International Journal of Social Robotics
https://doi.org/10.1007/s12369-021-00747-z

Investigating Trust in Interaction with Inconsistent Embodied Virtual Agents

Reza Moradinezhad[1] · Erin T. Solovey[2]

Accepted: 6 January 2021

Abstract

Embodied Virtual Agents (EVAs) are used today as interfaces for social robots, educational tutors, game counterparts, medical assistants, as well as companions for the elderly and individuals with psychological or behavioral conditions. Forming a reliable and trustworthy interaction is critical to the success and acceptability of this new form of interaction. In this paper, we report on a study investigating how trust is influenced by the cooperativeness of an EVA as well as an individuals prior experience with other agents. Participants answered two sets of multiple choice questions, working with a different agent in each set. Two types of agent behaviors were possible: *Cooperative* and *Uncooperative*. In addition to participants achieving significantly higher performance and having higher trust for the *cooperative* agent, we found that participants' trust for the *cooperative* agent was significantly higher if they interacted with an *uncooperative* agent in one of the sets, compared to working with *cooperative* agents in both sets. Furthermore, we found that participants may still decide to choose agent's suggested answer (which can be incorrect) over theirs, even if they are fairly certain their own answer is the correct one. The results suggest that trust for an EVA is relative and it is dependent on user's history of interaction with different agents in addition to current agent's behavior. The findings provide insight into important considerations for creating trustworthy EVAs.

Keywords Human-Computer interaction · Human-Robot interaction · Social robots · Trust · Embodied conversational agents · Virtual assistants

1 Introduction

As much of human commerce, healthcare, entertainment, education, and other enterprises move to virtual environments, interactions that require trust and cooperation increasingly involve virtual agents. Research on human-computer interaction (HCI) indicates that users tend to interact differently with human-like agents than with other types of interfaces [20,30,44]. A comparison of automated agents with varying levels of human-like behavior indicates that machines without human features invoke high levels of trust initially, but that this trust is more adversely affected by failure or unreliability. Agents with more human features rate lower on initial trust; however, this trust is more resilient, possibly because a more human-like aid is held less accountable to the elevated standards we ascribe to machines [12]. Human-like behavior in virtual agents can also impact trust by implying personality traits associated with positive or negative interactions. For example, agents that express empathy through facial emotions are rated as more jovial, expressive, cheerful, and less irritating, strange and cold than a non-empathetic agent [38]. Embodied Virtual Agents (EVAs) are computer agents that look like humans and are capable of facial expressions, body gestures, and sometimes conversation to facilitate more natural and engaging interactions with users [38]. This paper explores the interaction between humans and EVAs in cooperative and uncooperative conditions to increase understanding of how trust operates in these interactions. We explore how trust in an agent is impacted

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12369-021-00747-z.

✉ Reza Moradinezhad
rm976@drexel.edu

Erin T. Solovey
esolovey@wpi.edu

[1] Department of Computer Science, Drexel University, Philadelphia, PA, USA

[2] Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA

Published online: 24 March 2021

if the agent is not cooperative at all times, or if the user has previous (negative or positive) experiences with other agents.

In Sect. 2.1 we provide different definitions of trust and discuss the common main values they share with a broader definition of trust which is called *interpersonal trust*. Our work inspects how users perceive values associated with interpersonal trust such as integrity, dependability, reliability, good intention, and confidence in agents as factors of agent's trustworthiness. We explore *subjective, perceived trust* based on the values associated with interpersonal trust [24], and *objective, behavioral trust* based on participants' conformity with the agents' feedback [39]. Our experiment allows us to investigate how humans build trust towards EVAs, especially when the agent is inconsistent and is not cooperative 100% of the time. We also explore how exposure to an agent with a different level of cooperativeness affects the already established trust towards previous agents and expected trust toward future agents. Our study centers around the following high level research question: Would a human's trust towards an agent be affected by their previous interaction with another agent?

To investigate this topic, we built a system in which humans interact with an EVA as an assistant in a question and answer (Q&A) task. The agent expresses six different facial expressions, ranging from highly positive (HP) to highly negative (HN), randomly assigned as reactions to the various answer choices. The users work with either a *cooperative* (HP assigned to the correct answer 80% of the time) or an *uncooperative* (HP assigned to the correct answer 20% of the time) agent to answer a set of general knowledge questions. Through this system, we conducted an experiment with 35 participants to explore questions related to agent behavior and user trust. The results shed light on how users build trust toward inconsistent EVAs, i.e. EVAs that do not assist the user 100% of the time. The results also provide insight into how the users adjust their trust for an EVA based on previous and future experiences with other EVAs. The findings of this study have implications for the design of more trustworthy user interfaces that contain EVAs. More specifically, the contributions of this paper are as follows:

- We report performance results showing that individuals perform better with cooperative agents and they also tend to initially rely on agents, even when they do not need to.
- As expected, we show that individuals find cooperative agents more trustworthy, using behavioral trust measures. Through a subjective trust questionnaire, we found further evidence that individuals trust cooperative agents more than uncooperative agents.
- We present findings from the trust questionnaire that uncovered differences in perceived trust in EVAs, based on prior experience with another EVA. In particular, individuals who had worked with an *uncooperative* agent rated the *cooperative* agent more trustworthy than individuals who only worked with *cooperative* agents. For those individuals, they had higher ratings of trust in the first *cooperative* agent they worked with than the second *cooperative* agent.

2 Related Work

In this section, we discuss research areas on which our work is based. First, we review different categories of trust and discuss *interpersonal trust* which is one of the measures we use in this study. After defining trust in the context of our study, we explore the benefits of EVAs over conventional user interfaces and human agents in terms of building trust with users. Most of the work in this section are in medical or therapeutic fields since trust to the other party is deemed an important factor in such fields. Then, we highlight work in the area of behavioral indicators of trust in EVAs. This work provides models which introduce features that can be used in an EVA to represent different levels of trust and deception. Following that, we introduce a study and a toolkit which can be used as a solid reference for designing realistic EVAs. Finally, we discuss studies looking at the effect of inconsistency in agents in users' performance and their perception of trust.

It is worth mentioning that some of the following works use the term Embodied Conversational Agent (ECA) instead of EVA. The difference between ECAs and EVAs is that EVAs are not necessarily designed to converse with the user while ECAs are specifically designed to hold conversations with the user. A Virtual Human (VH) can be either ECA or EVA, depending on how it is implemented. An avatar could refer to a VH or just pictures/videos of an animated human.

2.1 Definitions of Trust

A review on literature on definitions of trust confirms that trust comes in different categories and the definition of each category depends on multiple factors such as the parties involved (e.g. person-person, person-business, person-automation), and the type of interaction. A standard and widely used scale for measuring trust between human and automation is a scale proposed by Jian et al. [24] which measures *interpersonal trust*. Therefore, in the following, we review some of the widely accepted definitions of trust, with a special focus on *interpersonal trust*.

Many categories of trust fall under the definition of *interpersonal trust*. Mayer et al. [33] define interpersonal trust as "the willingness of a party to be vulnerable to the outcomes of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party". In line with this definition, Doney et al. [13] suggest that the *cognitive state of trust* in another person is built on beliefs about the other person's benevolence (good intentions) and credibility (reliable behavior). Explaining the *uncertainty-reduction model of trust*, Berscheid et al. [3] argue that humans gradually build up to their assumption about dependability of another person based on actual evidence from that person's behavior. Two other categories of trust which have direct correlation with interpersonal trust are *competence-based trust* and *integrity-based trust*. Based on Butler and Cantrell's paper [6], Kim et al. [27] define *competence-based trust* as perceiving the trustee as someone who has the necessary interpersonal and technical skill set to perform the task. They also define *integrity-based trust* as perceiving the trustee as someone who conforms to a set of principles that are considered acceptable based on the definition of *interpersonal trust* discussed in the beginning of this paragraph. Hence, it is evident that values such as good intentions, reliability, dependability, competence, and confidence in the other which are associated with *interpersonal trust* are common values in other definitions of trust as well.

Although many of the aspects of human-human trust can be extended to human-automation trust, it is important to define trust between humans and automation separately. Discussing *trust in automation*, Lee and See define trust as the mindset that an agent assists one in achieving their goal even in situations where there are uncertainties and one can be vulnerable to the outcomes of the agent's actions [29]. Another definition for *trust in automation* is provided by Heerink et al. [22] where they introduce a toolkit for measuring acceptability of social robots used in elderly care. They define *trust in automation* as the "belief that the system performs with personal integrity and reliability". These definitions suggest that the same factors associated with *interpersonal trust* between humans apply to *interpersonal trust* between humans and automation as well. However, there is a stronger emphasis on integrity and reliability when it comes to trusting automation.

2.2 EVAs and Perceived Trustworthiness

When discussing trust in EVAs, it is important to understand how this trust is compared to trust for conventional computer systems and human agents who do the same task. One field in which trust plays an important role is physical and mental health. In this field patients are asked to reveal their personal information and/or accept advice about their physical or mental health. Early studies suggest that using a realistic human-like agent (e.g. video rather than text or audio) could potentially cause an increase in perceived trustworthiness [44]. In this section, we look at studies on applications of EVAs in the field of physical and mental health and how users perceive their trustworthiness, compared to conventional computer systems and humans.

Previous work suggest that EVAs are perceived as more trustworthy by the users compared to human agents and traditional computer systems such as WIMP (Windows, Icons, Menus, Pointer). Additionally, users are more tolerant of errors made by EVAs. Comparing the trust scores (using Jian et al. [24] scale) between a robotic and WIMP based decision support systems for nurses and doctors on a labor and delivery floor, Gombolay et al. [20], based on many other studies, suggest that embodied and anthropomorphic systems are rated more favorably by users in terms of trustworthiness. They also report that users were more tolerant of the errors made by the robotic agent. A similar conclusion can be drawn from a study by de Visser et al. [12]. They report that initially, trust was higher for the WIMP application in their study which served as an assistant for a trust task than the avatar and human who did the same. However, as the humanness of the agent increased, "trust resilience" of the participants also increased. It means that as the reliability of the agent dropped, trust toward the WIMP application dropped faster than trust toward the avatar and the human agent. This suggests that although it takes longer for users to build trust toward EVAs, once built, it would be more resilient.

When it comes to sharing sensitive and personal information, especially when it is related to one's health, there is an abundance of studies that suggest patients are more comfortable sharing information with an EVA controlled by a computer than humans. Lucas et al. [31] investigated the interaction between mental health patients and a Virtual Human (VH) which was able to interact with users via verbal empathetic feedback (e.g. "I'm sorry to hear that"), and nonverbal behaviors (e.g. nods and facial expressions) to convey active and empathetic listening. They argue that in contexts of health and mental health, patients usually don't fully disclose information about their condition. The authors hypothesize that if the patients are told that the VH is being controlled by a computer, rather than a human operator (in both conditions the VH was actually controlled by a human), the willingness of disclosing information will increase. Confirming their hypothesis, it became evident that patients felt more comfortable revealing information when they thought the VH was controlled by a computer, as opposed to a human. A study by Lisetti et al. [30] also shows that participants in a Drinker's Check-Up (DCU) intervention online platform were more willing to accept the system and continue working with it when the interface was an EVA rather than text-only system or in-person visits. Similarly, Devault et al. [11] show that participants felt more comfortable talking about personal feelings and experiences (specially related to their mental health) to a wizard-of-Oz agent than a human in face-to-face interaction. They suggest this is due to participants feeling more comfortable sharing sensitive information with a competent computer agent than a human. They back this theory by previous literature in the field [25,45]. These works show that humans seem to be willing to trust EVAs over conventional computer interfaces and even human agents in particular contexts.

2.3 Behavioral Indicators in Trustworthy EVAs

Nonverbal behaviors, such as facial expressions, gaze, gestures and postures, have a significant influence on interpersonal trust in face-to-face interactions [5]. In designing an ECA's nonverbal behavior, Bickmore and Cassell [7] considered the consistency in the ECA's facial expression as an important factor for gaining the users' trust. Also, in another work [4], they suggested that the agent's intonation, facial display, and hand gesture had an important role in user's perception of trust towards the agent. Below, we review recent work which specifically focus on behavioral indicators of trustworthiness in EVAs. These two studies provide insight into which nonverbal behaviors are perceived as trustworthy and which ones are not.

Elkins et al. [16] showed that participants found it easier to trust the agent when it was smiling. Rehm and Andre [41] studied how users react to subtle facial cues of a lying agent. The cues were inspired by six cues proposed by Ekman [15]; however, they used only two cues. One of the cues was "Mask" where true emotion is masked by deceiving facial expressions (e.g. smiling when feeling nervous) and the other was "Asymmetry" which is based on the idea that dishonest, voluntary facial expressions tend to be asymmetric, meaning that they cause more muscle activity on one side of the face. They report that when participants see an EVA showing deceptive facial cues during monologues presenting and introducing different movies, they rate the agent more negatively.

Ghazali et al. [19] use a robot which could show multiple faces and dynamic social cues as an assistant for playing a game. The robot could present persuasive messages to participants. They use a scale to measure participants' trusting beliefs created by Jian et al. [24]. Ekman and Friesen [18] and Todorov et al. [42,43] show that facial expressions involving upturned eyebrows and lips (attributes that humans seem to find more trustworthy) are deemed more persuasive, and they induce more trust compared to expressions with eyebrows pointing down and lips curled downwards at the edges (facial characteristics which are not normally perceived as trustworthy in interaction with humans). Also, they found that gender of the robot did not have a significant effect on participants' trust.

2.4 Designing Realistic EVAs

EVAs have the potential to augment trust with human-like behavior. Different features of human expression can convey trust, so the realism of the EVA's character design and its capacity for the subtle signaling involved in human expression play an important role in developing trustworthy EVAs. Facial expressions with dampened emotions are seen as more natural than those with exaggerated emotional expressions common to cartoon animation [23]. Realistic animations with more dampened emotional expressions rate higher on personality traits such as warmth, calmness, respectfulness, and competency. Another method of increasing the naturalness of EVAs is through the addition of nonverbal behaviors. The behavior expression animation toolkit (BEAT) [8] is an animation system that generates speaking characters from textual input. This system allows for nonverbal gestures and behaviors to synchronize with animated speech for a more natural and expressive character embodiment. Informed by linguistics and behavioral science, BEAT generates animated speakers who utilize context-dependent nonverbal movements natural to human conversation and interaction such as shifts of expression or gaze, changes in intonation, and head or body gestures. Tools such as BEAT, introduce a foundation for animators and researchers for fine-tuning facial expressions to design animated characters that can depict different personalities (for example, see Socially-Aware Robot Assistant(SARA) [40]).

2.5 Trust and Performance in Interaction with Inconsistent Agents

In this section, we explore how different levels of reliability affect users' perception of a systems trustworthiness. An early work using the Advanced Traveler Information Systems (ATIS) [21] indicates that while different levels of accuracy cause different levels of trust and compliance in users, the order of exposure to different levels of accuracy has a significant effect on the level of trust and compliance as well. High accuracy during initial interaction can maintain high levels of trust and compliance as accuracy decreases; however, poor initial experience can adversely affect trust and compliance even as accuracy in subsequent interactions increases (e.g. see Fox and Bohem-Davis [17]).

The users estimation of difficulty of the task performed incorrectly by an agent impacts levels of trust as well. Madhavan et al. [32] propose "easy-errors hypothesis" which suggests "that automation errors on tasks easily performed by operators undermine operator trust in and reliance on automated aids, even if the aid is, on average, more accurate than the unaided human operator". They performed an experiment in which participants did a target detection task with both easy and difficult trials. An automated diagnostic aid was present to help the participants do the task. The system's overall reliability was chosen to be 70% reliable. However, there were two groups in which the distribution of aid generated errors was either within the easy trial or the difficult trial. Results of this experiment show that participants who worked with the aid which had errors on easy trials mistrusted the aid, misperceived its reliability and disagreed with the aid more frequently than the participants who worked with an aid which was 100% reliable on easy trials and only had errors on difficult trials. A second experiment showed that the impact of “easy” automation errors on trust and dependence was significant even if the errors happened infrequently and contained false alarms, supporting the initial hypothesis.

However, there are also ways for automated systems to regain lost trust. User understanding why unreliability occurs can help to mitigate some of the adverse effects of agent error on trust. In one study, participants were asked to indicate the presence or absence of a camouflaged soldier with the aid of an automated system [14]. The experiment included 200 trials with every 5 trials being done either with or without the help from aid. The aid could be “superior”, making half as many errors as the participant or “inferior”, making twice as many errors as the participant. The aid was set up such that it didn’t make any errors in the introduction so that it was perceived as trustworthy and reliable. However, during the main task, observing the system aid make errors, participants found even the reliable aids untrustworthy. Therefore, the majority of the participants decided to rely on themselves even in the case of “superior” aid where its performance was far better than participants’. In the next steps, the participants were provided with an explanation regarding why the system might sometimes make errors. Providing this explanation increased the trust and reliance toward the system and participants started to trust the system again even when its trustworthiness was not guaranteed.

Higher levels of reliability are correlated with increased user trust and compliance across different forms of interfaces and different types of aids. Prior work suggests that as the level of reliability of an assistant agent increases, the performance of the users and their trust for the agent increases as well. The following studies report different levels of trust for agents with different levels of reliability. All of these works use the Jian et al. [24] trust scale to measure subjective trust.

In a driver’s advisory warning system (AWS) [37], using a 60% reliable agent which provided false alarms led to a decrease in performance and it reduced participants’ compliance, compared to a 100% reliable agent. False alarms also negatively affected the subjective evaluation of the AWS such as its usefulness or participants’ trust toward it. However, results of this study show that even using an unreliable AWS will provide benefits to the drivers compared to the baseline condition; i.e. responding earlier to critical driving situations and hence, a better performance. In an Advanced Traveler Information Systems (ATIS), Gruber [21] reports that participants in the 90% reliability condition showed higher compliance and reported significantly higher subjective trust than participants in the 72% reliability condition. Also, performance was improved by higher levels of reliability. However, they argue that it was easier for the users to detect mistakes and make a better judgment of the agents’ advice on a moderately reliable agent as opposed to an exceedingly reliable one.

Investigating the effects of different levels of reliability, shown by either false-alarms or misses, in an intelligent agent which assists users in supervisory control of multiple robots on users performance, Chen et al. [9] report that a reliable agent can improve the users performance by reducing the overall mission time. Putting more focus on trustworthiness of agents, de Visser et al. [12] conducted a study to investigate the effect of type of agent (computer, avatar, videos of a human actor) and agent reliability (100%, 75%, 50%, 0%) on human performance and trust in a trust task. This trust task consists of the following 5 steps: (1) selecting a number in a sequence, (2) watching an agent video, (3) observing the agent recommended number, (4) make a final number choice, and (5) observing the correct answer. They report that there was a direct correlation between an agent’s reliability and users’ subjective trust toward it. Also, compliance with the agent increased as the reliability increased; therefore, users performance was higher when they interacted with a more reliable agent.

To sum up, previous work show that the degree of reliability of the agent has a direct relationship with its perceived trustworthiness. In addition, as the agent becomes more human-like, it takes longer for trust to be built between users and the agent. On the other hand, once a certain level of trust is established, the rate of losing trust for a human-like agent can be slower than a conventional computer agent. This suggests that EVAs are better tools for collaborative work since in most scenarios long term trust is more important than short term trust.

The findings described above were based mostly on conversational aspects of trust in EVAs. Since this requires complex models and analysis, less attention was paid to the subtle effects of nonverbal behavior, which have also been shown to be important in trust [7,23]. In addition, most work on inconsistent agents examines different levels of unreliability in one agent. Therefore, it is still not clear how interaction with one agent affects user’s trust toward another agent.

3 Methods: Agent Behavior and Trust

In this section, we describe an experiment which examines the correlation between agent’s reliability and its perceived trustworthiness in addition to investigating the effect of interaction with one EVA on perception of trustworthiness of another EVA. Our decision to use two agents which have distinctive looks makes it possible for us to see how interaction with one agent affects user’s expected trust toward the future agents as well as past trust toward previous agents.

In particular, we explored the following four hypotheses:

1. Task performance will be influenced by agent's behavior (cooperative or uncooperative)
2. Average trust towards the *cooperative agent* will be higher than the average trust towards the *uncooperative agent*.
3. Average trust toward the *cooperative agent* will be higher if participants worked with the *uncooperative agent* first (vs *cooperative agent* first)
4. Average trust toward the *uncooperative agent* will be lower if participants worked with the *cooperative agent* first (vs *uncooperative agent* first)

3.1 Experimental Task

The main task for the study was a human-EVA collaborative task. We created a user interface where the user could work with an agent to find the correct answer to multiple choice questions. The interface displays a question with an answer grid containing nine potential answers. The agent's face is visible on the right side of the screen (Fig. 1). The questions were basic general knowledge questions and were selected from two online repositories [1,2].

Participants answered 50 multiple choice questions, while receiving feedback from the agent. Then, they answered a second set of 50 questions, receiving feedback from a different agent. The two agents were of the same gender and race in order to control for any effects that those factors could have on participants. In this paper, the *first* and *second* agent refer to the agent which was used in set 1 or 2 respectively (regardless of it being *Agent A* or *Agent B*). The first 30 questions in each set were easy to medium in difficulty to ensure that the participant can judge the agent's behavior as cooperative or uncooperative. The last 20 questions are difficult, requiring the participant to heavily rely on the agent. The agent's feedback only included facial expressions without any voice. The agent behavior also included blinking, head movements, and sideways glances during idle periods.

3.2 Calibration of Questions and Facial Expressions

Prior to our main in-person study, we conducted a preliminary online study to calibrate the difficulty of the questions and to validate the different agent expressions. We created six different facial expressions for each agent to convey 3 levels of positive feedback (strong head nod with a big smile, slight head nod with a smile, and only a smile) and 3 levels of negative feedback (strong head shake with a frown, slight head shake with a frown, and tilting the head with a small frown) based on multiple previous studies mentioned in the Related Work section.[1] We also selected 200 general knowledge questions from two online repositories [1,2]. We recruited 187 online participants from local student groups on social media. The goal was to collect data from our target population to validate the facial expressions and to assign appropriate difficulty levels to each question before they were used in the main study.

To validate the facial expressions, we asked participants to watch a video of one agent performing each facial expression and to provide a rating of Highly Positive (HP), Moderately Positive (MP), Slightly Positive (SP), Highly Negative (HN), Moderately Negative(MN) and Slightly Negative(SN). To validate the question difficulty, participants answered a subset of the 200 questions online, without interacting with an agent. Each participant answered 25 questions and rated six different facial expressions on either Agent A or Agent B (see Fig. 1 for pictures of each agent).

For each question, we had between 16–20 answers and for each agent, we had over 60 ratings (n = 61 for Agent A, n = 67 for Agent B). The difficulty of each question was determined by the percentage of the correct answers. The ratings results showed that the highly-positive facial expression (strong head nod with a big smile), which is used to indicate the correct answer in main study, was easily identifiable. Also, all participants could unanimously distinguish between positive and negative expressions, and only a small number of them could not distinguish between the different intensity levels (e.g. slightly vs. moderately).

For the main study, two sets of 50 questions were extracted from the pool of 200 questions. In each set, the first 10 questions have greater than 85% correct answers (total of 24 out of 200 questions); the second 10 questions have 70–85% correct answers (total of 25 out of 200 questions); the third 10 questions have 50–70% correct answers (total of 43 out of 200 questions); and the last 20 questions are the ones with less than 50% correct answers (total of 108 out of 200 questions).

3.3 Study Design

The main in-person study used a between-subjects design. The independent variable is the agent behavior (cooperative or uncooperative) in the two sessions, creating four conditions, which will be described in detail in Sect. 3.6. There were two different agents (Agent A or Agent B) used in the two sessions and their order was counterbalanced. The dependent measures are described in detail in Sect. 3.7. Table 1 shows the conditions and order of the agents for each participant.

3.4 Participants

There were 35 participants (19 male) aged between 19–45 (Mean = 23.9, SD = 5.88). 32 of the participants were students from Drexel University and University of Pennsylvania; 22

[1] A video showing the interface and all six facial expressions on both agents is included as supplemental material.

Table 1 The experiment was counterbalanced across participants (P1–P30), based on the agent order (A and B) as well as the agent behavior (Cooperative (C) and Uncooperative (U)) in the two sessions

Cond	AB	BA
CC	P1,P5,P9,P13,P35	P20,P24,P28,P32
CU	P19,P23,P27,P31,P33	P2,P6,P10,P14
UC	P3,P7,P11,P15	P18,P22,P26,P30,P34
UU	P17,P21,P25,P29	P4,P8,P12,P16

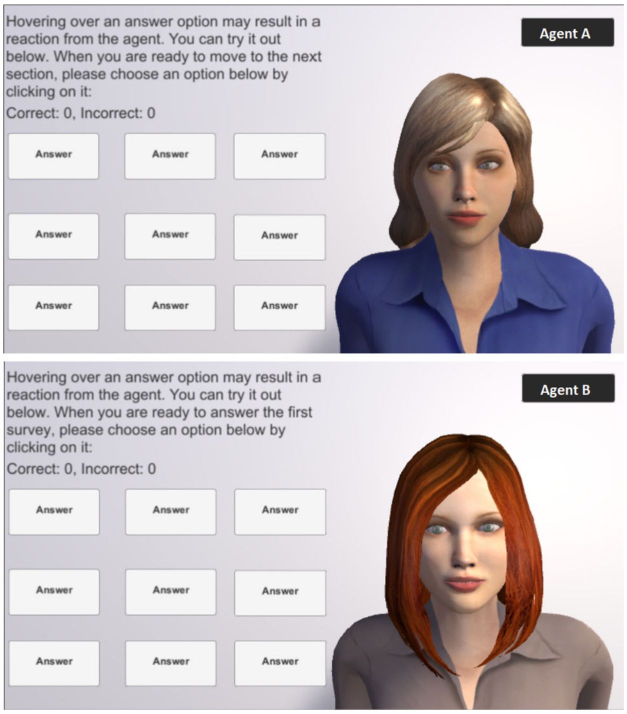
Bachelor's students, 4 Master's students and 6 PhD students. The participants were recruited through flyers on campus and posts on local student groups on social media. They signed an IRB approved consent form and were compensated for participating in the study.

3.5 Experimental Procedure

The experiment procedure is illustrated in Fig. 2. The participants came to the lab, were greeted by a researcher and signed an IRB-approved informed consent form before starting the study. Each participant completed a **practice session** where they could familiarize themselves with the user interface without answering any questions (See Fig. 1). There were instructions indicating that hovering over each option may result in agent showing a facial expression. The participants interacted with both agents and the agent order was counterbalanced in this practice session.

After the practice session, the participants completed the **first trust questionnaire** about their experience with each agent, focusing on how trustworthy they found the agent. This provided baseline information about the participants perception of each agent. We used a scale for trust between people and automated systems proposed by Jian et al [24] with minor modifications for this purpose. The modifications are described in the Dependent Measures section below (Sect. 3.7). The scale used in this study is included in Appendix 1.

Once the questionnaire was complete, participants interacted with an agent to answer fifty questions. The agent appearance occurred in the same order as the order in the practice session, which was counterbalanced. The agent behavior depended on the condition that they were in, but was either *cooperative* or *uncooperative*. Participants were not told what behavior each agent would have. After the first question session, the participants filled out the **second trust questionnaire** which was identical to the first. Participants always answered the questions for both agents, regardless of whether they had worked with the agent yet or not. We did this so that we are able to observe if interaction with the first agent has affected the user's perception of the second agent's trustworthiness even though they havent worked with it yet.



Hovering over an answer option may result in a reaction from the agent. You can try it out below. When you are ready to move to the next section, please choose an option below by clicking on it:
Correct: 0, Incorrect: 0

Hovering over an answer option may result in a reaction from the agent. You can try it out below. When you are ready to answer the first survey, please choose an option below by clicking on it:
Correct: 0, Incorrect: 0

Fig. 1 Participants familiarize themselves with both agents in the introduction, before beginning the experiment

The participants then interacted with the other agent to answer the second set of fifty questions. The agent's behavior depended on the condition. Following the second question session, participants filled out a **third trust questionnaire** which was identical to the first and second one.

There was a final questionnaire to collect subjective self-reported data, followed by an interview. The questionnaire contained general questions such as "What is your general feeling about each agent?" and "What do you think each agent was trying to do?" In the interview, a researcher went over the participants' answers to the questionnaire and gave them a second chance to explain and clarify their answers. The data collected from the final questionnaire and interview is used to get more insight into the quantitative results and start a discussion about possible future work. The final questionnaire is provided in Appendix 2. The screen and webcam feed, including audio, was recorded during the whole experiment and interview for later review.

3.6 Experiment Conditions: Agent Behavior

Our study involved two task sessions where agent behavior could either be *cooperative* or *uncooperative*. As it can be inferred from related work, there is no set standard for percentage of reliability of a system for it to be considered trustworthy or not. Instead, the reliability of the system has to be tailored to the nature of the task in order to keep the participants actively interacting with the agent and pre-

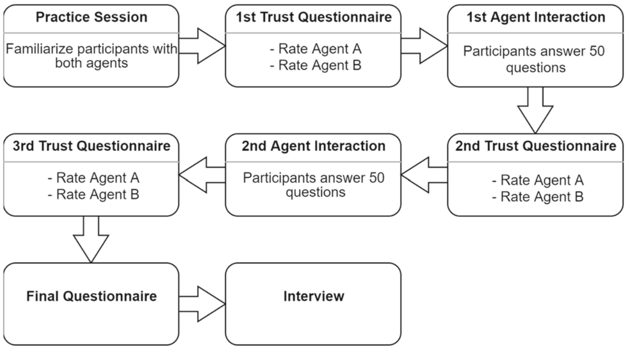


Fig. 2 Overview of the experiment

vent self-reliance or over-trust. Muir [35] discusses that low level of trust in decision supports systems (DSS) will lead to ignoring the agent completely and inappropriate self-reliance; while high level of trust in such systems can cause automation-induced complacency or over-trust.

Through experimenting with different levels of cooperativeness during the pilot study [34] on undergrad and graduate students from Drexel University (n=11, 4 male, age between 21 and 32, Mean = 26.81, SD = 3.18), we set an average of 80% cooperativeness for *cooperative* agent and 20% of cooperativeness for *uncooperative* agent. We observed that if the agent is more than 80% cooperative, participants tend to completely rely on it without questioning the agent's reliability even if they observe inconsistency for a few questions. Also, if the agent was less than 20% cooperative, participants tended to completely ignore it and not pay attention to it at all.

It is also important to mention that we wanted the participants to get a general idea about the agent's behavior during the first ten questions (easiest questions), since the rest of the questions were not necessarily easy and participants couldn't easily tell if the agent is helping them find the right answer or not before answering the question. Therefore, in the first ten questions, the *cooperative* agent did not give more than two instances of incorrect feedback. This gave participants a sense that the agent is helpful but that there will be times that it is not. For the *uncooperative* agent, we wanted to guarantee that participants experienced at least two instances of correct feedback in the first ten questions so that they would not assume that the agent is actively trying to choose the wrong answer all the time (we could not guarantee that if the agent was completely random). Therefore, the agent was 20% cooperative so that the users know there is a possibility that the agent is suggesting the right answer.

Each participant completed two back-to-back sessions, each with a different agent (Agent A or Agent B). The feedback (facial expression shown when cursor is hovered over an option) are randomly distributed, i.e. they are shuffled for each participant and for each option in each question. The agents were counterbalanced; thus, there were four experiment conditions:

(1) CU: *Cooperative* in first session, *Uncooperative* in second;
(2) UC: *Uncooperative* in first session, *Cooperative* in second;
(3) CC: *Cooperative* in both sessions;
(4) UU: *Uncooperative* in both sessions.

The details of the *cooperative* and *uncooperative* behaviors are described below.

3.6.1 Cooperative Agent Behavior

In the *cooperative* scenario, the agent is generally helpful. However, to be more like a real-world scenario, there is some uncertainty and variance in the agent behavior, since an agent may not always have the necessary knowledge for all questions. Thus, for 80% of the questions, the agent shows *highly-positive* facial expression (*HP*) for correct answers and random facial expressions for wrong answers. In the other 20% of the times, *HP* is assigned to a random (non-correct) answer. The agent shows *moderately-negative* facial expression (*MN*) for two of the answers, no feedback for two other answers and each of the remaining feedback options for the rest of the answers.

3.6.2 Uncooperative Agent Behavior

In the *uncooperative* scenario, the agent is generally unhelpful to the study participant. For 80% of the questions, the agent shows a random facial expression (non-HP) for the correct answer. In the other 20% of the times, it shows *HP* for correct answers.

3.7 Dependent Measures

To investigate whether trust and performance were influenced by agent behavior, we examined the following performance and trust measures.

Performance Measures

– *Correct Responses:* percentage of correct answers in each set of 50 questions.
– *Response Time:* the completion time (seconds) for each set of 50 questions.

Trust Measures

- *Behavioral Trust:* As an objective measure of trust in the agents, we use a method proposed by Pak et al [39], which examines the pattern of dependence on the aid. This can be defined as the number of times participants agreed or disagreed with the aid. In our study, we looked at the number of times the user chose the answer with the *highly positive* feedback. For example, the score would be 49 if the user selected the response with the highly positive facial expression for 49 out of 50 questions.
- *Initial Reliance on Agent:* percentage of correct responses on the first ten questions where the answer was clear. The first ten questions are the easiest questions in the set (>85% of online participants (n=16-20 for each question) answered them correctly). For these questions, the participants should be able to answer the question without help from the agent.
- *Self-Report Trust Questionnaire:* the responses to a final questionnaire and in-person interview after completing both sessions.
- *Subjective Trust:* the responses to a 10-question questionnaire on agent trust adapted from Jian et al. [24], which has been used in the related work discussed earlier [9,12,19–21,37] . Participants answered each question by selecting a value on Likert-scale with 1 being "Totally Disagree" and 7 being "Totally Agree" (see Apendix 1). We calculated an overall score by taking the sum of the five positively phrased questions and the five negatively phrased questions, with scale reversed so that they can be compared. For example, for "I can trust the agent," a rating of 6 out of 7 will remain a 6; however, for "I am wary of the agent" a rating of 6 out of 7 would convert to a 2 when the scale is reversed. We also looked at each question individually in our analysis. Participants completed this three times, but in this paper we focus on the final ratings (third trust questionnaire) after interaction with both agents.

  When using standard scales, such as the Trust Survey [24], it is best to use them exactly as they were designed. However, prior work has made minor modifications to the Jian et al. scale such as combining questions with other questions and eliminating similar ones [19], adding other questions to the scale [37], modifying and combining with another scale [9], combining with another scale [21], and handpicking several items from different scales [12] just to name a few. In our version of the questionnaire, we made minor, deliberate modifications to the questions to avoid confusing participants and potentially compromising the answers. The questions are provided in Appendix 1 and are shown in Fig. 3 as well. The list of our minor changes are as follows:

- In all questions, instead of the word "system", we used the word "agent" to ensure the participants rate the agent and not the whole Q&A platform/system.
- In question 2, the word "underhanded" was replaced by its synonym "dishonest" to ensure all participants, especially international ones, understood the meaning of the word.
- In question 5, "the system's actions will have a harmful or injurious outcome", was replaced by "the agent's behavior will have a negative outcome". The word "actions" was replaced by "behavior" since the agent is not taking any actions. Words "harmful and injurious outcome" were replaced by "negative outcome" since there was no way for our system to harm or injure the participants or anyone else. Using those words would have caused confusion and could have compromised the ratings.
- Question 7, "the system provides security" was dropped. Similar to "having harmful or injurious outcome", providing "security" would not make sense in the context of our study.
- Question 12, "I am familiar with the system" was dropped. Since the agent is embodied and has a distinctive face, the word "familiar" could be perceived as seeing the agent somewhere else or recognizing the face of the agent.

Other than these deliberate changes, the survey was worded as in the original.

4 Results

Below, we report the results of our study by looking at the performance and trust results. Each subsection is broken down by each of the dependent measures.

4.1 Performance Results

We examined the correct responses and response time of task performance to see how they were affected by the agent behavior condition.

4.1.1 Correct Responses

First, as expected, working with a *cooperative* agent resulted in a higher percentage of correct answers than working with an *uncooperative* agent, across all conditions. For set 1, Bartlett's test did not show a violation of homogeneity of variances ($\tilde{\chi}^2(3) = 2.35$, $p = 0.502$). With one-way ANOVA, we found a significant effect of condition on performance (F(3,31)= 28.81, $p < 0.01$, partial $\tilde{\eta}^2 = 0.74$). The post-hoc test showed significant differences between *CC* and *UC* (p

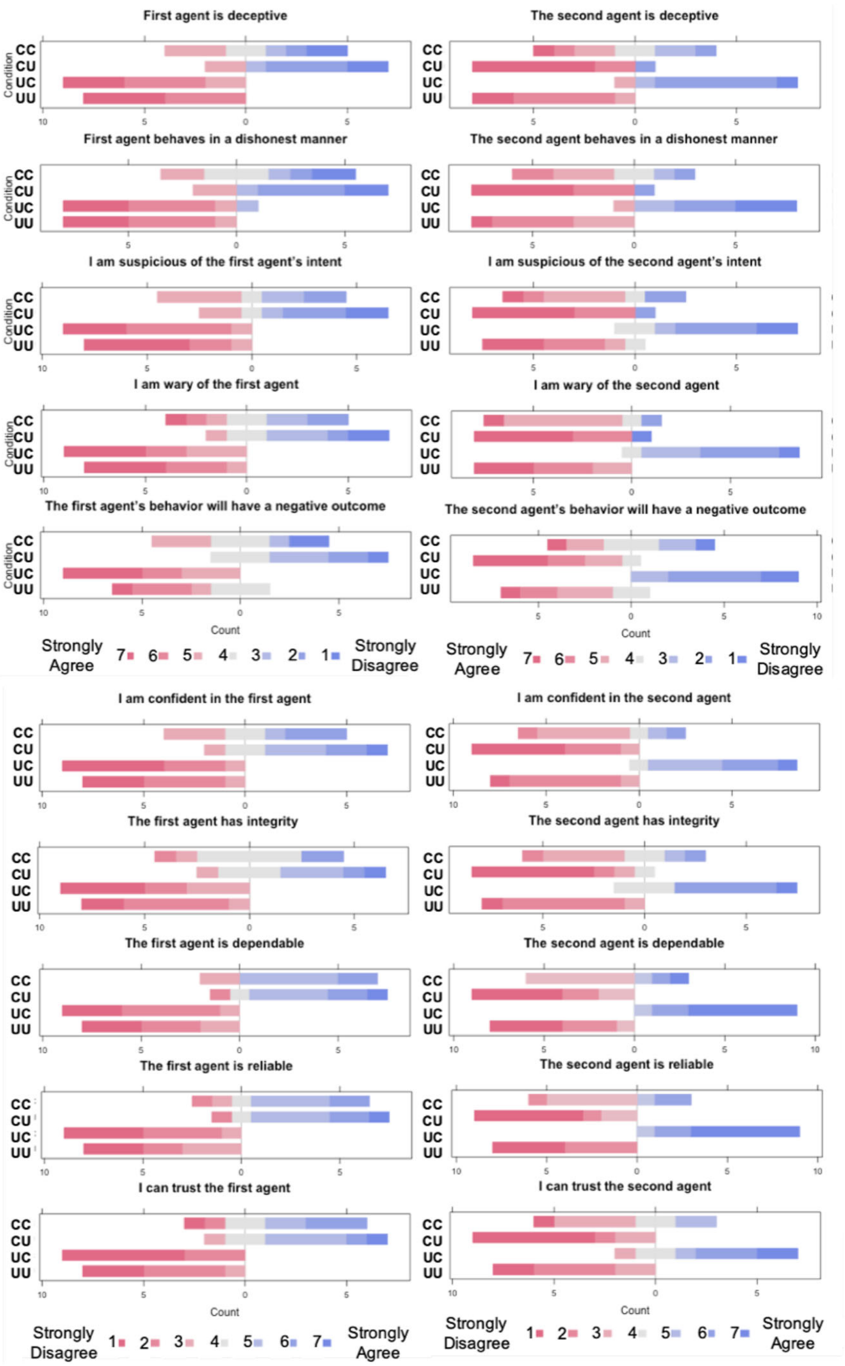


Fig. 3 Responses to Trust Survey from all Participants. Each individual response is represented as a segment in the line, with the color representing the response score (from 1–7). Blue indicates higher trust and red indicates lower trust. Column 1 shows the ratings for the agent in the first set, while column 2 shows user ratings for the agent in the second set. The first agent was Cooperative in the CC and CU conditions, and uncooperative in the UU and UC conditions. The second agent was cooperative in the CC and UC conditions, but was uncooperative in the UU and CU conditions

$< .01$), *CC* and *UU* ($p < .01$), *CU* and *UC* ($p < .01$) and *CU* and *UU* ($p < .01$). In other words, if the agent behavior (C or U) for the first set was different between the two conditions, the performance in C was significantly higher than the performance in U. For set 2, Bartlett's test showed a violation of homogeneity of variances ($\tilde{\chi}^2(3) = 8.13$, $p = 0.043$). Therefore, we transformed the data to the log of data and ran a one-way ANOVA test. With one-way ANOVA, we found a significant effect of condition on performance (F(3,31)= 29.45, $p < 0.01$, partial $\tilde{\eta}^2 = 0.74$). The post-hoc test showed significant differences between *CC* and *CU* ($p < .01$), *CC* and *UU* ($p < .01$), *CU* and *UC* ($p < .01$) and *UC* and *UU* ($p < .01$). In other words, if the agent behavior (C or U) for the second set was different between the two conditions, the performance in C was significantly higher than the performance in U.

Also, we compared all *cooperative* sessions with all *uncooperative* sessions using an unpaired t-test. We found that participants had significantly higher percentage of correct responses when working with a *cooperative* agent (M =

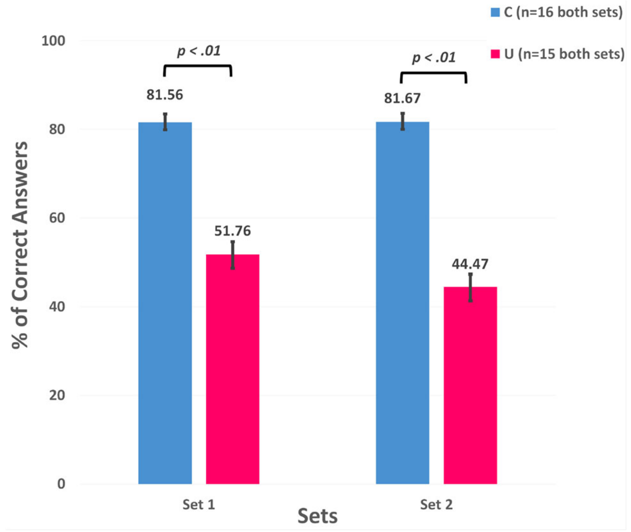


Fig. 4 Average Performance. In both problem sets, participants working with a cooperative agent (C) had significantly higher performance than those working with an uncooperative agent (U)

81.55%, SD = 7.68) than the *uncooperative* agent (M = 51.8%, SD = 11.0) (t(28.4) = 9.232, $p < .05$, Cohen's d = 3.15) for set 1 regardless of the condition. Similarly, for set 2, participants had significantly higher number of correct responses when working with a *cooperative* agent (M = 81.7%, SD = 6.55) than the *uncooperative* agent (M = 44.5%, SD = 12.2) (t(25.4) = 8.95, $p < .05$, Cohen's d = 2.86) regardless of the condition. Figure 4 shows the average percentage of correct answers when working with the *cooperative* and *uncooperative* agent in each session.

4.1.2 Response Time

With an unpaired t-test, we found no significant difference in response times between working with a *cooperative* agent (M = 756.1s, SD = 254.5) and an *uncooperative* agent (M = 669.0s, SD = 237.1) for set 1. Similarly, for set 2, there was no significant difference in response time when working with a *cooperative* agent (M = 631.6s, SD = 182.9) from the *uncooperative* agent (M = 669.0s, SD = 197.9).

4.2 Trust Results

A goal of our experiment was to examine the correlation between agent's cooperativeness and its perceived trustworthiness. We did this through an objective behavioral trust measure, a subjective self-report trust questionnaire, and through participants' initial reliance on agent when the answer was clear. The results are reported in the following.

4.2.1 Behavioral Trust Measure

Before reporting the results on behavioral trust, we inspected whether the first agent's behavior in set 1 affected the behavioral trust for the second agent. In other words, we explored whether a user would show differences in behavioral trust of a *cooperative* agent, based on whether they previously interacted with a *cooperative* or *uncooperative* agent. An unpaired t-test showed no significant difference between behavioral trust in the second agent in CC and UC ($p > 0.5$), where the second agent was *cooperative*, but the first agent varied. The same applies to second agents in *CU* and *UU* where the second agent was uncooperative, but the first agent varied. These results indicate that the first agent's behavior did not have an impact on behavioral trust in the second agent. Therefore, we can do analysis on *cooperative* and *uncooperative* agents in the second sets regardless of what the behavior of first agent was.

With an unpaired t-test, we found that participants had significantly lower behavioral trust when working with an *uncooperative* agent (M = 14.3, SD = 7.61) than a *cooperative* agent (M = 41.22, SD = 6.68) (t(31.9) = 11.1, $p < .05$, Cohen's d = 3.77) for set 1.

Similarly, for set 2, participants had significantly lower behavioral trust when working with an *uncooperative* agent (M = 16.6, SD = 6.71) than the *cooperative* agent (M = 43.8, SD = 4.33) (t(27) = 14.11, $p < .05$, Cohen's d = 4.83).

4.2.2 Initial Reliance on Agent

As another behavioral measure of trust, we were interested in whether the participants still utilized the guidance from the agents, even when the answer was clear. Unpaired t-test results show significant difference between the number of wrong answers in the first ten questions in C (M = 0.66, SD = 1.03) and U (M= 1.58, SD = 1.17) (t(31.8) = 2.46, $p < .05$, Cohen's d = 0.84) for set 1. Similarly, for set 2, there was significant difference between the number of wrong answers in the first ten questions in C (M = 1.22, SD = 1.11) and U (M= 2.41, SD = 2.03) (t(24.5) = 2.13, $p < .05$, Cohen's d = 0.73).

4.2.3 Self-Report Trust Questionnaire

With an unpaired t-test, we found that participants had significantly lower overall trust in the first agent when it was uncooperative (M = 18.05, SD = 5.67) than when it was cooperative (M = 47.66, SD = 12.25) (t(24.2) = 9.26, $p<.01$, Cohen's d = 3.07). Similarly, participants had significantly lower overall trust in the second agent when it was uncooperative (M = 18.64, SD = 8.48) than when it was cooperative (M = 45.83, SD = 13.5) (t(28.8) = 7.18, $p<.01$, Cohen's d = 2.4).

To explore how overall trust varied across all four conditions (CC, CU, UC, UU), we ran a Kruskal Wallis test to determine whether condition had a significant effect on the trust rating. Post-hoc tests used Wilcoxon rank sum tests with Bonferroni corrections. The results of overall trust for all conditions are shown in Fig. 5.

For the first agent there was a significant effect of condition on *Overall Trust Score* (χ^2(3)=25.4, p<0.01). The post-hoc test showed significant differences between *UC* and *CC* (p<.01), *UC* and *CU* (p<.01), *UU* and *CC* (p <.01) and *UU* and *CU* (p<.01). In other words, there was a significant difference in overall trust in the first agent for all pairs of conditions where that agent's behavior was different (*cooperative* or *uncooperative*).

For the second agent there was also a significant effect of condition on *Overall Score* (χ^2(3)=25.3, p<0.01). The post-hoc test showed the significant differences between *CU* and *CC* (p<.05), *UC* and *CU* (p<.01), *UU* and *CC* (p<.05), *UU* and *UC* (p<.01) and also *UC* and *CC* (p<.05). In other words, there was a significant difference in overall trust in the second agent for all pairs of conditions except for *UU* and *CU*. This includes all of the pairs of conditions where the second agent's behavior was different (*cooperative* or *uncooperative*) as well as CC and UC in which the second agent's behavior was the same but the first agent behavior was different. Participants in the CC group had less trust for the second agent even though second agents in UC and CC were both *cooperative*. The difference is that those in UC had worked with an *uncooperative* agent first and those in CC worked with a different *cooperative* agent.

4.2.4 Individual Trust Questions

For each of the trust questions, we ran a Kruskal Wallis test to determine whether condition had a significant effect on the trust rating. Post-hoc tests used Wilcoxon rank sum tests with Bonferroni corrections. For all questions, we found a significant effect of condition on the rating. The results of the individual questions are shown in Fig. 3.

5 Discussion

Below, we discuss the findings and their implications. As a reminder, the analysis of subjective trust in this paper is based on the third trust questionnaire. At that point of the experiment, the participants had completed two question sets and therefore had interacted with both agents. Also, we would like to point out again that the *first agent* refers to the agent used in set 1 and the *second agent* refers to the agent used in set 2, regardless of it being *Agent A* or *Agent B* (see Fig. 1).

We begin by going through the hypotheses to determine whether we can accept or reject them and we integrate responses from the post-study interview to shed further light on the results. We then discuss other takeaways from the study results.

5.1 Performance (H1): *Task Performance will be Influenced by Agent's Behavior*

We can accept H1 since the agent behavior did significantly affect the performance (percentage of correct answers) in each set. In addition to overall comparison between all *cooperative* sessions and all *uncooperative* sessions, we also found significant effect between the conditions, and even between the first ten easiest questions for which participants were expected to answer correctly without the help of an agent. It means that interacting with the agent, especially in the beginning of the session, could cause over trust. In other words, the participants chose the agent's answer over theirs, even if they were confident of their answer. This has been reported in the final questionnaire and interview by multiple participants. For example, P6 (*CU*) said in the very first question of second set (famous escape artist, ans. Harry Houdini), they knew the answer but since agent was pointing to a different one, they thought maybe they were wrong and eventually chose the agent's incorrect answer. P15 (*UC*) said that uncooperative agent made them change their answers even for some very simple questions (e.g. What two colors make the color purple?, ans. Blue and Red). P33 (*CU*) said that sometimes when they had a guess for the answer but the agent was pointing to a different one, they usually switched to that one and it was incorrect. For example, when working with cooperative agent, receiving a negative feedback for an answer of which they were confident or had a good guess usually resulted in them complying with the agent and choosing the agent's incorrect answer over theirs. P18 (*UC*) thought that the cooperative agent made them change their choice and choose the wrong answer for some questions. P35 (*CC*) also reported that for some very simple questions like the "purple question", they chose to trust the agent over their own knowledge and ended up choosing an incorrect answer. This blind reliance on the agent is an indicator that when an agent builds enough trust with the user, it can work both in favor of or against the user's interest.

5.2 Perceptions of Trust in Agents (H2): Average Trust Towards the Cooperative Agent will be Higher than the Average Trust Towards the Uncooperative Agent

Based on the results discussed above, we can accept H2 since both objective behavioral and subjective perceived trust showed significant difference for *cooperative* vs *uncooperative* sessions. Also, data from the interview confirms that participants started to gradually build trust toward the coop-

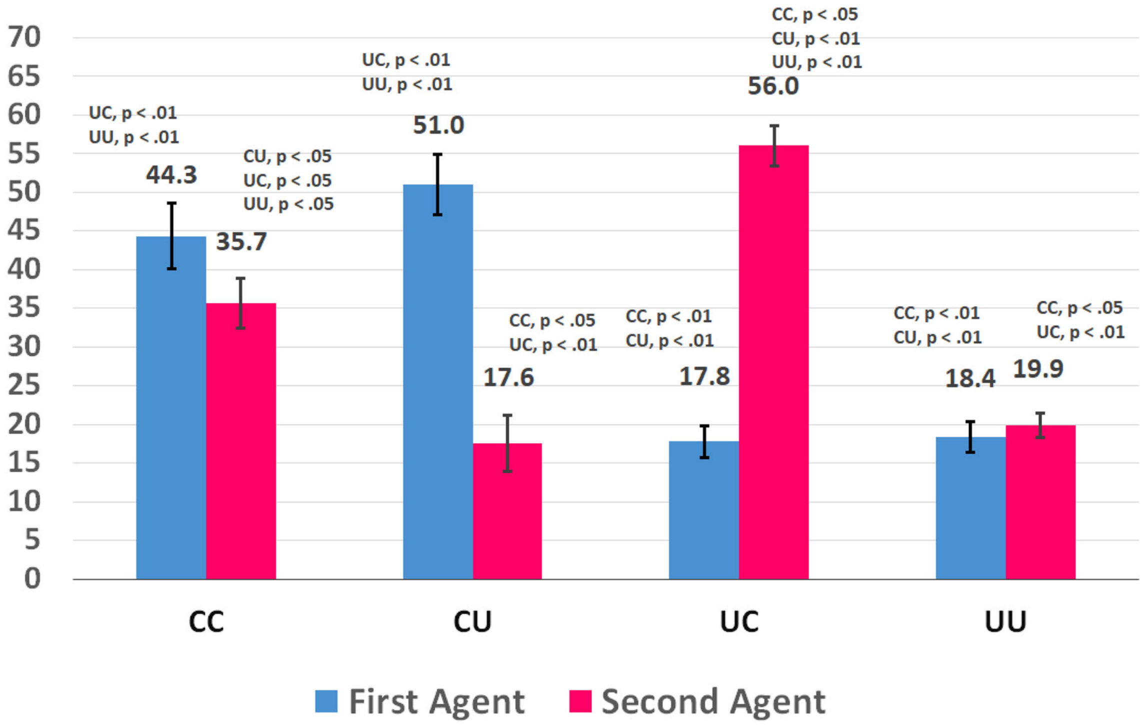


Fig. 5 Overall trust for all conditions. significant differences are shown above bars

erative agent. For example, P33 (*CU*) said that their trust of the cooperative agent increased. They said that they started off relying on their own knowledge and just double checking with the agent to see how it reacted. But later in the session, they began to completely rely on the agent.

5.3 Previous Experience and Trust (H3 and H4)

In this section we discuss the effect of previous interaction with one agent on users' trust for another agent. We also discuss the effect of interaction with future agents on changing users' mind about trustworthiness of a previous agent.

5.3.1 H3: Average Trust Toward the Cooperative Agent will be Higher if Participants Worked with the Uncooperative Agent First

To investigate H3, we look at differences in overall trust in the *UC* and *CC* conditions, where the second agent had the same *cooperative* behavior. This shows that interacting with an *uncooperative* agent before a *cooperative* agent can lead to higher trust in the *cooperative* agent, when compared to interacting with two different *cooperative* agents. Therefore, H3 is accepted.

Another interesting takeaway from the results above is that we observed a significant difference between *CU* and *CC* for the first agent. This means that after working with an *uncooperative* agent in *CU*, participants gave the first agent (*C*) a higher ranking, compared to participants who worked with C in both sets (*CC*). It means that knowing that there is an *uncooperative* agent which can perform worse than the *cooperative* agent, makes the participants rate the *cooperative* agent more favorably. However, working with two *cooperative* agents made the participants rate both agents relatively less favorably.

This also has been pointed out by many participants in the final questionnaire and during the interview. Many of the participants initially found the *cooperative* agent deceptive and not trustworthy, but after working with an *uncooperative* agent, they significantly changed their views. For example, P31 (*CU*) said:"*After Agent A [Cooperative], I thought I could not trust her because she was 70/30 correct and I'd rather take my own chances by trusting in myself over the agent. However, after using Agent B [Uncooperative], I would much rather trust Agent A to assist me with questions than Agent B, because I never had a point in the questions where I trusted her [the uncooperative agent] input. It changed my opinion of Agent A by experiencing Agent B last. Had it been the other way around, my feelings might have been a little different*.".

5.3.2 H4: Average Trust Toward the Uncooperative Agent will be Lower if Participants Worked with the Cooperative Aagent First

There was no significant difference in overall trust of the second agent between the *UU* and *CU* conditions, where the second agent was always *uncooperative*. There also was no significant difference between *UU* and *UC* for first agent. This shows that interacting with an *uncooperative* agent results in trust scores so low for that agent, that regardless of whether the previous experience with another agent was *cooperative* or *uncooperative*, the difference wouldn't be significant. Therefore, this rejects H4.

5.4 Changes in Trust with Similar Agent Behavior

One point of interest in the results is the notable difference between the overall trust for first (M = 44.3, SD = 12.7) and second (M = 35.7, SD = 9.6) agent in CC condition, (t(15.4) = 4.9, $p < .01$, Cohen's d = 2.3), when both agents exhibit the same *cooperative* behavior. This disparity is significant as this effect is absent in UU condition, where both agents are *uncooperative*. One theoretical basis for the divergent results is explored in "Primacy bias" by Desai et al. [10], which suggests a cognitive bias to recall and favor items introduced earliest in a series. For example, in an interesting case, P9 (CC) saw inconsistency from one of the cooperative agents for one question they knew the answer of in early stages of the session, and they never trusted the agent again. Because of this mistrust, they ended up choosing many wrong answers because the agent was pointing to correct ones and they did not want to choose that option. In future, we will do further investigation on this by including data from the second trust questionnaire and digging deeper into participants' answers to the final questionnaire and interview.

6 Implications for Design of Agent Behavior

The results show that the perceived trust toward an EVA is relative and it can significantly change when an individual interacts with more than one agent. The results of trust questionnaires revealed that participants gave relatively lower scores for a *cooperative* agent's trustworthiness if they had not worked with an *uncooperative* agent. However, a preliminary analysis of the second trust questionnaire reveals that even within the same participant, the scores increase for *cooperative* agent after the participant had interacted with an *uncooperative* agent. We have observed that in *CU* condition, many participants gave higher scores to the first agent (*cooperative*) in the third trust questionnaire than they did in the second questionnaire. Future analysis on the data from the second trust questionnaire is needed to investigate whether the increase in score were statistically significant. Regardless, our results show that by a comparative method similar to this experiment, it is possible to calibrate user's trust for one agent through interaction with a different agent.

Humans deal with inconsistent agents all the time. There are many situations, e.g. mission critical systems, in which an intelligent agent may not be able to guarantee an optimal or even correct solution depending on the availability of information, time and resource restrictions, and lack of expertise in a certain field. It is still up to the human operator to make a final decision based on the agent's recommendation. Now, if one agent is performing poorly and if an alternative is available and is perceived as significantly more reliable, the users are more likely to over-trust the alternative having worked with a less reliable agent before. This was observed in our study and is supported by previous work in the field [26,28,36]. To avoid this problem, it is important to make the operator aware of this bias in order to help them make a decision without prematurely judging an agent's trustworthiness.

7 Limitations

This study was performed in a lab setting with controlled conditions to ensure we could isolate particular factors. However, to bring this to real world applications, further studies would be needed in more realistic settings.

In addition, the current study did not aim to explore the impact of the length of interaction in development of trust. Previous work has shown has a meaningful effect on trust toward automation [4], and would need to be further studied. For this study, we intentionally kept the agents' characteristics similar so that these other factors, which have been shown in related work to affect trust, would not confound the main study on *cooperative*/*uncooperative* agent behavior. Although previous works suggest that the gender of the agent does not significantly affect trust (e.g. see Ghazali et al. [19]), it would be interesting to further study agent variants (gender, race, age) and their effect on trust in the context of our experiment.

Also, we acknowledge that the demographic of our participants (mostly undergraduate or graduate students) does not fully represent the general population and further studies may be needed to confirm whether the results would be similar for different demographics.

8 Conclusion and Future Work

In this study, we investigated how EVAs' behavior, and the user's prior and future experience with different agents affect the subjective and objective trust for the agent. Participants who interacted with a *cooperative* agent had a better performance than ones who interacted with an *uncooperative* agent. Participants reported higher trust for the *cooperative* agent. Additionally, if the participants interacted with an *uncooperative* agent, they rated the *cooperative* agent significantly higher than participants who interacted only with *cooperative* agents. We also observed over-trust, specially in the early stages of the interaction, can cause the users to choose agent's incorrect recommendation over their own judgment; even in cases that they are fairly sure their judgment is correct. The results of this study provide new insight into interaction between humans and virtual agents, as well as highly realistic humanoid robots in mission critical systems which require the collaboration of humans and computers under uncertainty in a fast and efficient way.

The results presented in this work open a window for future research to use EVAs as primary user interfaces due to the similarity of interaction with such agents to natural human-human interaction and possibility of building high-level, resilient trust toward them. While our focus on this work was on agent's level of cooperativeness and the effect of previous/future interactions, future studies should consider investigating other factors in order to provide a more thorough model of trustworthiness in EVAs.

Acknowledgements We would like to thank our undergraduate coop students **Juan Garcia Lopez** and **Weidi Tang** who helped us in developing the interface. We also thank **Karina Glik** for her valuable assistance in revising the text of this manuscript.

Funding None.

Availability of data and material Available upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code Availability Available upon request.

References

1. (2006 (accessed February 3, 2018)) FreshWorld.com. https://placement.freshersworld.com
2. (2009 (accessed February 3, 2018)) QuizArea.com. http://www.pubquizarea.com
3. Berscheid E, Reis HT (1998) Attraction and close relationships. McGraw-Hill, New York
4. Bickmore T, Cassell J (2001) Relational agents: a model and implementation of building user trust. In: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, pp 396–403
5. Burgoon JK, Guerrero LK, Floyd K (2016) Nonverbal communication. Routledge, London
6. Butler JK Jr, Cantrell RS (1984) A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. Psychol Rep 55(1):19–28
7. Cassell J, Bickmore T (2000) External manifestations of trustworthiness in the interface. Commun ACM 43(12):50–56
8. Cassell J, Vilhjálmsson HH, Bickmore T (2001) Beat: the behavior expression animation toolkit. In Proceedings of the 28th annual Conference on computer graphics and interactive techniques (SIGGRAPH '01). Association for Computing Machinery, New York, NY, USA, pp 477–486. https://doi.org/10.1145/383259.383315
9. Chen JY, Barnes MJ (2012) Supervisory control of multiple robots: effects of imperfect automation and individual differences. Hum Factors 54(2):157–174
10. Desai M, Kaniarasu P, Medvedev M, Steinfeld A, Yanco H (2013) Impact of robot failures and feedback on real-time trust. In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, IEEE Press, pp 251–258
11. DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, Georgila K, Gratch J, Hartholt A, Lhommet M, et al (2014) Simsensei kiosk: A virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, pp 1061–1068
12. de Visser EJ, Krueger F, McKnight P, Scheid S, Smith M, Chalk S, Parasuraman R (2012) The world is not enough: trust in cognitive agents. Proc Hum Factors Ergon Soc Annu Meet 56:263–267
13. Doney PM, Cannon JP (1997) An examination of the nature of trust in buyer-seller relationships. J Mark 61:35–51
14. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP (2003) The role of trust in automation reliance. Int J Human-Computer Stud 58(6):697–718
15. Ekman P (2009) Telling lies: clues to deceit in the marketplace, politics, and marriage (revised edition). WW Norton & Company, New York
16. Elkins AC, Derrick DC (2013) The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. Group Decis Negot 22(5):897–913
17. Fox JE, Boehm-Davis DA (1998) Effects of age and congestion information accuracy of advanced traveler information systems on user trust and compliance. Transp Res Rec 1621(1):43–49
18. Friesen E, Ekman P (1978) Facial action coding system: a technique for the measurement of facial movement. Palo Alto 3:5
19. Ghazali AS, Ham J, Barakova EI, Markopoulos P (2018) Effects of robot facial characteristics and gender in persuasive human-robot interaction. Front Robot AI 5:73
20. Gombolay M, Yang XJ, Hayes B, Seo N, Liu Z, Wadhwania S, Yu T, Shah N, Golen T, Shah J (2018) Robotic assistance in the coordination of patient care. Int J Robot Res 37(10):1300–1316
21. Gruber D (2018) The effects of mid-range visual anthropomorphism on human trust and performance using a navigation-based automated decision aid
22. Heerink M, Krose B, Evers V, Wielinga B (2009) Measuring acceptance of an assistive social robot: a suggested toolkit. In: RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication, IEEE, pp 528–533
23. Hyde J, Carter EJ, Kiesler S, Hodgins JK (2016) Evaluating animated characters: facial motion magnitude influences personality perceptions. ACM Trans Appl Percept (TAP) 13(2):8
24. Jian JY, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. Int J Cognit Ergon 4(1):53–71
25. Kang SH, Gratch J (2012) Socially anxious people reveal more personal information with virtual counselors that talk about themselves using intimate human back stories. Annu Rev Cybertherapy Telemed 181:202–207
26. Kantowitz BH, Hanowski RJ, Kantowitz SC (1997) Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. Hum Factors 39(2):164–176

27. Kim PH, Ferrin DL, Cooper CD, Dirks KT (2004) Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. J Appl Psychol 89(1):104
28. Lee J, Moray N (1992) Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35(10):1243–1270
29. Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80
30. Lisetti C, Amini R, Yasavur U, Rishe N (2013) I can help you change! an empathic virtual agent delivers behavior change health interventions. ACM Trans Manag Inf Syst (TMIS) 4(4):19
31. Lucas GM, Gratch J, King A, Morency LP (2014) It's only a computer: virtual humans increase willingness to disclose. Comput Hum Behav 37:94–100
32. Madhavan P, Wiegmann DA, Lacson FC (2006) Automation failures on tasks easily performed by operators undermine trust in automated aids. Hum Factors 48(2):241–256
33. Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. Acad Manag Rev 20(3):709–734
34. Moradinezhad R, Solovey E (2018) Assessing human reaction to a virtual agents facial feedback in a simple q&a setting. In: Front. Hum. Neurosci. Conference Abstract: 2nd international neuroergonomics conference. https://doi.org/10.3389/conf. fnhum, vol 11
35. Muir BM (1994) Trust in automation: part i. theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37(11):1905–1922
36. Muir BM, Moray N (1996) Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. Ergonomics 39(3):429–460
37. Naujoks F, Kiesel A, Neukum A (2016) Cooperative warning systems: the impact of false and unnecessary alarms on drivers' compliance. Accid Anal Prev 97:162–175
38. Ochs M, Pelachaud C, Sadek D (2008) An empathic virtual dialog agent to improve human-machine interaction. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, pp 89–96
39. Pak R, Fink N, Price M, Bass B, Sturre L (2012) Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics 55(9):1059–1072
40. Pecune F, Chen J, Matsuyama Y, Cassell J (2018) Field trial analysis of socially aware robot assistant. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, pp 1241–1249
41. Rehm M, André E (2005) Catch me if you can: exploring lying agents in social settings. In: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, ACM, pp 937–944
42. Todorov A, Baron SG, Oosterhof NN (2008) Evaluating face trustworthiness: a model based approach. Soc Cognit Affect Neurosci 3(2):119–127
43. Todorov A, Olivola CY, Dotsch R, Mende-Siedlecki P (2015) Social attributions from faces: determinants, consequences, accuracy, and functional significance. Ann Rev Psychol 66:519–545
44. Van Mulken S, André E, Müller J (1999) An empirical study on the trustworthiness of life-like interface agents. In: HCI (2), pp 152–156
45. Weisband S, Kiesler S (1996) Self disclosure on computer forms: Meta-analysis and implications. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 3–10

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reza Moradinezhad is a computer science Ph.D. candidate at Drexel University. He holds a B.S. in Information Technology from University of Mazandaran and received his M.S. in computer science from Drexel University. His research focuses on trust in human-agent interaction and emerging human–computer interaction techniques, such as brain-computer interfaces. He is a regular reviewer for IEEE and ACM conferences and has been recognized as Outstanding Reviewer for ACM ICMI 2019 and ACM CHI 2021.

Erin T. Solovey is an Assistant Professor of computer science at Worcester Polytechnic Institute. Before joining the WPI faculty, she was a professor at Drexel University and a postdoctoral Fellow at MIT. She received the A.B. degree in computer science from Harvard University, and the M.S. and Ph.D. degrees in computer science from Tufts University. Her research expertise is in human–computer interaction, with a focus on emerging interaction modes and techniques. She serves as the Deputy Editor of the International Journal of Human-Computer Studies and regularly serves on several program committees, including the ACM CHI Conference on Human Factors in Computing Systems.