33] and deception [34], leads to reduced trust and, hence, less willingness to collaborate in the future. Communication skills of agents play a significant role in maintaining the trust relationship [26]. Furthermore, familiarity and personalization of agents have been shown to positively influence human trust [36].

Recently, researchers have examined the effects of different agent representations, such as avatars and robots [8, 36], and the effects of external factors, such as information representation [2] and reputation. Previous research demonstrated that positive reputation led to greater trust [19, 30] and the direct experience reduced the influence of reputation [13].

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 negotiation [33], or giving advice [9, 31]. Research on teamwork without explicit coordination and pre-planing, e.g., crowd-work [18], massively multiplayer online games, is significantly limited. This kind of human-agent teamwork has been rarely investigated from the aspect of factors affecting human trust, such as blame behavior for team failures [24]. Our research extends these studies as follows: we consider teams of human and agent rather than mere interactions between two players, e.g., [8, 34, 36]; we focus on teamwork environments in which there is neither explicit communication between human and agent in contrast to those in [33] nor the embodiment of agents in contrast to those in [8, 36]; we explore repeated, rather than one-shot, interactions of fixed teams, rather than dynamic teams [33]; the domain provides real team tasks for evaluating human-agent teamwork rather than the standard synthetic environments [8, 33–35]. To the best of our knowledge, this is the first study on agent reputation affecting trust in human-agent teamwork within a repeated virtual team game scenario where agents are autonomous, peer level team members and without prior coordination.

## 3 HUMAN-AGENT TEAMWORK MODEL

Our goal is to understand and characterize the human trust development in agent teammates over initial repeated interactions, but without any prior experience of interaction 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 emergency situations, 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 towards 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 agent teammate reduces the uncertainty over that agent's independent actions which positively correlates with the truster's utility towards achieving team goals*[29]. According to this interpretation, human trust in agent teammate can both reduce uncertainty about agent's contribution and improve team performance through more effective agent contribution and better team coordination.

### 3.1 The Game of Trust

*The Game of Trust (GoT)* is a team game in which two players form a team and have *n* 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, may be to recruit a given number of volunteers or to collect a number of specimens that fit a given description.

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.

We use the term *effort level*, the portion (percentage) of the total work units completed by this team member, as a standard metric that can be used to compare player performances over interactions. If we denote the amount of work units accomplished by human player in the $i^{th}$ interaction by $w_h^i$ and the team task size in this interaction by $|t_i|$, then the effort level of a human player in the $i^{th}$ interaction is $\frac{w_h^i}{|t_i|} \epsilon [0, 1)$.

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 own individual payment accounts which have an initial balance, $b_{init}$, at the beginning of the game. 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, however, the combined number of subtasks accomplished by the players is equal to or greater than the size of the team task, it means the players successfully completed the team task. Then 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 players accomplished is less than the team task, no reward is given.

By *utility of a player* we refer to half of the team reward, if the team completed the task, minus the cost of performing subtasks individually. If they cannot achieve the team task, both players 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 team members 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 to avoid, to the extent feasible, the possibility of participants having additional motivations. Based on these considerations, we chose an audio transcription domain for the human-agent teamwork goal instances. Hence, in this domain, the *task* that is assigned to the team corresponds to transcribing a number of English words from audio to text and the *atomic subtask* corresponds to transcribing one word. We will use the term *task size* to refer to the number of words to transcribe, i.e., number of subtasks, in an interaction.

Though we have no interest in the transcribed words, the participants 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[1] over the length of the transcribed word. This is done to ensure a minimum quality of participant 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 in a sequence, where each game consists of several 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 the risk of not achieving the team goal. Therefore, the number of words they need to transcribe is a critical decision that they have to make in each interaction and is based on their trust in the teammate for contributing to the team task.

# 4 EMPIRICAL METHODOLOGY

We attempt to investigate the relationship between agent reputation and resulting trust in agent teammates. A number of studies have shown that people are inclined to trust a system more when it is introduced as a reputable system [2, 17, 19, 30]. However, this effect may gradually decrease as the experience of the trustor with the trustee increases. Therefore, we expect agent reputation to affect the interactions between humans and agents. In particular, positive reputation may contribute to building trust in agent teammates, while negative reputation may cause human trust to deteriorate. For instance, trust in an agent teammate can be fostered if the agent is introduced as fair, trustworthy, and capable. This kind of reputation allows people to build trust more effectively because the uncertainty about the agent's attitude can be reduced to some extent.

Hypothesis 1. *Participants' trust in agent teammate can be increased when positive agent reputation is provided.*

Hypothesis 2. *Participants' trust in agent teammate can be reduced when negative agent reputation is provided.*

## 4.1 Experimental Setup

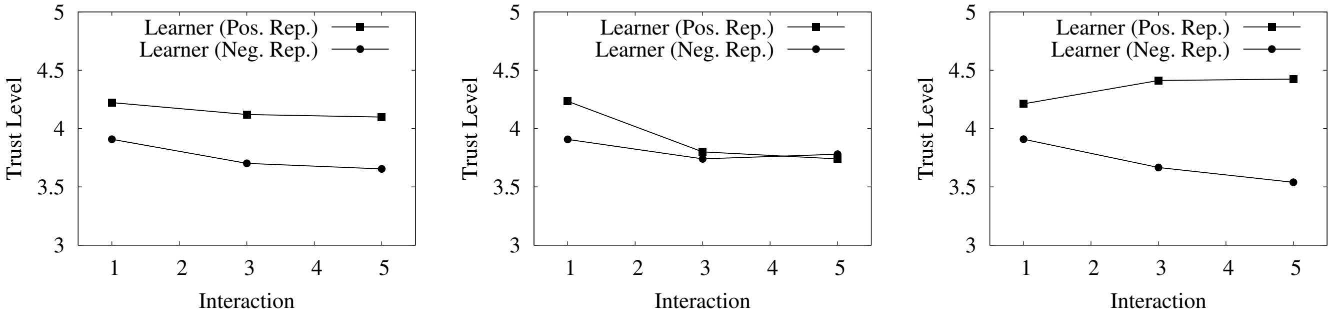
**Game Configuration:** The number of interactions in a game is five (as in [3, 7, 27]), which is a short enough duration to avoid participants becoming bored and yet still allows team members to adapt to teammates with predictable behavior. The size of team task is incremented by two in each interaction, i.e., the sequence of task sizes is $\langle 6, 8, 10, 12, 14\rangle$.

Both the human player and the agent have their own artificial account with the initial balance set to 45, which should be sufficient to perform all the tasks in the sequence. The cost and reward per work unit are set to 1 and 1.75, respectively. The players are allowed to choose a task size between one and the size of team task minus one.

**Reputation:** We consider two conditions of reputation: *positive* reputation and *negative* reputation. In the GoT framework, the participants were provided positive (negative) agent reputation with the instruction: *"This automated computer player is known to be a trustworthy (untrustworthy) teammate"* prior to the game. The reputation was presented anonymously rather than pointing out a reputation source to avoid any biases.

**Agent Teammate:** The *Learner agent* is trained offline to predict the human player's task choice by utilizing the linear regression method with the data collected from the teamwork experiences of humans in our prior work [15]. It completes half of the team task in the first interaction. In the subsequent interactions, it makes a prediction of the teammate's task choice given the prior interactions in the game. Based on this prediction, it determines its own individual task choice to complete the remainder of the task to achieve the team goal optimally, without redundancy or falling short of the team task.

[1] http://en.wikipedia.org/wiki/Wagner-Fischer_algorithm



(a) Overall

(b) G1

(c) G2

**Figure 1: Trust in the Learner Agent for positive and negative reputations**

Given the irrational, unpredictable, and "noisy" behavior of the human players, it is a challenge to develop a learning agent teammate that can produce optimal social utility over repeated interactions. This is particularly true given that any adaptation by agents can elicit responsive adaptation by the human, which significantly complicates the task of the agent learner. This "moving target" learning problem is well-recognized in the multi-agent learning literature [28]. The current situation is, if anything, of even greater challenge because of the very different biases, knowledge, cognitive load, and expectations of the human and agent players.

The Learner agent in our experiments only selects the number of words to transcribe, but actually does not transcribe any words, yet the human players are told that it does so. Additionally, we assume that the Learner agent transcribes all words accurately. The reason for this decision is that we are only interested to see how the agent's choice of number of subtasks to complete affects human trust attitude and behavior.

**Experiment:** We designed the experiment to investigate how negative and positive reputation of agent teammates affect the perception of participants and the interactions. The experiment consists of two games with the Learner agent. However, the participants were instructed that they will play with a new automated computer player in each game. The Learner agent was introduced as a trustworthy teammate (positive reputation) in one game and as an untrustworthy (negative reputation) player in the other game. In fact, neither positive nor negative reputation reflects the actual trustworthiness of the Learner agent. We intended to bias the participants and then observe how positive or negative reputation biases influence the interactions. A final note is that the participants were debriefed about the deception on fictitious reputation after completing the study.

To neutralize the influence of the order, we experimented with two groups of participants based on the order of reputation, where *G1* (*G2*) was introduced positive (negative) reputation in the first game and vice versa in the second game.

**G1:** Learner Agent (Pos. Rep.), Learner Agent (Neg. Rep.);
**G2:** Learner Agent (Neg. Rep.), Learner Agent (Pos. Rep.).

**Survey:** The game includes a short survey on trust that is adopted for measuring human players' perceived trustworthiness and fairness of their teammates. Participants were asked to complete this survey after the first, third, and fifth interactions of a game (similar to [31]) after they were shown the outcome of the most recent teamwork. This short questionnaire, adapted from [1], 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.

The *trust level* of a participant in agent teammate is computed as the average of the responses to these three items.

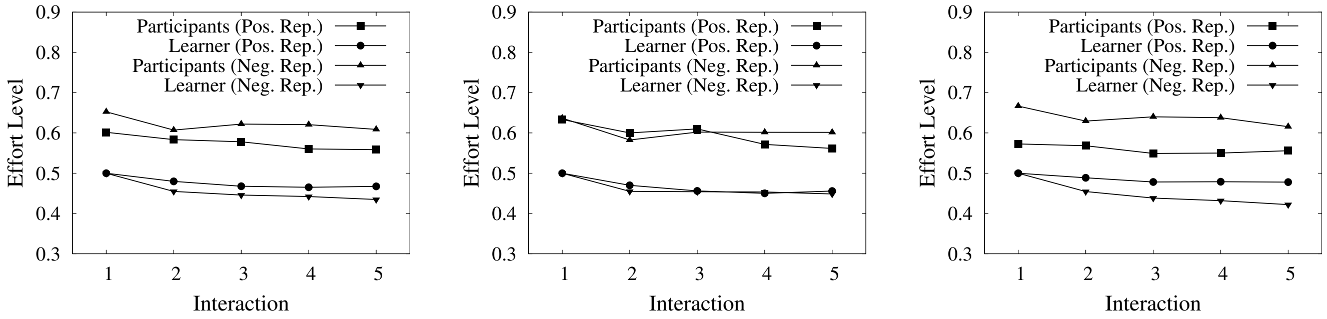
**Participants:** We recruited 115 participants through Amazon Mechanical Turk and the study data of 10 participants is eliminated due to insufficient attention based on the consistency criteria. We analyzed the data collected from 105 participants. There were 50 and 55 participants in groups G1 and G2, respectively. Approximately 35% of the participants were female. Age distribution was as follows: 18-24 years, 12%; 25-34 years, 49%; 35-44 years, 20%; 45-54 years, 13%; and 55-64 years, 6%. The distribution based on education level was as follows: high school degree, 9%; some college experience, 24%; associate's degree, 14%; bachelor's degree, 42%; and graduate degree, 11%. The ethnicity distribution was as follows: White, 80%; Hispanic-Latino, 6%; African-American, 7%; Asian, 5%; and other ethnicities, 1%.

## 5 RESULTS

This section presents the participants' trust levels in agent teammates, effort levels by the participants and the Learner agent, and the cumulative team performance with respect to positive and negative reputation.

### 5.1 Trust Analysis

Figure 1 presents the participants' trust in the Learner agent. In Figure 1(a), trust in the agent teammate slightly decreases over the game in both conditions. This decline can be attributed to the effort levels of the Learner agent that decrease over interactions. This is because of adaptive nature of the Learner agent, which delivers less than half of the team task when the participant completes more



(a) Overall

(b) G1

(c) G2

**Figure 2: Effort levels for positive and negative reputations**

than half of the team task in previous interactions (see Figure 2). When comparing the two conditions, positive reputation led to significantly greater trust compared to the negative reputation after the first ($F(1, 208) = 11.98,\ p < 0.001$), third ($F(1, 208) = 8.69,\ p < 0.01$), and fifth ($F(1, 208) = 6.99,\ p < 0.01$) interactions.

In G1 ( Figure 1(b)), trust in the Learner agent decreases significantly over interactions ($F(2, 147) = 3.59,\ p < 0.05$) in the positive reputation condition, i.e., the first game (within condition). Furthermore, the Learner agent is trusted significantly more in the positive reputation condition ($M = 4.23,\ SD = 0.53$) compared to the negative reputation condition ($M = 3.91,\ SD = 0.72$) initially ($F(1, 98) = 6.70,\ p < 0.05$) (between condition).

In G2 ( Figure 1(c)), trust in the Learner agent slightly increased (decreased) over interactions for positive (negative) reputation (within condition). The positive reputation condition led to significantly greater trust compared to trust in the negative reputation condition after the first ($F(1, 108) = 5.35,\ p < 0.05$), third ($F(1, 108) = 17.52,\ p < 0.001$ ), and fifth ($F(1, 108) = 17.03,\ p < 0.001$) interactions (between condition).

Interestingly, a significant decrease in trust in the game with positive reputation agent is observed when this was the first game of the experiment (G1) (see Figure 1(b)), while only a slight increase is observed if the positive reputation was provided in the second game (G2) (see Figure 1(c)). This contrasting behavior of trust development between the two groups highlights the importance of past experience in that it significantly affected participants' trust and the influence of reputation on the subsequent interactions.

## 5.2 Effort Level Analysis

Figure 2 presents the variation in effort levels by the participants and the Learner agent over interactions.

Figure 2(a) illustrates that the participants' initial effort level was significantly higher ($F(1, 208) = 6.14,\ p < 0.05$) for the negative reputation compared to the corresponding effort levels in the positive reputation (between condition). This difference demonstrates their caution as a result of the negative reputation received. Subsequently, effort levels in the negative reputation condition continued to be significantly higher compared to the positive reputation condition in the third ($F(1, 208) = 5.04,\ p < 0.05$), fourth ($F(1, 208) = 9.04,\ p < 0.01$), and fifth ($F(1, 208) = 6.63,\ p < 0.05$) interactions.

The Learner agent adjusted its effort levels by significantly decreasing with both positive ($F(4, 520) = 4.67,\ p < 0.01$) and negative reputation ($F(4, 520) = 11.05,\ p < 0.001$) over the course of the game (within condition). Between the two conditions, the Learner agent put significantly less effort in the negative reputation condition compared to its efforts in the positive reputation condition in second ($F(1, 208) = 6.99,\ p < 0.01$), third ($F(1, 208) = 4.55,\ p < 0.05$), fourth ($F(1, 208) = 3.59,\ p < 0.1$), and fifth ($F(1, 208) = 5.59,\ p < 0.05$) interactions. Since the Learner agent adapts to its teammate, it put less effort in the negative reputation condition due to the fact that the participants put significantly greater effort.

In G1 (Figure 2(b)), the variation in effort levels by the participants is not significant in either conditions. Additionally, there are no significant differences in effort levels between the two conditions throughout the game. Effort levels by the Learner agent significantly declined in the positive ($F(4, 245) = 3.09,\ p < 0.05$) and negative reputation ($F(4, 245) = 3.22,\ p < 0.05$) conditions (within condition).

In G2 (Figure 2(c)), effort levels by the participants do not change significantly over interactions in both games. Between condition, negative reputation led to significantly greater effort by the participants compared to effort levels in for the positive reputation condition in the first ($F(1, 108) = 11.86,\ p < 0.001$), second ($F(1, 108) = 7.82,\ p < 0.01$), third ($F(1, 108) = 14.7,\ p < 0.001$), fourth ($F(1, 108) = 14.14,\ p < 0.001$), and fifth ($F(1, 108) = 7.15,\ p < 0.01$) interactions. Additionally, the Learner agent's effort levels decreased significantly in the negative reputation condition ($F(4, 270) = 8.97,\ p < 0.001$) but not in the positive reputation condition over the game (within condition). Between the two conditions, the Learner agent's effort level was significantly higher in the positive reputation condition compared to the negative reputation condition in the second ($F(1, 108) = 9.49,\ p < 0.01$), third ($F(1, 108) = 8.43,\ p < 0.01$), fourth ($F(1, 108) = 9.97,\ p < 0.01$), and fifth ($F(1, 108) = 11.62,\ p < 0.001$) interactions.

**Table 1: Results of the games for positive and negative reputations**

| | Positive Rep. | Negative Rep. |
|---|---|---|
| Goals Achieved | 4.69 ± 0.67 | 4.68 ± 0.63 |
| Words Transcribed | 46.74 ± 7.43 | 46.51 ± 7.02 |
| Redundancy | 2.84 ± 2.96 | 3.98 ± 3.13 |
| Participant Utility | 12.18 ± 8.39 | 9.59 ± 8.40 |
| Agent Utility | 17.26 ± 6.96 | 18.21 ± 7.05 |
| Social Utility | 29.44 ± 12.87 | 27.80 ± 12.31 |

### 5.3 Performance Analysis

Table 1 provides the cumulative game results: the number of team goals achieved (successful interaction), total number of words transcribed by the team, the number of words transcribed redundantly (the number of excess unit tasks performed), and participant/agent/social utilities in a game. A significant difference is observed in redundancy ($F(1, 208) = 7.41,\ p < 0.01$) between the two conditions. In particular, participant utility was significantly higher when positive reputation was provided ($F(1, 208) = 5.00,\ p < 0.05$). Likewise, agent utility was slightly higher when negative reputation. Finally, social utility was slightly higher in the positive reputation condition as a result of the higher participant utility. On the other hand, no significant difference is observed in the number of team goals achieved and the number of words transcribed between the two reputation conditions.

## 6 DISCUSSION

*Trust in Agent Teammates:* Our objective was to understand the relation between agent reputation and resulting trust in agent teammate. Specifically, we examine whether trust increases (decreases) when the agent teammate has a reputation of being a (an) trustworthy (untrustworthy) teammate. Our results indicate that reputation draws significant differences in the participants' perception of their agent teammate's trustworthiness. That is, positive reputation led to significantly greater trust compared to negative reputation (Figures 1(a) and 1(c)) and the difference in trust between the two conditions remained significant over the course the game, i.e., the impact of reputation on trust did not disappear, supporting **Hypothesis 1** and **Hypothesis 2** as well as the results from previous studies [20].

However, trust in the Learner agent in the positive reputation condition differed significantly between the groups G1 and G2. In particular, trust in the Learner agent that was introduced with positive reputation declined significantly when the game played was the first game (G1). In fact, this is the case causing the overall trust to decrease slightly (see Figure 1(a)). We posit that this unanticipated result is due to lack of experience in GoT and expectations which is discussed below.

In the positive reputation condition, the interplay between lack of experience in the GoT and high expectations, due to positive reputation, led to significant decrease in trust (Figure 1(b)). Lack of experience in the GoT led to higher initial efforts, i.e., the participants had tendency to contribute more at the beginning of the first game compared to the second game (Figure 2(b) and 2(c)). The participants in G1 had greater expectations from the Learner agent introduced as a trustworthy teammate. In turn to high initial effort by the participants, the Learner agent put relatively less effort and did not fulfill expectations in the subsequent interactions, hence participants' trust reduced significantly as a result of their high expectations. The effort level by the participants in G1 is initially indistinguishable between the positive and negative reputation conditions, and this difference remains insignificant throughout the game. Therefore, similar effort levels by the Learner agent in the two conditions (Figure 2(b)) led to similar trust scores (Figure 1(b)). These results also suggest that teammate's contribution is more influential on trust than reputation in case of adaptive partners. This significant decrease in trust is in accordance with Merritt and Ilgen's [23] findings suggesting that the greater the sense of violation, the greater damage to subsequent trust.

In G2, on the contrary, initial effort by the participants was significantly less for the positive reputation condition compared to the negative reputation condition. This difference remained significant throughout the game. Thus, the Learner agent put significantly greater effort in the positive reputation condition as the participants began playing with significantly less effort compared to negative reputation condition. Therefore, notably higher effort levels by the Learner agent (Figure 2(c)) were rated as significantly more trustworthy by the participants (Figure 1(c)) compared to the trust in the negative reputation condition.

*Effort Level Distribution:* The adaptive nature of both participants and the Learner agent, i.e., they adapt their effort levels to reduce the amount of redundant work, plays an influential role on the effort levels and the resulting trust. This is a multi-agent learning process in which each player estimate the trustworthiness of their teammate and adjust their behavior in each interaction, while their teammate also attempts to predict their task choices and determines their task choices accordingly.

Lack of experience in GoT and uncertainty about the agent teammate made the participants cautious and hence begin the first game by putting greater effort. In the first interaction, delivering more than half of the team task regardless of reputation type demonstrates the participants' cautious attitude towards achieving the team goal. Given that the initial interactions are the basis of the trust relationship and future effort level distribution between teammates[2], this uncertainty at the beginning of the first game becomes more critical and influential on teamwork. Reputation, in general, reduces uncertainty about the trustworthiness of teammate and, hence, positive (negative) reputation leads to less (more) effort. Our results are in agreement with this general trend. The participants exert maximum amount of effort when negative reputation is provided in the first game. On the other hand, minimum amount of effort is observed when positive reputation is provided in the second game (Figure 2(c)).

Positive reputation encouraged the participants to rely on their agent teammates towards achieving team goals, i.e., positive agent reputation led the participants to reduced effort levels compared to the no and negative reputation conditions. Subsequently, it led

[2] The impact of initial interactions is of utmost importance when the players are adaptive because the initial choices are the seeds and the process of learning teammate behavior builds on it.

the Learner agent to put significantly greater effort in the positive reputation condition relative to negative reputation due to the complementary nature of the effort levels of the Learner agent and the participants. The empirical results reveal significant differences in effort levels between the *positive* and *negative* reputation conditions (Figures 2(a) and 2(c)). The only exception to this trend emerged when the positive reputation was provided in the first game (Figure 2(b)). As discussed above, lack of experience in GoT and higher expectations canceled out the contributions of positive reputation in relying on agent teammates.

Another aspect of positive reputation is that it lead to a reduction in the variation in effort levels by the participants (Figure 2(c)) and the Learner agent. The underlying reason is that without reputation, the participants started playing with unnecessarily high effort levels. In the subsequent interactions, the Learner agent reduced its effort significantly as a result of participants' initial great effort. Furthermore, participants, too, were likely to reduce their effort either slightly, e.g., negative reputation condition (see Figures 2(a) and 2(c)), as long as there was redundant work in the previous interaction.

The results indicate that opposite changes occur in the negative reputation condition. The participants put significantly greater effort compared to the positive reputation condition (Figures 2(a) and 2(c)). On the other hand, the Learner agent put significantly less effort over the course of the game except the first interaction. Furthermore, the Learner agent decreased its effort levels significantly with each interaction in the negative reputation condition (see Figures 2). However this is not always the case in the positive reputation condition (see Figures 2(c)). Our findings suggest that agent reputation significantly affects the effort levels in different games as well as the variation in the effort levels over the course of the game.

*Team Performance:* The results indicate that positive reputation led to a decrease in redundancy in team efforts and hence an increase in social utility by reducing the cost of redundant work. Given the cautious nature of participants, the average number of achieved team goals was high for all reputation conditions. In the GoT framework, loss of utility due to redundancy is lower compared to loss from failure to achieve the team goal. For instance, missing one word out of ten words results with social utility −9, while transcribing one extra word results with social utility 6.5. Given the fact that maximum social utility is 7.5, redundancy is not as unpleasant as the failure to complete the team tasks in terms of utilities. This explains why the differences in social utilities are not significant.

## 7 CONCLUSION

This study is an empirical investigation of the growth of human trust in human-agent teamwork in virtual environments without explicit communication based on agent reputation. The novel aspect of this study that distinguishes it from previous work is that human and agent have the same level of autonomy in a team. We introduced a formal team game, the Game of Trust, and its use for studying human trust over repeated interactions without explicit coordination. Positive/negative reputation of the agent player is provided at the beginning of the game to bias the participants.

Empirical findings show that positive reputation reduces the uncertainty about and the reliance on the agent teammate. Hence, positive reputation led to significantly greater trust compared to negative reputation. The impact of positive reputation did not disappear throughout the game. Furthermore, positive reputation led to significantly lower effort levels, i.e., more reliance on agent teammate, compared to the negative reputation. The only exception to the trend above occurred when the positive reputation condition was the first game where the effect of positive reputation on leading to increased reliance on the agent teammate was canceled out by the initial uncertainty in the GoT environment, i.e., lack of experience led participants to be skeptical. Regarding team performance, positive reputation led to significantly reduced redundancy in teamwork.

Our highest priority for future research is to study the human-agent teamwork with complex tasks in ad-hoc scenarios. Such complex tasks comprise of subtasks that require different abilities as is experienced in many real-life teamwork scenarios. Furthermore, some of the subtasks may be dependent on others. The ad-hoc scenarios are particularly challenging and interesting because humans and agents neither know each other's abilities regarding different task types nor the alignment of their own abilities and teammate's abilities.

## ACKNOWLEDGMENTS

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

## REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Riccardo Boero, Giangiacomo Bravo, Marco Castellani, and Flaminio Squazzoni. 2009. Reputational cues in repeated trust games. *The Journal of Socio-Economics* 38, 6 (2009), 871–877.

[6] Iris Bohnet and Steffen Huck. 2004. Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change. *The American Economic Review* 94, 2 (May 2004), 362–366.

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

[8] 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.

[9] 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.

[10] Hongying Du and M.N. Huhns. 2013. Determining the Effect of Personality Types on Human-Agent Interactions. In *Joint Conferences (WI) and (IAT).* 239–244.

[11] 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.

[12] Feng Fu, Christoph Hauert, Martin A. Nowak, and Long Wang. 2008. Reputation-based partner choice promotes cooperation in social networks. *Phys. Rev. E* 78, 2 (August 2008), 026117.

[13] 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.

[14] 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.

[15] Feyza Merve Hafizoglu and Sandip Sen. 2015. Evaluating trust levels in human-agent teamwork in virtual environments. In *HAIDM Workshop at the Autonomous Agents and Multiagent Systems (AAMAS'15)*. 1–16.

[16] Galit Haim, Yaakov Kobi Gal, Sarit Kraus, and Michele Gelfand. 2012. A Cultural Sensitive Agent for Human-Computer Negotiation. In *AAMAS*. Valencia, Spain, 451–458.

[17] JinBaek Kim. 2012. An empirical study on consumer first purchase intention in online shopping: integrating initial trust and TAM. *Electronic Commerce Research* 12, 2 (2012), 125–150.

[18] 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.

[19] F. Javier Lerch, Michael J. Prietula, and Carol T. Kulik. 1997. Expertise in Context. MIT Press, Cambridge, MA, USA, Chapter The Turing Effect: The Nature of Trust in Expert Systems Advice, 417–448.

[20] Poornima Madhavan and Douglas A. Wiegmann. 2007. Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49, 5 (2007), 773–785.

[21] 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.

[22] 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.

[23] Stephanie M. Merritt and Daniel R. Ilgen. 2008. Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 2 (April 2008), 194–210.

[24] 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 *CSCW*. ACM, 685–688.

[25] Manfred Milinski, Dirk Semmann, and Hans-Jurgen Krambeck. 2002. Reputation helps solve the /'tragedy of the commons/'. *Nature* 415 (January 2002), 424–426.

[26] 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.

[27] 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.

[28] Liviu Panait and Sean Luke. 2005. Cooperative Multi-Agent Learning: The State of the Art. *Autonomous Agents and Multi-Agent Systems* 11, 3 (November 2005), 387–434.

[29] 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.

[30] Randall D Spain. 2009. *The Effects of Automation Expertise, System Confidence, and Image Quality on Trust, Compliance, and Performance*. Ph.D. Dissertation. Old Dominion University, Norfolk, VA.

[31] 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.

[32] Claudio Tennie, Uta Frith, and Chris D. Frith. 2010. Reputation management in the age of the world-wide web. *Trends in Cognitive Sciences* 14, 11 (2010), 482–488.

[33] 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.

[34] Arlette van Wissen, Jurriaan van Diggelen, and Virginia Dignum. 2009. The Effects of Cooperative Agent Behavior on Human Cooperativeness. In *AAMAS*. 1179–1180.

[35] 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.

[36] Frank M. F. Verberne, Jaap Ham, and Cees J. H. Midden. 2014. Familiar faces: Trust in a facially similar agent. In *HAIDM Workshop at AAMAS*.