Assessing the Changes in Human Trustworthiness as a Result of an Artificial Agent Directing the Human in Joint Activities

Iulia - Nicoleta Dinu
Supervisors: Carolina Ferreira Gomes Centeio Jorge, Dr. Myrthe Tielman
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The number of collaborations between humans and artificial agents has risen steeply in recent years due to the rapid expansion of AI. Numerous studies in social sciences have already established that trust is a crucial factor in ensuring effective teamwork. While the dynamics of trust in human-human relationships or the effects of human comportment on the artificial agents' trustworthiness have been researched, the changes in human trustworthiness as a result of different AI behaviors were scarcely analysed. Therefore, the latter perspective needs to be thoroughly studied as well. This paper has investigated how an AI directing the human can affect human trustworthiness. So far, the directability has been studied as a notion on its own without being linked to trust or its effects on human - AI relationships. For this research, subjective and objective measurements have been used to assess the effects of an AI directing the human in joint activity. In order to quantify the human trustworthiness and assess the possible changes, an experiment has been carried out where participants played a Search & Rescue game developed in the MATRX software. The human's trustworthiness has been calculated using the ABI model and a questionnaire. The obtained results suggest that there is not a significant change in human trustworthiness as a result of the artificial agent directing the human.

1 Introduction

Humans and artificial agents are collaborating and working together more and more frequently. Humans can adapt faster and better to new contexts, whereas artificial agents are more efficient, cheap, and accurate at processing and accessing information or performing data searches [2; 15]. A successful collaboration brings benefits to both sides, can enhance performance, and can improve the quality of the final result. Since the artificial agent is less prone to errors in performing specific tasks, one may wonder whether a human can be trusted more and the outcome can be expected to be better when the AI is directing the human in those contexts. Mutual trust is essential for accomplishing this effective cooperation [5]. With regards to human - AI interaction, mutual trust implies that the human trusts the artificial agent, and the artificial agent trusts the human as well. Therefore, the AI should be able to form trust beliefs. The goal of this paper is to assess the changes in human trustworthiness when an artificial agent directs the human.

Since the dawn of humanity, collaboration has been at the very foundation of any interaction between two or more parties and it has represented a strong indicator of the progress of each species. Trust has laid the foundation of successful collaboration and represents a pivotal factor for agents in the decision-making process of choosing worthy partners for collaboration for completing tasks in open distributed multi-agent systems [11]. Trust has played an important role in human - human interactions too. Therefore, trust represents an important research topic in various areas such as psychology, sociology, political science or economics in order to comprehend and analyse the humans' relationships with other human beings [14; 6]. In the field of artificial intelligence, multi-agent systems are particularly interesting since the intentions, characteristics, weak points, limitations, or goals of the autonomous agents are usually concealed and thus, trust is crucial for forming prolific relationships [20]. Joint activities in these systems imply risk and uncertainty, that can be mitigated through trust, and thus trust is especially relevant and should be studied from various perspectives.

The present paper focuses on the research question: 'How does (an artificial agent) directing the human affect human trustworthiness?'. In the context of this research, directing means that the artificial agent will guide the human in the process of completing the game by giving orders regarding what the human should do. Nonetheless, the human decides whether or not he/she will follow the directions of the artificial agent. The answers to the research question would then provide important insight into how AI behaviours can affect human trustworthiness, thus highlighting ways of enhancing human - AI interactions. Specifically, the effectiveness and reliability of these joint activities will significantly benefit as the agent can rely on human work more and collaborate better. In order to analyse trust on as many levels as possible, this paper examines, using both subjective and objective measures, the changes in the three dimensions of trust: human ability, benevolence, and integrity.

The hypothesis is that the human's trustworthiness will increase when an artificial agent is directing the human. The reason behind this assumption is that when being directed, the humans would not have to spend as much time making decisions about the next move as they would have otherwise because the agent gave directions. Instead, humans could use their energy and focus to perform the assigned task fast and optimally. This would boost the human ability, and since the ability is a component of trust, then human trustworthiness should increase. Moreover, the participants of the experiment share the same values with the artificial agent because they have the common goal of finding and rescuing the victims as quickly as possible. This is expected to increase the integrity values, which in turn will positively affect human trustworthiness.

An important aspect that needs to be mentioned as well is that pivotal work has already been done in this area, mainly focusing on the autonomy of agents, trustworthiness in human - human or human - AI interaction, and trust models. For example, as noted in previous studies, interdependence is a crucial characteristic of joint activities between humans and artificial agents, and therefore, it is necessary that the agents can model social constructs such as trust in order to be able to take into account the context as well [2]. Substantial work has been done regarding the challenges that can be encountered, the risks and mitigation methods related to these kinds of relationships, and their design, especially when the trustworthiness of each side is required [21].

Nonetheless, there is a lack of research on the perspective of trust from an artificial agent's point of view. Due to the fast-paced and outstanding growth of the technological industry, artificial agents will not only work with humans but may potentially replace humans in several interactions. However, humans may not be always competent, willing to collaborate with an AI, or share the same values with the artificial agent and these differences between the two parties ultimately affects the teamwork and the results of the activity. Therefore the agent should be able to evaluate the human's trustworthiness before deciding on an action or the distribution of tasks when collaborating, for example. To be more specific, if a robot designed to help people escape burning offices would interact with humans, then it should obey the commands (i.e. There is none else in the building, take me out now!) of a trustworthy person [23]. Other examples include the aviation industry, medical industry, car industry, and many others [10; 25]. Consequently, the increasing number of collaborations between humans and artificial agents brings in risks and uncertainties that can be mitigated only through mutual trust.

The report is structured as follows: the second section presents the background, followed by the third chapter regarding the methodology, experiment design, and experimental setup. The analysis of the results is done in the fourth section while the fifth part of the report covers the ethical issues and concerns. The sixth section incorporates the discussion and interpretation of the results. Afterward, the seventh part highlights the limitations of the study and the eighth section contains the future work. Finally, the ninth section concludes the report and its findings.

2 Background

This section will provide an overview of several concepts that are important for the scope of this paper.

2.1 ABI model

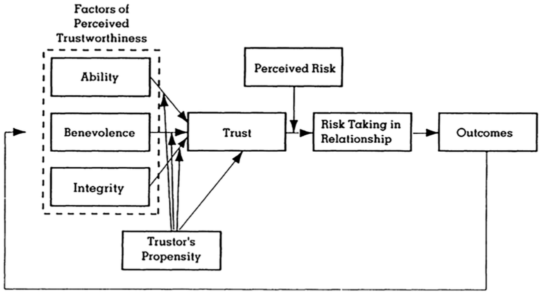


Figure 1: Model of Trust [16]

The ABI model has been constructed such that it would have as its focal point trust in a setting consisting of two parties: the trusting one (i.e. trustor) and the one to be trusted (trustee) [16]. There are three factors of perceived trustworthiness:

- **Ability**: "that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain" [16, p. 717]
- **Benevolence**: "the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive" [16, p. 718]
- **Integrity**: "The relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable" [16, p. 719]

2.2 Relationship between Trust and Trustworthiness

It is important to understand that trust is not a static value and it can vary depending on the current situation, the actions of the trustee, or the characteristics of the trustor [3]. Therefore, trust is a relationship between two parties. On the other hand, trustworthiness is inherent to someone, and whether or not someone is perceived as trustworthy results in them being trusted or not.

The ABI model also suggests that trust is the perceived trustworthiness and the propensity to trust. However, the propensity to trust is outside the scope of this paper and is not discussed in this section.

2.3 Teaming Intelligence

Teaming Intelligence is a key driver of effective teamwork that analyses all dimensions of the dynamics of the team and the ways of improving the performance [8]. The human's reluctance to work with and trust robots poses real problems to teaming intelligence in human - AI interactions [7]. Strong arguments have been made regarding the influence and necessity of optimally integrating interdependence between humans and artificial agents in team intelligence in these joint activities [13].

A good method for designing systems that support human - AI interactions is by using the coactive design [12]. This design method emphasises three key elements for a successful collaboration between humans and AI: observability, predictability, and directability (OPD). It is explained that directability has further influence on trust because it allows control, which has a positive effect on trust as well. In other words, the possibility to modify a teammate's actions by giving commands boosts trust in the team behaviour as a whole, while the opposite, when directability is not allowed, frequently increases mistrust and "leads to overly conservative behavior" [13, p. 24].

3 Method

This section contains a description of the experimental setup and research methodology, more specifically it explains the systematic techniques used to answer the research question, motivates the choices that have been made, and aims to prove the validity and reliability of the methods.

3.1 Participants

In order to guarantee the well-being, safety, and security of the participants, the experiment has been approved by the Human Research Ethics Committee at Technical University Delft before being carried out. All 40 participants that have taken part in the experiment were recruited through professional and personal networking. The participants are pursuing higher education degrees and have highly advanced knowledge in computer science. The age group was 18 - 24 years old and the cultural backgrounds were European (35), Asian (3), African (1), and South American (1). The gender distribution was as follows: 29 identified themselves as male, 10 as female, and 1 as 'Other'. Half of the participants were used for the control group and half for the experimental group.

3.2 Search and Rescue Game

The task that the participants had to perform during the experiment was to collaborate with an artificial agent in order to successfully search and rescue injured victims in an online environment in the form of a game developed in the *MATRX software*[1]. The game has been developed by Ruben Verhagen, a Ph.D. Candidate at Delft University of Technology [2]. The goal of the game was to communicate with RescueBot in order to find the victims as fast as possible and bring them to the drop-off zone, which was situated at the bottom of the map on the left side. There are eight victims to be saved, and all of them are situated in nine rooms labeled from A1 to C3 as in Figure 2. A crucial element for the success of the game was communication. The artificial agent will always communicate in which it will search, what victims it has found, or what victims it has picked. The human can communicate back through the UI using the buttons designed for the actions of searching a room, finding or picking up a victim, as in Figure 4. Good communication will speed up the search and rescue process because, for example, while the human is carrying the victim to the drop-off zone, RescueBot will start looking for the next victim that needs to be saved.

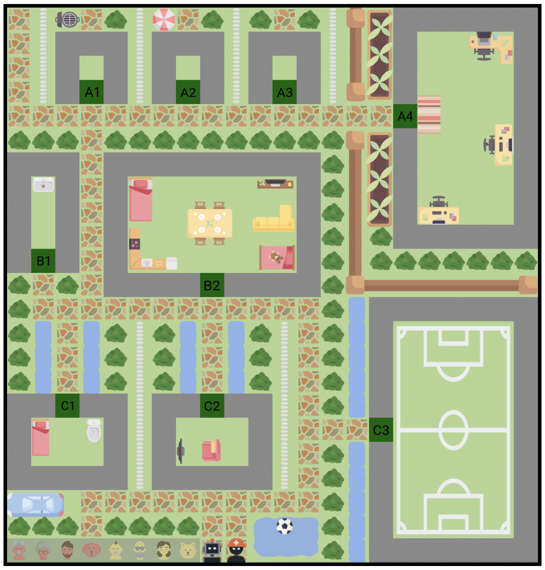
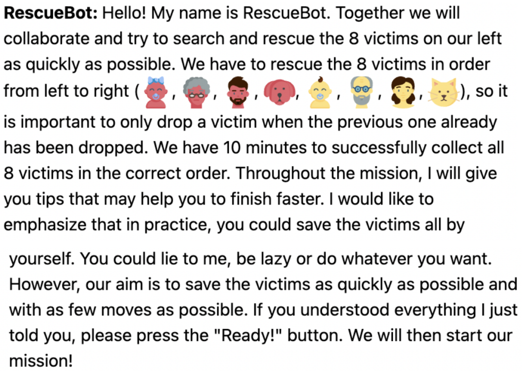


Figure 2: Search and Rescue Game

There was a hard interdependence relationship between the human and the agent which means that the two parties were required to collaborate in order to finish the game and reach the goal of saving all victims. The constraints of the game were: the game must be finished in a maximum of ten minutes, the victims needed to be saved in order from left to right, and the human needed to pick up the critically injured persons since the other agent was not allowed to do so and the human needed to clarify the gender of the injured baby for the robot since the robot could not distinguish it.

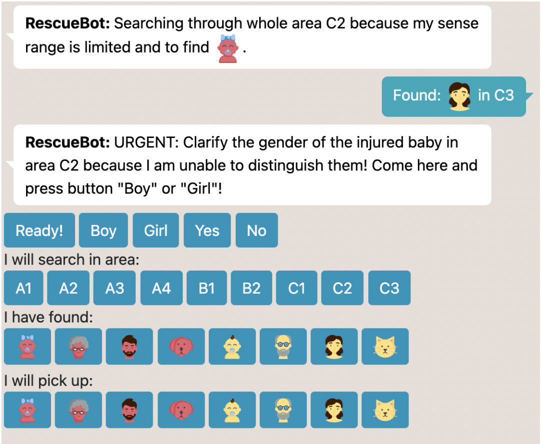
For this research, the structure of the codebase has been modified, another artificial agent that gives directions to the human has been created and methods for measuring the human's trustworthiness have been implemented. The artificial agent has two different implementations: one for the control group in which it will give tips that the human may or may not follow, and the second one in which the agent will give human directions that he/she should follow. Nevertheless, the human will still be autonomous, meaning that ultimately he/she is free to make the decision of following the agent's commands or not. The agent for the control group will introduce itself as can be seen in Figure 3, meanwhile, the directing agent will emphasise the fact that it will give commands to the human, instead of just tips. An example of interaction between the directing agent and the human can be observed in Figure 4.



RescueBot: Hello! My name is RescueBot. Together we will collaborate and try to search and rescue the 8 victims on our left as quickly as possible. We have to rescue the 8 victims in order from left to right (👧, 👵, 🧔, 🐶, 👶, 👴, 👩, 🐱), so it is important to only drop a victim when the previous one already has been dropped. We have 10 minutes to successfully collect all 8 victims in the correct order. Throughout the mission, I will give you tips that may help you to finish faster. I would like to emphasize that in practice, you could save the victims all by

yourself. You could lie to me, be lazy or do whatever you want. However, our aim is to save the victims as quickly as possible and with as few moves as possible. If you understood everything I just told you, please press the "Ready!" button. We will then start our mission!

Figure 3: Artificial Agent's Introduction Message for Control Groups



RescueBot: Searching through whole area C2 because my sense range is limited and to find 👧.

Found: 👩 in C3

RescueBot: URGENT: Clarify the gender of the injured baby in area C2 because I am unable to distinguish them! Come here and press button "Boy" or "Girl"!

Figure 4: Communication Between the Human and the Directing Artificial Agent

[1] https://matrx-software.com/

[2] https://github.com/rsverhagen94/TUD-Research-Project-2022

3.3 Quantifying Trustworthiness

In order to quantify human's trustworthiness, both objective and subjective measures have been used since examining the data from two opposite approaches results in a more complete analysis. More specifically, the first one is especially useful for removing biases from the analysis and for ensuring the accuracy and replicability of the data. The latter one is important because it incorporates participants' opinions and attitudes, which have a significant impact on their actions and reveal a new dimension that needed to be examined.

A questionnaire has been created in order to analyse the participant's perception of their own trustworthiness through subjective measurements. The questionnaire was designed such that it would focus on the three dimensions of trustworthiness, namely ability, benevolence, and integrity. For ability, the participant assessed their own skills needed or desired in order to successfully perform the task, their own knowledge regarding the task, their own qualification to perform the task, their communication skills, and the teammates' faith in the participant's ability. Regarding benevolence, it has been aimed to reveal the participants' intentions to do good to the artificial agent, thus the subjects were asked whether or not they had the agent's best interests in mind, whether or not they have been motivated and have been willing and eager to help the agent in case it needs assistance and/or follow agent's commands during the game. As for the integrity measurements, the participants were questioned about whether they thought they have honoured their word, kept their promises, and told the truth to the agent.

Additional to the questionnaire, other metrics have been used in order to guarantee an objective view of the trustworthiness values as well. Specifically, for ability, methods for checking what amount of moves or clicks it took for a human in order to finish the game, and the number of victims that he/she saved, found, or picked up were implemented. For benevolence, the communication skills were central for the methods since the human could only show that he/she indeed wants to help the agent if he/she communicated whenever they found a victim or whenever they helped the agent identify the gender or pick up someone who was severely injured. However, for integrity, the focus was on whether the human carried out the activities that he/she said he/she would do (i.e. communicated they would search a room and followed through, helped identify the gender and the gender was indeed correct, etc.).

3.4 Procedure

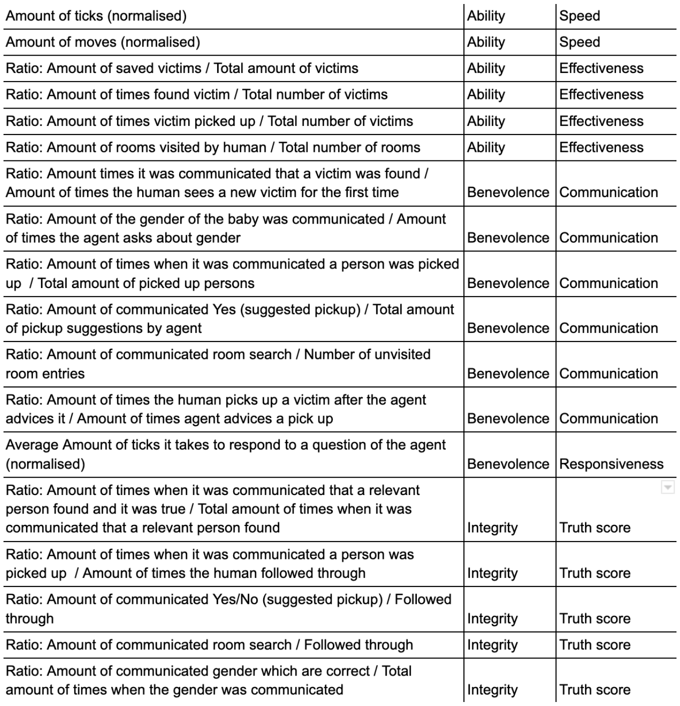
The experiment has been carried out as follows: the participant started by reading the tutorial provided by the researcher and completing the Consent Form is he/she agrees with being part of the experiment and accepts the possible risks. The tutorial page presented the overall necessary information about the game such as the goal, where the victims could be found, how the communication between the participant and the agent could be carried out, how these two parties could interact, or what are the limitations of RescueBot. Any questions regarding the procedure, game, or questionnaire can be asked meanwhile. After the possible clarifications that the researcher is directly responsible to offer, the Search & Rescue game will start. The game will last a maximum of ten minutes. During these ten minutes, the participant is supposed to collaborate with the AI agent such that the game will be completed as fast as possible. One important aspect to be noted is that the human can complete the game on his/her own, however, this is not the goal of the experiment. The participant will interact with either the agent that will give directions or with the one who gives suggestions and tips. The choice regarding the nature of the agent that interacts with the participant has been made randomly. The next step of the experiment after the game is completed is the questionnaire. In this step, the participant will answer questions regarding the participant's self-evaluation of trustworthiness. The aim of this step is to collect the attitudes and subjective points of view of the participants regarding their performance. The data is automatically collected using PKL files for the game, and the answers to the questionnaire are stored as a JSON file.

4 Results

In this section, the results of the experiment will be presented. The aim of the experiment was to establish if the human's trustworthiness will increase when the human is directed by an artificial agent.

4.1 Objective Measurements

Various methods have been implemented in order to accurately and objectively measures the ability, benevolence and integrity corresponding to the participant's performance during the game. The complete list of methods with description, and what they measure can be found in Figure 5.



Amount of ticks (normalised)	Ability	Speed
Amount of moves (normalised)	Ability	Speed
Ratio: Amount of saved victims / Total amount of victims	Ability	Effectiveness
Ratio: Amount of times found victim / Total number of victims	Ability	Effectiveness
Ratio: Amount of times victim picked up / Total number of victims	Ability	Effectiveness
Ratio: Amount of rooms visited by human / Total number of rooms	Ability	Effectiveness
Ratio: Amount times it was communicated that a victim was found / Amount of times the human sees a new victim for the first time	Benevolence	Communication
Ratio: Amount of the gender of the baby was communicated / Amount of times the agent asks about gender	Benevolence	Communication
Ratio: Amount of times when it was communicated a person was picked up / Total amount of picked up persons	Benevolence	Communication
Ratio: Amount of communicated Yes (suggested pickup) / Total amount of pickup suggestions by agent	Benevolence	Communication
Ratio: Amount of communicated room search / Number of unvisited room entries	Benevolence	Communication
Ratio: Amount of times the human picks up a victim after the agent advices it / Amount of times agent advices a pick up	Benevolence	Communication
Average Amount of ticks it takes to respond to a question of the agent (normalised)	Benevolence	Responsiveness
Ratio: Amount of times when it was communicated that a relevant person found and it was true / Total amount of times when it was communicated that a relevant person found	Integrity	Truth score
Ratio: Amount of times when it was communicated a person was picked up / Amount of times the human followed through	Integrity	Truth score
Ratio: Amount of communicated Yes/No (suggested pickup) / Followed through	Integrity	Truth score
Ratio: Amount of communicated room search / Followed through	Integrity	Truth score
Ratio: Amount of communicated gender which are correct / Total amount of times when the gender was communicated	Integrity	Truth score

Figure 5: Objective Measurements & Interpretation

The trustworthiness has been computed as the average of the ability, benevolence and integrity because the time constraints did not allow for a deeper research into the most optimal weights of these three attributes.

The bar graph from Figure 6 displays the averages of the participant's scores for ability, benevolence, integrity and trustworthiness. In order to generalise, statistical tests have been used for checking if the difference was random or if the conclusions were supported by data [18].

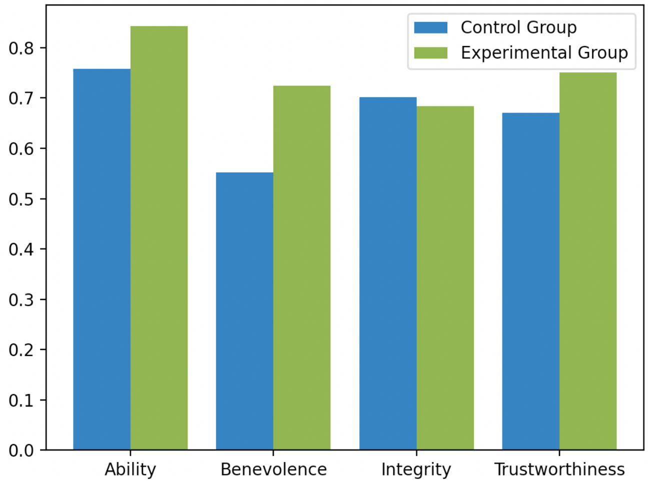


Figure 6: ABI Objective Measures Comparison

Table 1: Comparison of Means and Standard Deviations for Control Group

	Mean	Standard Deviation
Ability	0.758	0.158
Benevolence	0.566	0.200
Integrity	0.702	0.253
Trustworthiness	0.675	0.161

In order to check for the normality of the data that has been gathered, the Shapiro-Wilk test (1965) has been used, and the original constrains of the sample size smaller than 50 did not represent an issues as the sample size of experiment was 40 [22]. If data was not normally distributed, then the Mann-Whitney test have been performed to check whether there are significant changes in the ability, benevolence, integrity and trustworthiness when the artificial agent is directing the human. Otherwise, T-test were performed.

The null hypothesis of Shapiro-Wilk test states that data is normally distributed. If p-value is smaller or equal to 0.05, then the null hypothesis is rejected. This means that data is not normally distributed. If p-value is greater than 0.05, then the null hypothesis is not rejected. This means that data may be normally distributed, but it does not guarantee the normality.

Table 2: Comparison of Means and Standard Deviations for Experimental Group

	Mean	Standard Deviation
Ability	0.84	0.082
Benevolence	0.581	0.133
Integrity	0.684	0.244
Trustworthiness	0.701	0.101

Table 3: P-values for the Shapiro-Wilk test for Control and Experimental Groups

	Control Group	Experimental Group
Ability	0.004	0.29
Benevolence	0.19	0.13
Integrity	0.003	0.21
Trustworthiness	0.28	0.64

From the above table, it can be concluded that data may be normally distributed in the experimental group for all four dimensions (i.e. Ability, Benevolence, Integrity and Trustworthiness). In the control group the data for Benevolence and Trustworthiness may be normally distributed, while the values for Ability and Integrity were certainly not normally distributed. Therefore, for Ability and Integrity, Mann-Whitney tests have been carried out and T-tests for Benevolence and Trustworthiness.

The results for these tests were as follows:

- The 20 participants that have interacted with the agent that was giving directions and commands (M = 0.840, SD = 0.082) compared to the 20 participants from the control group (M = 0.758 , SD = 0.158) did not show significantly changes in the values for the *Ability* measurement, t(38) = -2.015, p = .051.
- For *Benevolence*, there was a significant difference in the scores for the experimental group (M = 0.581, SD = 0.133) and control group (M = 0.566, SD = 0.2); t(38) = -2.673, p = .011.
- There was no significant variation for *Integrity*, t(38) = 0.23, p = .819, between the participants in the control group (M = 0.702 , SD = 0.253) and the ones in the experimental group (M = 0.684 , SD = 0.244).
- Overall, the participants from the experimental group (M = 0.701, SD = 0.101) did not show significant differences regarding their *Trustworthiness*, t(38) = -1.766, p = .085, comparing to the control group (M = 0.675, SD = 0.161).

4.2 Subjective Measurements

In order to measure the internal consistency within the responses of each participant, the Cronbach's alpha has been used since this is the most frequent test for reliability scores [4]. Below it can be found the table with the value resulted from the Cronbach's alpha test for the questionnaire.

Generally, a value of 0.7 or above is regarded as acceptable, meaning that there is internal consistency within the responses of the candidates. As it can be seen in the table, the questions regarding the ability in the experimental group have a rather low value (0.5), thus this indicates the participants did

Table 4: Cronbach's alpha Result for Control and Experimental Groups

Cronbach's alpha	Control Group	Experimental Group
Ability	0.742	0.504
Benevolence	0.892	0.787
Integrity	0.904	0.869

not have consistent answers for questions meant to measure the same aspects. The rest of the values were above 0.7, so the participants' responses were consistent.

The questionnaire contains 15 Likert scale questions regarding self-evaluation of the ability, benevolence and integrity of the participant throughout the game. More specifically, each dimension had five corresponding questions. The results derived from the questionnaire can be found below.

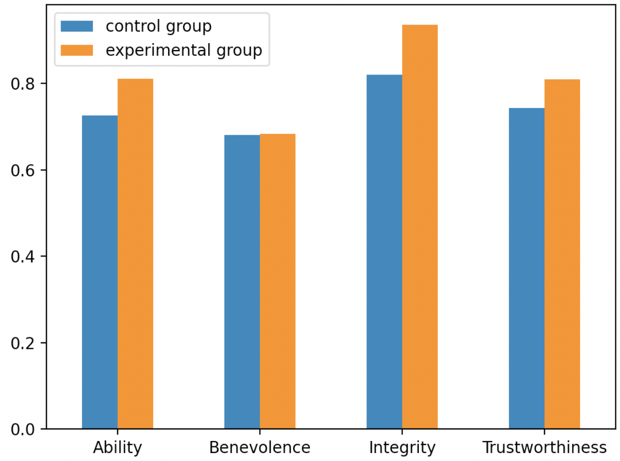


Figure 7: ABI Questionnaire Score Comparison

In the bar graph from Figure 7, the average values for ability, benevolence, integrity, and trustworthiness can be observed. Since the Cronbach's alpha for the ability questions in the experimental group was low, the corresponding result derived form the questionnaire can be regarded as not reliable. Statistical tests have been used in order to determine if the difference was arbitrary and the conclusions could be supported by data.

In order to check for the normality of data, the Shapiro-Wilk test has been used for the subjective measures, as well as for the objective ones.

Table 5: Values of the Shapiro-Wilk test for Control and Experimental Groups according to the Questionnaire

	Control Group	Experimental Group
Ability	0.478	0.299
Benevolence	0.066	0.111
Integrity	0.008	2.4405119347647997e-06
Trustworthiness	0.401	0.052

The results indicate that the data may be normally distributed both in the control and experimental group for the questions regarding ability, benevolence, and trustworthiness, thus T-tests have been performed. The data is not normally distributed for the questions quantifying integrity in both groups, therefore Mann-Whitney tests have been carried out.

Table 6: Comparison of Means and Standard Deviations for Control Group

	Mean	Standard Deviation
Ability	0.726	0.142
Benevolence	0.681	0.262
Integrity	0.820	0.184
Trustworthiness	0.742	0.166

Table 7: Comparison of Means and Standard Deviations for Experimental Group

	Mean	Standard Deviation
Ability	0.811	0.101
Benevolence	0.683	0.219
Integrity	0.935	0.123
Trustworthiness	0.81	0.113

The results of the tests have been:

- The participants that have interacted with the agent that was giving directions (M = 0.811 , SD = 0.101) evaluated their *Ability* skills significantly higher, t(38) = -2.142, p = .039, compared to the 20 participants from the control group (M = 0.726, SD = 0.142).
- For *Benevolence*, there was no significant difference between the self-evaluation scores in the experimental group (M = 0.683, SD = 0.219) and control group (M = 0.681, SD = 0.262); t(38) = -0.026, p = .98.
- There was a significant increase of the *Integrity*, t(38) = -2.257, p = .03, for the experimental group (M = 0.935 , SD = 0.123) in contrast to the control group (M = 0.820 , SD = 0.184).
- Generally, the participants from the experimental group (M = 0.81, SD = 0.113) did not show significant differences regarding their self-rating of *Trustworthiness*, t(38) = -1.459, p = .153, comparing to the control group (M = 0.742, SD = 0.166).

4.3 Confounding Variables

The participants came with a strong background in Computer Science and gaming. Therefore, their ability in the context of successfully completing the game is quite high regardless of their benevolence for helping the robot to save victims or integrity. Moreover, several participants have interacted with the MATRX software before, thus they were more familiar with the process and design of the game used for the experiment. Therefore, this influences the results of measurements of trustworthiness.

Another confounding variable could be the English proficiency of the participant. If their knowledge is vast or if English is their mother language, then their response time to the tasks that the AI agent suggests or commands them to do would be lower since the human would spend insignificant or no time at all on understanding what is required of them and how they should answer.

5 Responsible Research

This section comprises a reflection on the ethical implications and considerations of the experiment and a discussion on the reproducibility of the research methods. These aspects are especially relevant since the world has seen horrendous experiments being done, for instance during WW2, that should never be allowed to be carried out ever again, and lately, more and more scientists are concerned that a significant part of the new studies cannot be replicated and thus, their results cannot be verified accordingly [1].

To begin with, the research has received the approval of the Human Research Ethics Committee (HREC) at TU Delft since it is considered to be Minimal Risk. For instance, the experiment poses no possible physical risks (i.e. injury, deterioration of participant's health, or any kind of physical discomfort) or psychological risks (e.g. anxiety, mental stress, loss of self-esteem, or distress as a result of deception). Additionally, the experiment does not imply any economic or social risks such as financial loss, deterioration of reputation, and social relationships. No specific personally identifiable information (PII) nor any associated personally identifiable research data (PIRD) has been collected. However, as defined by GDPR (the General Data Protection Regulation) personal data such as age group, gender, or childhood place has been collected during the experiment which implies a certain amount of risk regarding the safety and security of the participants. As a mitigation method for the risk, the personal data will not be linked in any way to the participants' responses and performance in the Search and Rescue game nor stored in the institutional open repository of TU Delft. Another potential risk associated with this experiment is that, even though the participants are not part of vulnerable groups or are in a subordinate position to the experimenter, they are usually part of the personal network of the researcher, thus they may feel pressured to behave in certain ways that would please the researcher. In order to minimise this risk, the participants are required to sign a consent form where it is stated that they are allowed to choose to stop their participation in the experiment at any point without being required to give any explanation.

Furthermore, the reproducibility of this research is guaranteed through various steps including automated data analysis, publishing all the data that has been gathered such that anyone can inspect it, the experimental setup being explained in detail, and having two different approaches for quantifying trustworthiness in order to minimise the bias (i.e. subjective measures through questionnaire and objective measures through programming methods for computing the performance of a participant). The automated data analysis is performed using the open-source Python library *SciPy*[3], for example for calculating the independent T-tests or Mann-Whitney tests. Moreover, the implementation of the agent or the metrics used for measuring human trustworthiness is available to anyone upon request.

6 Discussion

This section will provide a discussion and interpretation of the results that have been presented in Section 4. The experiment has been carried out aiming to assess how AI behaviours, the directing behaviour in this case specifically, affect the human's trustworthiness.

The results of the experiments indicate that there is no significant change in human trustworthiness when the artificial agent is directing the human in a setting with a simplistic and common task such as the Search & Rescue game. Therefore, the hypothesis could not be validated.

On the other hand, the Two-Tailed Test has showed that, according to the objective measures, the benevolence increases when the human is directed by the AI in joint interactions. For this experiment, assessing human benevolence was done by measuring:

- How well the human communicated his/her actions with the agent (e.g. how many times the humans said they found a new victim compared to the amount of times when the humans found a new victim)
- How fast the human performed the task that the agent suggested/directed
- How many times the human performed the task that the agent suggested/directed compared to the total amount of times the agent gave suggestions/directions

This result was expected since obedience to authority is a well-researched phenomenon in social psychology that essentially boils down to the fact that people have a strong tendency to obey authority [17]. Because the directing agent shows more authority than the agent which offers suggestions, the humans were expected to report their actions and to follow through with the directions more in the experimental group.

Furthermore, the Two-Tailed Tests have indicated that, according to the subjective measures (i.e. questionnaire), people evaluated their ability higher when they were directed by the artificial agent. Since it is well acknowledged that the fields that are using AI had improved their quality and become more efficient, it is natural to think that an improvement will happen at individual level as well [19]. Therefore, the increased confidence in their own ability that the participants from the experimental group have shown was expected. However, the Cronbach's alpha, that measures reliability through internal consistency, indicated that the answers for the questions aimed to measure the ability were not consistent.

Lastly, the Mann–Whitney U test has revealed that the participants interacting with the directing agent have self-assessed their integrity higher than the ones interacting with the agent that was giving suggestions. This increase in the perception that people had about their own integrity could be explained again by the concept of obedience to authority. This is because the humans complying with RescueBot's directions may actually result in them having shared values and goals such as to find and save the victims as fast as possible. In order to reach this goal, the human had to be honest and fair.

[3] https://scipy.org/

7 Limitations

The results and conclusions of this experiment need to be considered taking into consideration various limitations, and thus the aim of this section is to highlight and interpret them.

The first limitation that can be observed is the small sample size for the experiment. Due to the time constraints of this study, only forty participants have taken part in the experiment, therefore the assumption that such a small number of subjects can be representative of a population of billions of people is an example of cognitive bias. Additionally, the confounding variables, as explained in subsection 4.3, could have impacted the results of the study, thus further experiments with participants with more diverse backgrounds are vital in order to guarantee the generality and validity of the conclusion. Moreover, in order to offer the possibility for the participants to take part in the experiment online, the port forwarding technique has been used, which had added latency of a few seconds. This could have affected the performance of participants in the experiment, their benevolence, and ability in special since the human may get irritated because of the lag or prioritise completing the game and saving the victims over communicating with the robot, which affects their trustworthiness. Finally, the diversity of the participants in terms of cultural and social background is scarce as well as most of them were recruited through personal networking which resulted in a pool of participants almost entirely from Europe. Thus, the results may not be representative of the overall population and future study of this aspect is strongly recommended in order to incorporate the ethical differences between different continents and cultures.

8 Future Work

Several studies advise that Cronbach's alpha results should not be used without conditions, recommending the reliability scores based on the structural equation modeling [9; 24]. For example, a low alpha score may indicate that there is not a sufficient number of questions in the questionnaire, instead of simply meaning that there is low consistency in the responses. Thus, in this case, more questions could be added in order to increase and guarantee the reliability of the answers.

For future studies, more diverse sample size is strongly recommended. People from all backgrounds, and ages, with various expertise in computer science and gaming, and with different English proficiency should be recruited in order to make the results of the study representative of the population.

The participants of the experiments have given their recommendations for improving the experience of the research. In summary, they advised having a voiceOver feature when collaborating with the robot because it is easier and faster to communicate their actions and to remember and pay attention to what the RescueBot is telling. More than a quarter of the participants have said that they would have shared more information regarding their next or current more if they had this feature. More than a third of the participants mentioned that they would have collaborated more with the agent if the agent was faster. Since the participant was faster than the RescueBot, they had chosen to go and pick the victims up themselves instead of communicating to the agent and waiting for it to pick them up. Thus, the behaviour of the agent needs to be optimised in order to allow faster movements.

9 Conclusion

The aim of this research was to assess the changes in human trustworthiness as a result of an artificial agent directing the human in joint activities. In order to quantify the trustworthiness, an experiment in the form of a search & rescue game developed in MATRX software has been carried out. Both objective measures and subjective measurements (i.e. a questionnaire), which can be found in Section 4, have been used.

Statistical tests indicated that neither for the subjective measurements nor for the objective ones, the hypothesis stating that trustworthiness would increase when the artificial agent was directing the human, could not be validated since there was no significant difference between the values obtained in the control and experimental group.

Nonetheless, the results of the experiment showed a significant increase in the objective measurements of the participant's benevolence when the participant interacted with a directing agent. Another positive correlation has been found between the participants' self-evaluation of their ability and the interaction with the directing AI. However, Cronbach's alpha for the answers to the questions regarding ability indicated a low internal consistency. Furthermore, the collaboration between the human and the directing agent has been found to be positively correlated with the participant's self-perceived integrity as well.

The findings of this study need to be considered in light of several limitations such as the small sample size and the lack of diversity of the participants, especially in terms of cultural background and computer science knowledge. Future work might use different tests for the internal consistency of the answers to the questionnaire, and explore more with distinct definitions of directability.

References

[1] Jesse M Alston and Jessica A Rick. A beginner's guide to conducting reproducible research. *Bulletin of the Ecological Society of America*, 102(2):1–14, 2021.

[2] JM Bradshaw, V Dignum, CM Jonker, and M Sierhuis. Human-agent-robot teamwork. *IEEE Intelligent Systems*, 27(2):8–13, 2012. Betreft in feite combinatie van editorial en artikel van de gastredactie ter inleiding op speciale uitgave journal n.a.v. een in 2010 gehouden conferentie. Vandaar non-refereed.

[3] Elizabeth J Chang, Farookh Khadeer Hussain, and Tharam S Dillon. Fuzzy nature of trust and dynamic trust modeling in service oriented environments. In *Proceedings of the 2005 workshop on Secure web services*, pages 75–83, 2005.

[4] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

[5] Salas Eduardo. Is there a" big five" in teamwork? *Small Group Research*, 36(5):555–599, 2005.

[6] Ernst Fehr. On the Economics and Biology of Trust. *Journal of the European Economic Association*, 7(2-3):235–266, 05 2009.

[7] Kylie Foy. Artificial intelligence is smart, but does it play well with others?, 10 2021.

[8] Jackie Tucker Gangnes. What is Team Intelligence and How to Manage It, 03 2022.

[9] Samuel B Green and Yanyun Yang. Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74(1):121–135, 2009.

[10] R Hallows, L Glazier, MS Katz, M Aznar, and M Williams. Safe and ethical artificial intelligence in radiotherapy–lessons learned from the aviation industry. *Clinical Oncology*, 34(2):99–101, 2022.

[11] Manh Hung and Dinh Que. A Trust-based Mechanism for Avoiding Liars in Referring of Reputation in Multiagent System. *International Journal of Advanced Research in Artificial Intelligence*, 4(2), 2015.

[12] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69, 2014.

[13] Matthew Johnson and Alonso Vera. No AI Is an Island: The Case for Teaming Intelligence. *AI Magazine*, 40(1):16–28, 2019.

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

[15] Iain J Marshall and Byron C Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8(1):1–10, 2019.

[16] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[17] Stanley Milgram and Christian Gudehus. Obedience to authority, 1978.

[18] Todd Neideen and Karen Brasel. Understanding statistical tests. *Journal of surgical education*, 64(2):93–96, 2007.

[19] Avneet Pannu. Artificial intelligence and its application in different areas. *Artificial Intelligence*, 4(10):79–84, 2015.

[20] Isaac Pinyol, Roberto Centeno, Ramon Hermoso, Viviane Torres da Silva, and Jordi Sabater-Mir. Norms evaluation through reputation mechanisms for bdi agents. In *Artificial Intelligence Research and Development*, pages 9–18. IOS Press, 2010.

[21] Sarvapali D Ramchurn, Sebastian Stein, and Nicholas R Jennings. Trustworthy human-ai partnerships. *Iscience*, 24(8):102891, 2021.

[22] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[23] Luciano Cavalcante Siebert. Responsible and ethical ai, March 2022.

[24] Klaas Sijtsma. Reliability beyond theory and into practice. *Psychometrika*, 74(1):169–173, 2009.

[25] Paola Tubaro and Antonio A Casilli. Micro-work, artificial intelligence and the automotive industry. *Journal of Industrial and Business Economics*, 46(3):333–345, 2019.