Evaluating Trust Levels in Human-agent Teamwork in Virtual Environments

Feyza Merve Hafızoğlu and Sandip Sen

Tandy School of Computer Science
University of Tulsa, Tulsa OK 74104, USA
{feyza-hafizoglu,sandip}@utulsa.edu

Abstract. With the improvement in agent technology and agent capabilities we foresee increasing use of agents in social contexts and, in particular, in human-agent team applications. To be effective in such team contexts, agents need to understand and adapt to the expectation of human team members. This paper presents our study on how behavioral strategies of agents affect the humans' trust in those agents and the concomitant performance expectations that follow in virtual team environments. We have developed a virtual teamwork problem that involves repeated interaction between a human and several agent types over multiple episodes. The domain involves transcribing spoken words, and was chosen so that no specialized knowledge beyond language expertise is required of the human participants. The problem requires humans and agents to independently choose subset of tasks to complete without consulting with the partner and utility obtained is a function of the payment for task, if completed, minus its efforts. We implemented several agents types, which vary in how much of the teamwork they perform over different interactions in an episode. Experiments were conducted with subjects recruited from the MTurk. We collected both teamwork performance data as well as surveys to gauge participants' trust in their agent partners. We trained a regression model on collected game data to identify distinct behavioral traits. By integrating the prediction model of player's task choice, a learning agent is constructed and shown to significantly improve both social welfare, by reducing redundant work without sacrificing task completion rate, as well as agent and human utilities.

Keywords: human-agent interaction, teamwork, trust, adaptation

1 Introduction

Researchers working on agent mechanisms and technologies are interested in developing tools and techniques that will enable the development and successful deployment of autonomous agent applications in high-impact and socially relevant domains. A particular class of such domains, that require collaborative partnerships and teamwork between humans and autonomous agents have received increasing attention. While standalone, independent agent applications

can deliver significant benefits in domains such as security enforcement, information collection, error correction, etc., we believe agent technology will receive much-needed recognition and appreciation with success in applications requiring active human-agent collaboration [3, 12, 13]. In particular, we need to develop agent technology that will enable developing agent applications in domains where humans recognize agents as *autonomous* and *key* partners and have to rely on agents as peer team-members.

One key aspect of teamwork is the necessity of team members to trust peers which allow them to rely on teammates as reliable partners with predictable and effective behaviors towards achieving team goals. In this study, we are interested in understanding the development of human's initial trust in agent teammates in virtual human-agent teamwork. 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. We consider the human's trust behavior based only on the agents' task performance or contribution towards achieving team goals over repeated interactions.

Our motivation stems from either of the following scenarios: (a) the human is new to a domain and has to rely on more experienced agent partners until she develops the necessary competency from her experiences, (b) the human is familiar with the domain but will need to work with autonomous teammates with whom they have had no prior collaboration experience. Such domains include ad-hoc teamwork scenarios, where unfamiliar individuals have to cooperate with new partners, which 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.

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 might not be possible to optimally allocate dynamically arriving tasks among team members. Rather, the team must be responsive to the emerging situations which can be achieved by team members adapting their behaviors and efforts based on expectations of contribution of team members. It is then critical to address the following questions:

- How does human's trust in automated, virtual partners change in their first few interactions with that partner?
- What is the associated behavior adaptation by the humans based on their prediction of the performance of the agent partner in future teamwork situations?
- How does interaction with one virtual agent inform or bias the human in subsequent interactions with other virtual agents?

The reason we investigate these issues is eventually for augmenting agents with the necessary capabilities to collaborate with humans. To do so, we developed a virtual teamwork game where human players interact for a small number of teamwork situations with an agent. In each interaction, the human knows about the total work units to be performed to achieve the team goal and has to choose its effort level without explicitly coordinating with its partner. The

effort level of the partner and the combined team performance are revealed to the players after the game. We collected data on human effort level 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 on the human's trust and on the resultant choice of work effort by the human. We performed experiments with the human workers where they were involved in several games with different agent teammates. The goal of these experiments is to observe how past experience biases a human player's trust in their partners in subsequent games.

These observations and analysis are further used to develop learning agents that adapt its behavior to optimize team performance with the given human. This task is particularly challenging as human players are adaptive and hence any adaptation by the agent partner runs the risk of being at cross-purposes with the adaptation by the human. This is similar to the well-known problem of learning a moving target function in the multi-agent learning literature. Our learning approach was to predict the following: (a) initial work effort chosen by a human at the start of a game, and (b) subsequent efforts as a function of the previous experiences in the game, any prior game experience of the human, and the demographic of the player.

The organization of the paper is as follows. First, we discuss related work and adress how our work differs. The human-agent teamwork model is presented in Section 3, Section 4 presents the experiments and discusses the results. Finally, Section 5 provides a brief conclusion and addresses the future work.

2 Related Work

Building and persisting trust is a critical component of successful relationships both between individuals and within groups. In other words, the outcome of an interaction often depends on trust, which has ability to handle risks that cannot be prevented in other ways. In domains of human-agent interaction, understanding dynamics of human trust and their attitude towards agents is a key insight for effective deployment of agents in human-agent teamwork, such as collaborative health care [3], emergency management [12], disaster response [13], and so on.

Developing a predictive model of human trust in agents, however, is an ongoing challenge problem. While humans routinely exhibit non-rational behavior, rationality has often been a key assumption in existing trust models within the domain of multi-agent systems. The consequence is the huge gap between what is predicted by formal models and the actual behavior of humans. Castelfranchi et al. [4] assert to broaden the foundation of *Rational Decision Theory* to a more complex decision making model in which emotions affect the goals, beliefs, decisions, and actions. In contrast to the trivial idea that always success (failure) increases (decreases) trust, Falcone and Castelfranchi [6] point out the necessity of a cognitive attribution process for updating trust based on the interpretation of the outcome of truster's reliance on trustee and trustee's performance. Simi-

larly, Hoogendoorn et al. [9] introduce a bias-based trust model by incorporating the human biases and empirically verify with humans.

In the mid-nineties, Nass et al. [10] conducted one of the pioneering human-computer teamwork studies in which computer is a peer cooperative partner. Their conclusion is that social dynamics of human-computer interaction are correlated with the social dynamics of interactions of solely humans. Though, recent findings suggest that people neither treat computers exactly like they treat to humans nor do people treat computers as simple asocial tools. People's perception of agents' characteristics has a significant influence on their attitude towards agents and the development of a continued effective partnership over repeated interactions. Some of the empirical findings on human trust in agents are as follows: agents with the appropriate emotion expressions can be perceived as more trustworthy and be selected more often for repeated interactions [1], people may prefer to cooperate with agents to human avatars when the agents are associated with more positive categories than the human avatars [5], people like and trust a facially similar agent more than a facially dissimilar agent [17], and people favor humans more than agents in reward sharing [14].

van Wissen et al. [15] explore how the cooperativeness levels of agents affect humans' cooperativeness in the context of repeated Ultimatum Game. The game is implemented with Colored Trails [7] which is a negotiation environment for studying decision-making behavior and dynamics between players. Their results support that cooperative (non-cooperative) behavior induces cooperative (punishment) behavior. Our research departs from [15] by considering the trust and fairness in a human-agent teamwork setting rather than cooperativeness between two players in a negotiation setting. van Wissen et al. [14] presents a behavioral study of fairness and trust in human-agent teams by examining how people choose their teammates and their commitment to the team in a dynamic environment. They observed that people offered lower rewards to agent teammates (with comparison to human teammates). However, people's defection behavior does not depend on whether it is an agent-led or a human-led team. Hanna and Richards [8] investigate the impact of multimodal communication on humans' trust in agents and the effect of trust on commitment that affects the performance of human-agent teamwork. Their empirical findings show that trust significantly correlates with commitment and human's commitment enhances team performance.

Our research extends these studies as follows:

- We are interested in human-agent teams rather than mere interactions between two players, e.g., [1, 5, 15, 17].
- We focus on teamwork environments in which there is neither explicit communication between human and agent in contrast to those in [8, 14] nor the embodiment of agents in contrast to those in [1, 8, 5, 17].
- We explore repeated interactions, rather than one-shot [16], of fixed teams, rather than dynamic teams [14].

To the best of our knowledge, we are the first to study trust in human-agent teamwork within a repeated team game scenario where agents are autonomous team members and no prior coordination exists.

3 Human-agent Teamwork Model

In this section, we introduce the virtual teamwork domain we have used in our experiments and the agents we have designed to work with humans to achieve team goals.

3.1 Game of Trust

The Game of Trust (GoT) is a team game in which two players form a team and have n interactions. In each interaction, players are assigned a task, t_i, as a team. The team task consists of n_{t_i} atomic subtasks of the same type. There are no dependencies between the subtasks. We assume these tasks do not require any specialized skills and hence both the human and the automated player can accomplish them. Though the general GoT framework can support more complex task types, for our experiments and for the remainder of the paper, we will consider tasks with non-differentiable subtasks where only the number of subtasks matter. For example, a task 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. Players choose the number of subtasks they want to perform without knowing the amount of task that other player will perform. After performing subtasks individually, players are told whether they have achieved the team goal, i.e., whether the two players combined have completed the required number of subtasks, and the number of subtasks that the other player completed.

There is a cost of performing tasks that is computed by the cost function, c, based on the number of subtasks accomplished. Both players have their own individual payment accounts, which have an initial balance, b, at the beginning of the game. Players are explained about the cost and reward functions. The cost of the subtasks that is performed by each player is withdrawn from the corresponding account. If the total number of subtasks accomplished by the players is equal to or greater than the size of the team task, it means the players achieved the goal. Then the reward, computed by the reward function r, is equally split between players, i.e., deposited to their individual accounts. If combined work that the players accomplished, is less than the team task, the interactions fails and no reward is given.

For the rest of the paper, by utility of a player we refer to half of the team reward (if the players achieved the task) minus the cost of performing subtasks individually. If they cannot achieve the team task, both players lose utility from this teamwork instance. Finally, *social utility* corresponds to the sum of the

utilities of the two players. Social utility is optimized when the total number of words transcribed by team members is precisely equal to the number of words assigned to the team.

There exist various games to study trust in literature. Among these, Public goods game[1] fulfills most of our requirements for studying teamwork without communication (The amount of tokens contributed to the public pot by an individual is analogy to the individual work and the tokens in the pot represents the teamwork). However, this game is not appropriate for our research due to the pre-defined size of the team task in our model.

3.2 Teamwork Domain

In our particular instantiation of the teamwork domain, a team consists of one human and one agent playing the *Game of Trust*. Though the monetary reward significantly affects humans' perceptions and decisions, the participants' effort and time spent on the study that are related to the nature of the task have a serious influence on humans' behavior. Therefore, we intend participants to perform a relatively real task rather than making mere decisions to achieve artificial goals such as Colored Trails [7]. This task should not require any special skills because complex tasks may impose extra constraints and undue burden on participants and add complicating factors like the of quality of work for subtasks that are not important for our work. 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 subjects having additional motivations that influence their choice of effort level or contribution to the team goal.

Based on these considerations, we choose an audio transcription task for the human-agent teamwork instances. In this domain, the human-agent team are tasked to transcribe a subset of a large collection of audio segments, each corresponding to a English word, to text. In order to reduce the experimental cost, a transcription task consists of a small number, e.g. ten, of words. Hence, in this domain, the task that is assigned to the team corresponds to the 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 growth of human trust from repeated interactions and their completion is of no intrinsic value to us, we simply count the number of words accurately transcribed and give credit even when team members transcribe overlapping words sets). We will use the term task size to refer to the number of words to transcribe, i.e., number of subtasks, in an interaction. A predetermined sequence of increasing task sizes[2] is used in the domain.

[1] In public goods game, participants choose how many of their private tokens to put into a public pot. The tokens in the pot are multiplied with a factor and equally shared between participants

[2] To evaluate human performance in a simple setting, we also conducted experiments where the task sizes did not change over generations. However, those results are not presented due to space limitations.

The purpose of the transcription task is to mimic a real teamwork environment where the participants have to collaborate with their automated partner to achieve their shared goal which they cannot achieve by themselves. Though we have no interest in the transcribed words, the subjects are still required to transcribe a word with at least 60% accuracy to receive credit for successful transcription. 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 sequence, where each game consists of several interactions with one of the several automated player types.

Both human and agent players are expected to be self-interested: the more words a player transcribes, the higher is that player's cost and the lower its gain from the teamwork is. 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 they have to make at each interaction and is based on their trust of their teammate for contributing to the team goal.

3.3 Agent Teammates

We employed several agent types with distinct behaviors to evaluate the growth of human trust in agent teammates over repeated interactions. The agents vary in their attitudes towards choosing the portion of the team task they will perform. These agents can be grouped into two categories: (1) static agents and (2) adaptive agents.

Static Agents Static agents exhibit the same behavior regardless of their teammate or their prior experience. The amount of work they perform[3] is a function of the task size. Three static agents described below put varying levels of effort over different interactions in a game. By the *effort level* of a team member we refer to the portion (percentage) of the total work units completed by this team member.

Fair agent completes half of the team task, i.e., $a^i_{fair} = \frac{|t_i|}{2}$, in every interaction.

Fixed Task Doer completes the same amount of work units in all interactions. The fixed amount, a^i_{fixed}, is a function of the total work units required in the overall game:

$$a^i_{fixed} = \frac{\sum_{j=1}^{n} 0.5 * |t_j|}{n}, \tag{1}$$

for all i. In fact, Fixed Task Doer and Fair perform the same total work over the course of the game but Fixed Task Doer's effort level varies throughout

[3] In reality, the agents do not transcribe any words, but the *GoT* framework reports that they do so to the human players. Additionally, we assume that agents transcribe all words accurately.

the game. Given that the size of the team task increases over interactions, the percentage of the team task that Fixed Task Doer performs is more than half at beginning, and then decreases to less than half towards the end of the game.

Complementary: The work units performed by Complementary agent is equal to the size of the team task minus a constant that is equal to what Fixed Task Doer performs:

$$a^i_{comp} = |t_i| - \frac{\sum_{j=1}^{n} 0.5 * |t_j|}{n}. \quad (2)$$

Just as Fair and Fixed Task Doer, this agent also delivers half of the efforts required in the overall game. In contrast to Fixed Task Doer, however, Complementary starts with an effort level less than half of the required effort and increases its contribution towards the latter interactions of the game.

The benefits of observing humans behavior with the teammates with static strategies is threefold: investigating humans reaction to agent teammates displaying varying effort levels, evaluating whether humans adapt their behavior to the static behavior of agents whose effort levels are easy to predict, and surveying humans perceived trustworthiness and fairness of their agent teammates.

Adaptive Agent Although static agents are useful for analyzing humans' reliance on their agent teammates, they are not designed to optimize the performance. Teamwork with such agents may result in repeated failures to complete the team task throughout the game or achieving the team goal with wasteful, redundant work. In both cases, both social and individual utilities will be adversely affected. To achieve the goal of effective agent teammates that can produce optimal social utility when paired with human players, we implement *Offline Learner* agent.

Offline Learner, as the name suggests, is trained by using the data we collected from the teamwork experiences of humans with static agents (see Section 3.3). In particular, we are interested in the effort level choices of individuals in teamwork when paired with different agent partner types. However, given the irrational, unpredictable, and "noisy" behavior of 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 [11]. 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.

When playing with a human player, there are two types of predictions the *Offline Learner* is tasked to make: the initial work units that human partner will undertake in the first interaction of the game, $\hat{h}^1$, and the work units in any of the following interactions, $\hat{h}^i$, $i = [2 \dots n]$. With an accurate prediction, the agent can then choose to complete the remaining of the team task, $a^i_{learner} = |t_i| - \hat{h}^i$,

to achieve the team goal optimally and without redundancy or falling short of the team goal.

To predict the initial effort level, the *Offline Learner* needs to know the previous experiences of the human partner [4]. Human players' decisions are affected by their experiences and biases. For example, when a human player first interacts with a new agent partner, her effort towards the team goalwill be influenced by her prior teamwork experiences with other agents in prior game(s). For example, interacting with an untrustworthy partner may cause her to be more cautious in subsequent teamwork scenarios. The Offline Learner may then be able to predict the human effort level of a human player in the first interaction using a regression function based on the human's total amount of work and total utility in the previous game. A similar approach, which additionally uses the player's effort level in the first interaction, is used to predict the human player's effort level in the second interaction. For predicting the human partner's work on the subsequent interaction, however, the *Offline Learner* needs only the efforts expended by the human in the previous interactions of the current game and does not have to depend on information from her previous games. Specifically, next task choice of human is predicted by using a function of the effort level of the human in the previous interactions in this game, the shortfall or redundancy in the recent teamwork, and the change in the amount of team members' work.

3.4 Hypothesis

In the experiments, we investigated the following hypothesis:

Hypothesis 1 People's trust in their agent teammate is proportional to their teammate's effort level. ▮

Hypothesis 2 People's prior experiences with other teammates affect their trust in the current teammate. ▮

Hypothesis 3 People's prior experiences with other teammates affect their choice of effort levels. ▮

Hypothesis 4 Performance of human-agent teams is correlated with people's trust in their agent teammates. ▮

4 Evaluation

4.1 Game Configuration

The number of interactions in a game is five, which is duration not too long for the participants to be bored and still allows team members to adapt to

[4] In situations where this information is not available, the learner can start by performing the half of the team task or make a probabilistic choice considering the previous teamwork experiences of other humans and agents that did not cause a serious loss of utility.

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.

Agents: The overall team task in the game is transcribing 50 words in five interactions, thus total work units in the overall game is 25 for each player if they were to split the team tasks equally. The sequence of individual task choices in five interactions is given for each static agent as follows:

Fair: $a_{fair} = \langle 3, 4, 5, 6, 7 \rangle$;
Fixed Task Doer: $a_{fixed} = \langle 5, 5, 5, 5, 5 \rangle$;
Complementary: $a_{comp} = \langle 1, 3, 5, 7, 9 \rangle$.

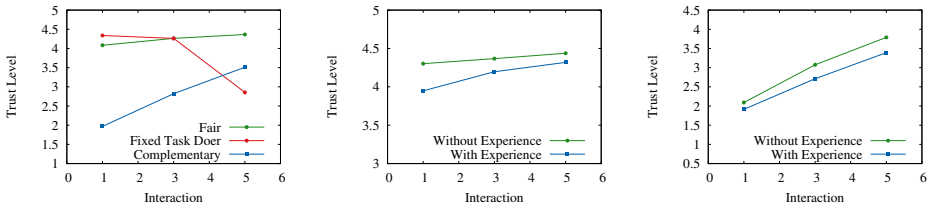
In the rest of the paper, we refer to an experiment which consists of a number of games by the term *session*. In order to investigate the humans' trust behavior, their adaptation to agent teammates, and the effect of previous experiences, we designed two sets of experiments as follows.

1. In the first set of experiments, the participants played games with three static agents, i.e., three games in a session. Since the order of agents matters for the homogeneity of the data, we conducted experiments for each possible ordering of three agents and collected data from 40 or more participants for each order.
2. In the second set of experiments, the participants played two games in a session. The purpose of this experiment is to compare the performance of the learner agent with that of the static agents. Therefore, we designed a series of sessions, in each of which the participant played with the *Complementary* in the first game and then played with one of the other agents in the second game. For each group, we collected data from 30 or more participants.

4.2 Survey

The game includes a short survey on trust for measuring human players' perceived trustworthiness and fairness of their teammates. Participants were asked to complete this survey at the end of 1^{st}, 3^{rd}, and 5^{th} interaction of a game after they were shown the outcome of the recent teamwork. This short questionnaire, that is adapted from [2], 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,
4. I believe my teammate trusts me to contribute fairly/effectively to our team's goal.



(a) Variation on individuals' trust in agent teammates (b) Trust in Fair (c) Trust in Complementary

Fig. 1. Individuals' trust in agent teammates

4.3 Recruitment of Participants

We recruited 260 participants through Amazon Mechanical Turk[5]. Approximately, 41% of the participants were female. Age distribution was as follows: 18 to 34 years, 48%; 35 to 54 years, 55 or higher 8%. Regarding to education, 50% of the participants had either high school degree or some college degree, 34% of the participants had bachelor's degree, and 14% of the participants had graduate degree or more. The data from 10 participants is eliminated due to insufficient attention, hence we analyzed the data collected from 250 participants.

The participants were paid $1 plus a bonus that is proportional to their total score (up to $1). We identified three qualifications for the eligibility to participate: (1) worker must reside in US (to avoid cultural biases), (2) worker must have completed more than 1000 tasks, and (3) worker's approval rating must be greater than 97%. Additionally, a worker is allowed to participate in the study only once. At the beginning of the experiment, the participants were told that their teammate is an automated computer player and the teammates in games have different attitudes towards collaboration.

4.4 Experimental Results

Perceived Trustworthiness of Agent Teammates Trust level of an individual in agent teammate is computed as the average of the individual's responses to the first three survey items (see Section 4.2). Though, we report the average of three responses, responses to each one of three items, follow similar patterns[6], as expected. We ran (one-way) ANOVA to verify that the changes in individuals' perceived trust in their agent teammate throughout a game and the differences between individuals' trust in three agents are statistically significant.

[5] http://www.mturk.com/

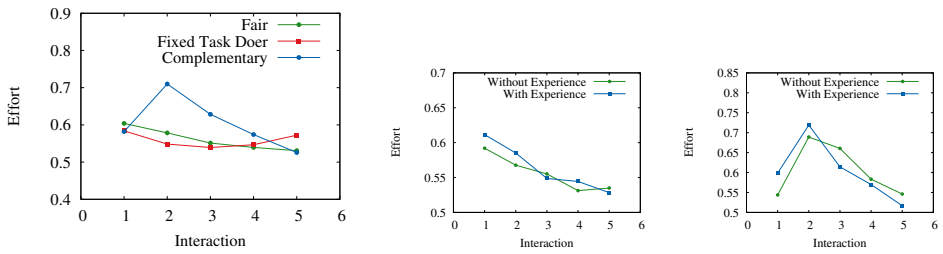
[6] The perceived fairness and perceived agent trust in individuals follow similar patterns as perceived trust in agent partners. These results are therefore not presented here due to space limitations.

Figure 1(a) shows individuals' perceived trust in agent teammates over interactions. Fair agent is perceived as a stable trustworthy partner, thus individuals' trust in it is increasing throughout the game ($p < 0.001$). Fixed Task Doer progressively reduces its effort level (0.83, 0.62, 0.50, 0.42, 0.36). Though its trustworthiness drops ($p < 0.001$), it does not drop off drastically and is still above the Complementary agent's trustworthiness ($p < 0.001$) at the end of third interaction. The perceived trustworthiness of the Complementary agent increases steadily ($p < 0.001$) as it increases its effort over the interactions (0.16, 0.37, 0.5, 0.58, 0.64). One key observation here is that though the Complementary agent performs equal or more than Fair agent from the third interaction onwards, it is never viewed to be anywhere close to be as trustworthy as the Fair agent($p < 0.001$)! This observation confirms the oft-quoted long-lasting effect of first impressions on people's perception. The correlation between individuals' trust in an agent teammate and the agent teammate's effort levels confirms *Hypothesis.1.*

Figure 1(b) shows the impact of prior experiences on individuals' perceived trust in their teammates. We present two cases: (1) Fair agent was the first agent that the participants played with (without any prior experience) and (2) Participants played with the Fair agent after playing with other agent(s)(with prior experience). Two (four) of the six groups, i.e., ordering of agents, fall into the first (second) case. Individuals' perceived trust in the same agent differs with respect to different order of partners, which confirms the fact that people's previous experiences bias their opinions about other agents they interact with at a later time [9]. Although, the difference in perceived trust levels between two cases is more remarkable after the 1^{st} interaction ($p < 0.001$), it decreases towards the end of the game($p < 0.1$ after three interactions and $p = .24$ after five interactions). In other words, the influence of individuals' biases on their perceptions decreases as their experiences increase with the current teammate.

Individuals' perceived trust in the Complementary agent also changes based on prior experiences as shown in Figure 1(c). However, the difference between perceived trust levels in two cases increases throughout a game in contrast to Figure 1(b). ($p = 0.20$, $p < 0.05$, and $p < 0.01$ after one, three, and five interactions, respectively). Consequently, these results confirm *Hypothesis.2* and demonstrates that human's perceived trust in the trustee is a complex cognitive process rather than a mere rational decision based on the objective quantities [6], even in this simple teamwork environment in which only the number of subtasks that is performed by the teammate matters.

Effort Distribution Figure 2(a) shows the variation in the average effort level of individuals over interactions for three static agents. We ran ANOVA to verify the adaptation of individuals to their agent teammates: Their effort level changes based on teammate's effort throughout a game ($p < 0.001$) as shown in Figure 2(a). Participants were adaptive to their teammates' effort levels over interactions, i.e., they adjusted their choice of individual task accordingly. Figure 2(b) and Figure 2(c) show individuals' effort over interactions when they



(a) Individuals' effort with three agents (b) Individuals' effort with Fair (c) Individuals' effort with Complementary

Fig. 2. Individuals' effort

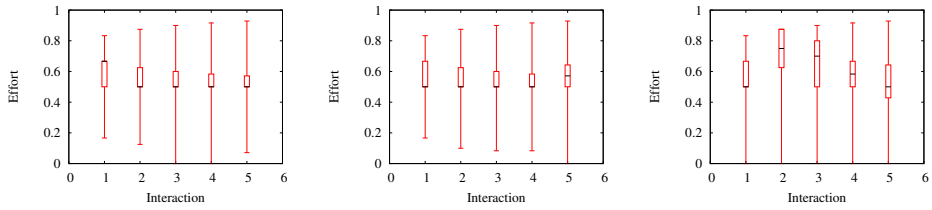
played with or without any prior experience with other agent(s). Although, we observe slight differences in task choices between two cases, they are not statistically significant ($p > 0.1$). Consequently, *Hypothesis. 3* is not confirmed.

Quartile analysis of individuals' effort distribution is presented in Figure 3. In games with all three static agent types, 75% of the participants completed half or more of the team task in the first interaction, which suggests that they were cautious when playing with a new teammate.

When playing with the Fair agent (see Figure 3(a)), the proportion of participants who performed more than half of the teamwork shrinks towards the end of the game as their trust in their partner increases. When playing with the Complementary agent (see Figure 3(c)), the majority of the participants reacted with a sharp increase in their effort in the second interaction as a result of disappointment in the first interaction. For the remainder of the game, half of the population adapted well to the Complementary agent.

In Figure 3(b) (playing with the Fixed Task Doer), the population of participants performing half or more of the team goal consistently shrinks up to the fourth interaction. Possibly they thought that they will continue to accomplish their goal even with reduced effort until they find out that Fixed Task Doer delivered less than half of the team task. In contrast, a minority of the participants increased their effort over interactions as they saw their teammate's effort is continuously decreasing. Similarly, most of the participants, who performed less than half of the team goal, increased their effort in fourth interaction, whereas a few of them gave up and delivered the minimum amount of work.

Adapting to Human Teammates In this section, we present the results of the second set of experiments, where all four groups of participants played the GoT game with the Complementary first and then each group played with a different agent in the second game. Figure 4 depicts the social utility obtained for each interaction of the second game. Interestingly, the *Offline Learner* performs almost as well or better than the Fair agent for most game interactions. We



(a) Fair (b) Fixed Task Doer (c) Complementary

Fig. 3. Quartile analysis of individuals' effort

also observe that the performance of teamwork with trusted agent teammates is higher than the performance of teamwork with less trusted agent teammates, which confirms *Hypothesis.4*.

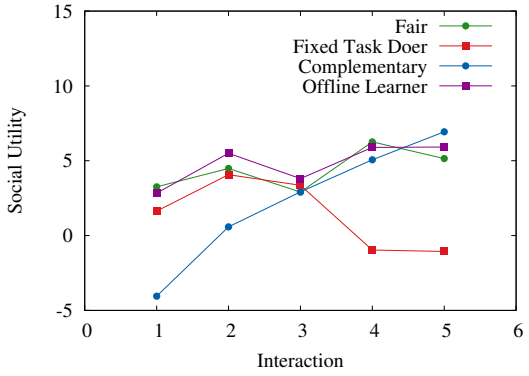


Fig. 4. Social utilities of teamwork with static and adaptive agents

Table 1 presents the cumulative results of teamwork with Fair and Offline Learner. Though both agents are able to achieve about the same number of team goals, the *Offline Learner* is able to achieve higher social utility by reducing redundant work. In particular, it is interesting to note that this efficiency improvement mostly benefits the human player through increased utility!

As a final note on *Offline Learner*, individuals' reported trust levels are 4.04, 4.13, and 4.17 at the end of 1^{st}, 3^{rd}, and 5^{th} interaction, respectively. Thus, the *Offline Learner* is perceived as trustworthy as the Fair agent by the participants.

Table 1. Game results

	Fair	Offline Learner
Goals achieved	4.44	4.43
Words transcribed	43.78	43.91
Redundancy	4.52	3.83
Human Utility	8.75	10.58
Agent Utility	13.31	13.56
Social Utility	22.01	24.23

5 Conclusion and Future Work

In this paper, we introduced a human-agent virtual teamwork model, *GoT*, to study the development of human trust in automated virtual partners over a few initial collaborations for achieving team goals. We performed experiments with this framework in a word transcription domain to evaluate the effects of initial impressions, changing task demands, and former agent teammates on the behavior of human players. The empirical results confirm that individuals' trust is correlated with the agent teammate's effort, prior experiences of individuals affect their interactions at a later time, and performance of human-agent teams is correlated with the individuals' trust in their agent teammates.

In spite of limited amount of data and hard to predict non-rational behavior of humans, the *Offline Learner* was able to produce higher social utility compared to the other agents by reducing wasteful, redundant activity. Of particular note is the observation that the learner's adaptation significantly improved the participants' and the social utility!

Our follow-up research has two directions. First, we will compare human players' behavior towards a human teammate and towards an agent teammate. This comparison is necessary because we still do not know the conditions in which people do (or do not) differentiate human teammates from agent teammates. Afterwards, our goal is studying human-agent teamwork with complex team tasks. A complex task comprises of subtasks that require different abilities that is similar to most of teamwork in real-life. In this domain, human and agent know neither each other's abilities for different task types nor the alignment of their own abilities with the abilities of others.

References

1. D. Antos, C. de Melo, J. Gratch, and B. Grosz. The influence of emotion expression on perceptions of trustworthiness in negotiation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI11)*, pages 772–778, San Francisco, CA, August 2011.
2. B. A. Aubert and B. L. Kelsey. Further understanding of trust and performance in virtual teams. *Small Group Research*, 34(5):575–618, October 2003.

3. M. Ball, V. Callaghan, M. Gardner, and D. Trossen. Achieving human-agent teamwork in ehealth based pervasive intelligent environments. In *Proceedings of 4th International Conference on Pervasive Computing Technologies for Healthcare*, pages 1–8. IEEE Xplore, 2010.
4. C. Castelfranchi, F. Giardini, and M. F. Relationships between rationality, human motives, and emotions. *Mind & Society*, 5:173–197, 2006.
5. C. de Melo, P. Carnevale, and J. Gratch. Social categorization and cooperation between humans and computers. In *The Annual Meeting of The Cognitive Science Society (CogSci'14)*, pages 2109–2114, July 2014.
6. R. Falcone and C. Castelfranchi. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 740–747, Washington, DC, USA, 2004. IEEE Computer Society.
7. B. J. Grosz, S. Kraus, S. Talman, B. Stossel, and M. Havlin. The influence of social dependencies on decision-making: Initial investigations with a new game. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, pages 782–789, 2004.
8. N. Hanna and D. Richards. "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)*, 2014.
9. M. Hoogendoorn, S. Jaffry, P. v. Maanen, and J. Treur. Modelling biased human trust dynamics. *Web Intelligence and Agent Systems Journal*, 11(1):21–40, April 2013.
10. C. Nass, B. Fogg, and Y. Moon. Can computers be teammates? *International Journal of Human-Computer Studies*, 45(6):669–678, April 1996.
11. L. Panait and S. Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, November 2005.
12. P. Robinette, A. R. Wagner, and A. M. Howard. Building and maintaining trust between humans and guidance robots in an emergency. In *AAAI Spring Symposium: Trust and Autonomous Systems*, pages 78–83, Stanford, CA, March 2013.
13. N. Schurr, J. Marecki, M. Tambe, P. Scerri, and J. Lewis. *Published Articles & Papers*, volume 42, chapter The Defacto System: Coordinating Human-Agent Teams for The Future of Disaster Response. Kluwer, 2005.
14. A. van Wissen, Y. Gal, B. Kamphorst, and M. V. Dignum. Human-agent teamwork in dynamic environments. *Computers in Human Behavior*, 28(1):23–33, 2012.
15. A. van Wissen, J. van Diggelen, and V. Dignum. The effects of cooperative agent behavior on human cooperativeness. In *Proceedings of eighth International Confefence on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 1179–1180, May 2009.
16. F. Verberne, J. Ham, and C. J. H. Midden. 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):799–810, May 2012.
17. F. M. F. Verberne, J. Ham, and C. J. H. Midden. Familiar faces: Trust in a facially similar agent. In *HAIDM Workshop at the Autonomous Agents and Multiagent Systems (AAMAS'14)*, 2014.