Assisting agent's effect on the trustworthiness of their human teammate

Alto Delia
Supervisor(s): Dr. Myrthe Tielman, C. Ferreira Gomes Centeio Jorge
EEMCS, Delft University of Technology,
Delft, The Netherlands

22-6-2022

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

Collaborative AI (CAI) is a fast growing field of study. Cooperation between teams composed of humans and artificial intelligence needs to be principled and founded on reciprocal trust. Modelling the trustworthiness of humans is a difficult task because of the ambiguous nature of its definition as well as the effect of team work dynamic. This research defines and measures human trustworthiness within the context of human-AI collaboration and tests how the artificial intelligence agent's action of offering help plays a role in it. The experiment is conducted through the MATRX framework, in which a human agent and an AI agent collaborate in order to search and rescue victims within the environment. The ABI trust model is used to determine the sub-components that define trustworthiness, which is ability, benevolence and integrity. Trustworthiness is measured in 2 ways, through objective measures which represent the AI agent's measure of the human trustworthiness and subjective measures that represent the human's measure of their own trustworthiness. The results show that help offered by the AI agent, improves the ability and benevolence of the human agent in objective metrics but not integrity. Subjective results show no statistically significant change. The research concludes that the trustworthiness perceived by the AI agent is indeed improved, but does not provide evidence of the same from the perception of the human agent.

1 Introduction

People and technology are becoming more interdependent by the day. Because of this, the interaction between the two in the context of teamwork, should be built around strong foundations. Contemporary AI systems (agents) have developed into systems with cognitive capabilities and autonomous behavior, which allows them to complete tasks without the need for human supervision. In a collaborative workspace, artificial agents may play the role of an assistance provider to the designated user, but they can also take initiative (Razmerita, Brun, & Nabeth, 2021). This paper tries to reason about the impact that these assisting agents have on the performance of humans in certain tasks. Judging this performance will enable the AI agent to model the trustworthiness of the human.

Trustworthiness of an agent is described as the property of the human agent that expresses the tendency to behave in a way in which they have promised to behave and other agents expect them to behave. (Hardin, 2002). There is various literature regarding the assessment of AI trustworthiness, however, that is not the case when it comes to assessing the human trustworthiness in a human-AI team. The applications of the latter are arguably just as important, as good trustworthiness modelling empowers artificial agents to take informed decisions when performing tasks with human counterparts.

The aim of this research project is to explore ways in which human trustworthiness is affected during a human-AI collaboration, and it focuses specifically on how bring offered help or assistance from the artificial agent plays a role in this. The research question is phrased as follows: **How does an artificial agent offering help affect human trustworthiness?** In this paper trustworthiness is defined through the ABI trust model (Lewis, Sycara, & Walker, 2018). The hypothesis that this paper will explore is that being offered help will allow for a better collaboration, which will lead to a better task performance and ultimately higher human trustworthiness.

Section 2 explores more in depth the problem at hand. In section 3 the methods to solve the problem are explained including the design of the experiment, the population of the participants, and the metrics being measured. The results are discussed thoroughly in section 4. In section 5 the ethical considerations of the research are laid out. The paper finishes with section 6 that consists of discussion of the results as well as conclusions which are discussed in section 8.

2 Literature and Background

In earlier days, research into artificially intelligent systems was aimed at contrasting strengths between them and humans, in order to identify tasks at which AI can replace human agents. This lead to tasks being divided into two categories: tasks that humans are better at and tasks that machines are better at, known as HABA-MABA (Bradshaw, Dignum, Jonker, & Sierhuis, 2012). With time however, researchers realized that there are tasks that cannot simply be solved through the autonomous work of these 2 counterparts, thus, possibilities of collaboration and teamwork between them were studied. In their research, Salas et al. defined teamwork as “a set of interrelated thoughts, actions, and feelings of each team member that are needed to function as a team and that combine to facilitate coordinated, adaptive performance and task objectives resulting in value-added outcomes” (Salas, Sims, & Burke, 2005, p. 562). Having a good understanding of what this pool of various thoughts, actions and feelings are helps teamwork become effective, by enabling the agents involved to make more informed decisions.

Concepts such as thoughts and feelings are normally ascribed to human teammates as opposed to artificial intelligence, who are normally judged upon other cognitive abilities instead, such as memory and logical reasoning. However, research has shown that humans prefer to interact with artificial intelligence systems that display anthropomorphic traits such as emotions ane expressions of good intent (Kiesler & Goetz, 2002). These findings were at the core of the development of social robotics, which is a field that deals with humans and robots interacting in ways humans typically interact with each other (Lewis et al., 2018). In the context of this paper, the AI agent will try to mimic human interaction when offering assistance in an attempt to invite reliance on the help it offers.

Furthermore, research indicates that being trusted is also a factor that plays a role in the trustee’s trustworthiness in general. (Falcone & Castelfranchi, 2004) conclude through a series of task delegation experiments that the action of trusting, indeed interferes with trustworthiness itself. This means that there is some indication that actions of the trustor within the teamwork affect the trustworthiness of the team partner. This is insightful to our research as it aims to further explore how the trust dynamics and specific actions of an artificial agent with a high trust propensity, such as help offering, influence the trustworthiness of the human agent.

As previously mentioned, in the context of collaboration between two or more agents, trustworthiness is usually modelled through trust. Trust is a directional transaction between two or more parties involved (Jacovi, Marasović, Miller, & Goldberg, 2021).(Mayer, Davis, & Schoorman, 1995, p.) described trust as follows: “If A believes that B will act in A’s best interest, and accepts vulnerability to B’s actions, then A trusts B”. This means that for the trustor, the trust it has in the trustee is a measure of the perceived trustworthiness of the trustee combined with the risks involved in relying on the trustee’s actions. Trust however, in the CAI setting, is a concept that has been modelled on the basis of different attributes. There is no universally accepted definition to it. For the scope of this project some models such as SWIFT (Haring et al., 2021) and Barber’s Model (Heineman, 1984) were considered, which looked at concepts such as purpose, competency and responsibility in order to quantify trust. Despite the wide range of components which trust is build upon, there is a convergence on three main ones, namely Ability, Integrity, and Benevolence (Lewis et al., 2018).

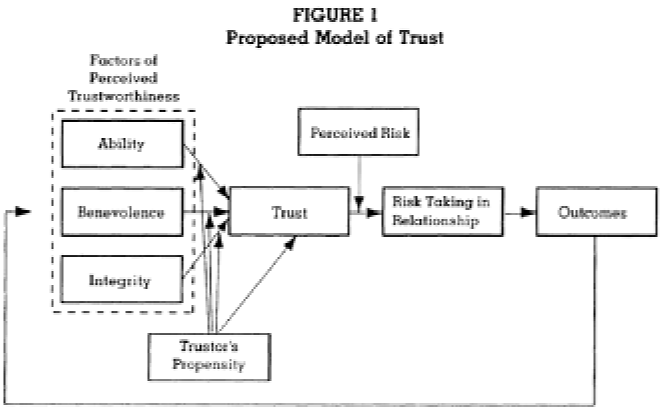

Figure 1: ABI Model Framework.
(Mayer et al., 1995, p. 715)

The ABI-model, as described by (Mayer et al., 1995), and graphically depicted in Figure 1, is a model of trust which is built upon these 3 concepts as its main pillars. They are formally defined as follows:

1. “Ability - is that group of skills, competencies, and characteristics that enable a party to have influence

within some specific domain" (Mayer et al., 1995, p. 717).

2. "Benevolence - is the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive" (Mayer et al., 1995, p. 718).

3. "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" (Mayer et al., 1995, p. 719).

When it comes to human-AI teams, there is a higher unpredictability and a wider range of factors that affect the ability, benevolence and integrity of humans as opposed to that of AI. Objectively quantifying these attributes will be one of the challenges that this research faces. This is mainly because human's behavior, emotions and intentions are more prone to be affected differently by changes in the social setting. The research serves the purpose of finding patterns on how these qualities change because of the AI's actions. That, coupled with lack of extensive research when it comes to modelling trustworthiness of a human agent, make this research of valuable contribution to the field of Collaborative AI. The methods on how the research tackles these challenges in order to contribute to the understanding of collaboration between humans and machines are described in the next section.

3 Methodology

In order to answer the question that this research poses, it is important to define some tools, measures and models. Here we give a definition of the experimental set-up, the technology used, the objective and subjective metrics that we will use to quantify trust and how results will be derived from these measurements.

3.1 Technology

For the purpose of this research a simulation of an AI-Team collaboration is needed in order to test for the trustworthiness. This will allow modelling the qualities of the human through the agent's perspective. To do this, MATRX, framework is used (van der Waa & Haije, 2022). The reason for this is that MATRX is popularly used within the institution of TU Delft. It has built-in functionality that is designed to facilitate simulations of this nature, which makes it suitable for this research. The language that will be used to construct the environment as well as develop the AI will be Python. The MATRX scenario with which the experiment is conducted will be an Urban Search and Rescue (USAR) (Verhagen, 2022), where the agents have to work together in order to find and rescue injured victims within a given time limit.

3.2 Experimental Design

In order to test the effect that offering help has to the human agent's trustworthiness it is important to explain the design of the experimental process and how the considerations that were taken into account during the design contribute to answering our research question. The design choices of the collaboration challenge as well as those of the AI agents are discussed further in this section along with the justification of these choices.

3.2.1 The USAR Task

The Urban Search And Rescue environment is essentially a map of rooms distributed in different locations, inside which there are victims/healthy characters. The human-AI team's goal is to search within the rooms for injured characters. The red colored ones are suffering from high severity injuries and therefore have the highest priority. The yellow ones have mild injuries, therefore they should also be rescued but to a lesser priority, whereas green ones are healthy thus the team should disregard them. The injured victims should be brought by the team members to the drop-off zone, which also describes the order in which the victims should arrive. An image of the environment is portrayed in Figure 2.

For a better team performance communication between the members is essential. For this reason the MATRX environment also offers a chat box and a list of messages that the human agent can send to its counterpart. This communication can then be used to optimize the team's strategy and reduce redundant

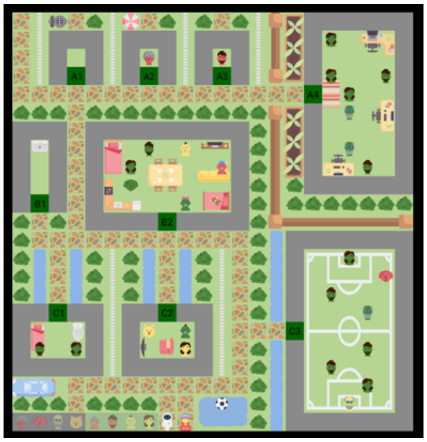

Figure 2: Urban Search And Rescue MATRX Environment.

work between the human and AI. In order to ensure collaboration between the two, a high interdependence setting was preferred. Research has shown that interdependence-focused frameworks enables research to portray human-machine teamwork in a more interpretative, understandable and generalizable way. (Johnson & Bradshaw, 2021). Thus, in the context of our research, interdependence was implemented through restrictions that insured the task could not be solved individually by the AI agent. The artificial agent was dependent on the skill set of one another in tasks such as identifying the gender of the characters, restrictions in what victims the agent can carry. The agent also offered tips, suggestions and help.

3.2.2 Design of Baseline and Helper Agent

As mentioned previously, the experiment was designed to fit a control group and experimental group. The two groups performed with two different agents, a baseline agent and a helper agent. The difference between the agents being that the helper agent was a more developed agent in terms of functionality, as it was designed to offer extra help in aspects the baseline agent was not able to. This was accomplished by implementing chat commands through which its human teammate could ask questions regarding the rules of the game, the current status of the game as well as ask the agent to pick up injuries.

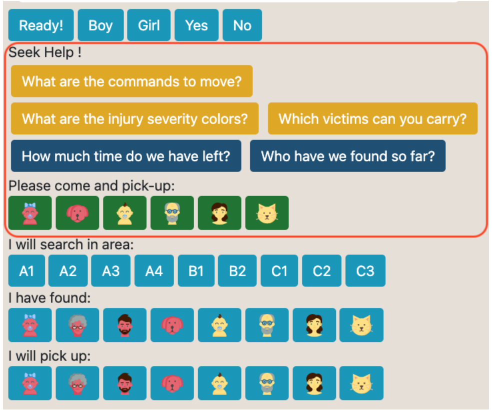
Ready!
Boy
Girl
Yes
No
Seek Help !
What are the commands to move?
What are the injury severity colors?
Which victims can you carry?
How much time do we have left?
Who have we found so far?
Please come and pick-up:
I will search in area:
A1
A2
A3
A4
B1
B2
C1
C2
C3
I have found:
I will pick up:

Figure 3: Communication buttons of Baseline and Helper Agents

In Figure 3 we can see the communication options that the human agent is offered. The sky blue commands are part of the baseline implementation whereas the gold, dark blue and green commands define the help that the helper agent offers to the human. The helper agent is also programmed to send a reminder message every minute in order to offer their teammate the possibility to seek for help in order to perform better.

Apart from this difference, the two agents behave the same when it comes to solving the task. They are programmed to go through a few operational phases such as identifying next goal victim, planning the room search operation, traversing the maps, identifying goal victims, picking up and transporting victims to the drop off locations. The map in which they operate is also kept the same within every participant in order to avoid creating extra variables to take into account. The difficulty of the challenge remains the same. The only difference between the two agents is the independent variable which is introduced in the form of extra help offered to the human agent as explained. The helper agent also introduces a new human action of seeking help and also tracks every time that the agent offers this help. Later this is used to test for a correlation between the frequency of offered help and the human trustworthiness. The extended hypothesis in this case would be that the more help is offered the higher the trustworthiness, as help enables a stronger relationship between the two agents and potentially increases the individual performance of the human agent.

3.3 Participants

Since the goal of the research is to gain insight on what affects human qualities and behavior, the experiment requires involvement of human participants. In order to ensure meaningful results it was decided to have a control group and experimental group of 20 participants each. The control group played the game with a baseline agent as a teammate, whereas the experimental group had an assisting agent to team up with for the task.

Category	Values	Control Group	Experimental Group
Age	18-24	14	16
	25-34	2	4
	35-44	2	0
	45-54	1	0
	55-64	1	0
	65+	0	0
Gender	Male	11	11
	Female	8	9
	Other	1	0
Place of Birth	Africa	1	1
	Asia	1	2
	Australia	0	0
	Europe	17	15
	North/Central America	0	1
	South America	1	1
Language Proficiency	Low	0	0
	Average	1	0
	High	19	20
Gaming Experience	Low	3	5
	Average	9	10
	High	8	5

Table 1: Participant Demographics

Table 1 displays the demographics and confounding variables of the control group population and the experimental group population in terms of age, gender and nationality, as well as proficiency in the language of the experiment and gaming experience. As seen in the table the population is made of a majority of young people between the ages of 18-34. In the experimental group, the population is strictly between this range. This is a limitation of the research which will be further discussed. When it comes to gender distribution the two groups are almost identical and the difference between male and females is almost 50-50. The nationality of the population is majority European, given that the research is taking place in Europe. In both populations there are instances of people coming from other continents such as Asia, Africa and South America. When it comes to language proficiency, almost all participants have a high language proficiency. Furthermore, when it comes to gaming experience the majority of participants answered with 'average' or 'high'.

3.4 Objective and Subjective Metrics

In the context of this project **objective** metrics represent metrics that are measured by the AI-agent in an attempt to model the human's trustworthiness. This is accomplished by keeping logs of all the human's actions and communication that occurs within the context of the game. Table 2 explains how the ABI scores were drawn using information from the logs.

Ability	Benevolence	Integrity
Amount of time spent (measured through in-game ticks)	Communication score on tasks agent needs assistance	Ratio of truthfully communicated baby gender
Amount of moves by human agent	Communication score on suggestions agent makes	Ratio of truthfully communicated actions (room search/pick-up, etc.)
Ratio of total saved victims over total (game completion)	Ratio of communicated actions over total actions	Ratio of truthfully communicated suggestion response
Ratio of discovered rooms over total	Average time (ticks) it takes the human to respond to teammate	
Ratio of victims dropped off by human over total		

Table 2: Objective metrics used to calculate Ability, Benevolence and Integrity scores.

As seen from the table the metrics are categorized to represent each component of the ABI trust model. Ability is determined through measures that are focused at team performance such as time of completion. The quickest participants will therefore score higher in this metric. Completion of the game is also important to measure especially for participants that do not finish the game within the 10 minute mark. This tests how many victims were rescued over the total amount that are supposed to be rescued. Furthermore, the human contribution to the rescued victims is also measured in order to identify how much of the completion was due to the human as opposed to being reliant on the AI to solve the game.

When it comes to benevolence, communication was the key attribute that was taken into account. Given that within the game the chatbox was the main channel of interaction between the members of the team, it was reasonable to consider frequent communication with their teammate as benevolent behavior. This choice is also reinforced by the fact that the AI agent relies heavily on this communication to optimize its rescue strategy. Communication score, as referenced in the table is calculated through averaging the ratios of communicated actions to total amount of actions undertaken by the human.

Integrity score is measured through the ratio that the human agent truthfully communicates an action. Being an honorable teammate is key to achieve a high integrity score, and through ratio of truths to lies, the model is able to quantify the human agent's integrity. The value of trustworthiness is generated by taking the mean of the ABI scores. After all three ABI component scores are produced as described, the average of all the scores represented the objective human trustworthiness score.

A **subjective** measurement of human trustworthiness was included in the methodology in the form of a questionnaire. For this, after the completion of the Search and Rescue game, the participants had to answer 15 questions through which they evaluated their performance and collaborative skills. The questionnaire covered 5 questions for each of the ABI components. The questions try to prompt the participant to evaluate their performance in the context of the team collaboration. Similarily to objective metrics, when the individual scores for ability, benevolence and integrity are computed the average of those results will produce the subjective trustworthiness value. To ensure the consistency and internal integrity of all the components of the questionnaire the Chronbach's alpha value was measured for every group and component.

	Cronbach's α – *Ability*	Cronbach's α – *Benevolence*	Cronbach's α – *Integrity*
Control Group	0.742	0.892	0.904
Experimental Group	0.735	0.645	0.831

Table 3: Cronbach's alpha values regarding ABI questionnaire

Table 3 shows the results of this measurement with the majority of Cronbach's α values falling over the 0.7 mark. These results ensure strong internal consistency within the sub-components of the questionnaire.The goal of the subjective metrics is compare and aggregate the results of trustworthiness from the perception of both members of the team. The two metrics are complementary to one another, in the sense that they complete their individual shortcomings. Relying on the objective metrics alone makes the results limited only to the interpretation of the experimenter and subsequently the implementation of the agent. On the other hand, the subjective measures, given that they are directed at one's self, have a personal bias factor to be accounted for. Thus, a combination of the two perspectives was deemed suitable for the research.

4 Results

After the experimental phase the data was collected through action pickle files which stored all the interaction between the team as well as their interaction with the environment. The answers to the questionnaire were stored in JSON files for further processing. The processing of the data was organized in a few steps. The distribution data of the data-sets was modelled. Using this distribution data we perform the Shapiro-Wilk test (Shapiro & Wilk, 1965) in order to check whether there is evidence that the data is not normally distributed. Depending on the test results, it is concluded whether there is significant evidence that the data is not normally distributed, otherwise we cannot reject the null hypothesis of normal distribution. In case no significance of normality is found, Mann-Whitney test("Mann–Whitney Test", 2008) is performed, to check whether the difference in the distribution means is random or not, whereas in the case of significant normality we perform the Student's T-test (Student, 1908). In both tests the alternative hypothesis is that the mean of the experimental group is higher than the mean of the control group as explained by the research hypothesis. In order to obtain test this hypothesis, we performed a two-tailed T-test on the data. The following subsections give a graphical representation of the data, and the resulting test values of both subjective and objective measures. For further reference, in graphs and tables the experimental group is always identified with the color green, and the control group with the color blue.

4.1 Objective Measurement Results For ABI

When it comes to the objective measures the results of the ability, benevolence and integrity scores were quite aligned with the expectations.

	Mean	**SD**	**Var.**	**Skew.**
Ability Distribution Control	0.772	0.157	0.025	-1.448
Ability Distribution Experimental	0.861	0.102	0.01	-1.39
Shapiro-Wilk Test	p <0.001		p <0.004	
Mann-Whitney Test	U($N_{control} = 20, N_{expr.} = 20$) = 109.0, p < 0.014 (significant)			

Table 4: Objectively measured ability distribution values(control vs. experimental)

	Mean	**SD**	**Var.**	**Skew.**
Benevolence Distribution Control	0.542	0.218	0.047	-0.236
Benevolence Distribution Experimental	0.698	0.14	0.02	-0.828
Shapiro-Wilk Test	p <0.910		p <0.145	
T-test	t(38) = -2.694, p < 0.010 (significant)			

Table 5: Objectively measured benevolence distribution values (control vs. experiment)

	Mean	SD	Var.	Skew.
Integrity Distribution Control	0.702	0.259	0.067	-1.473
Integrity Distribution Experimental	0.727	0.182	0.033	-0.696
Shapiro-Wilk Test	$p < 0.003$		$p < 0.182$	
Mann-Whitney Test	$U(N_{control} = 20, N_{expr.} = 20) = 208.5$, $p < 0.829$ (insignificant)			

Table 6: Objectively measured integrity distribution values (control vs. experimental)

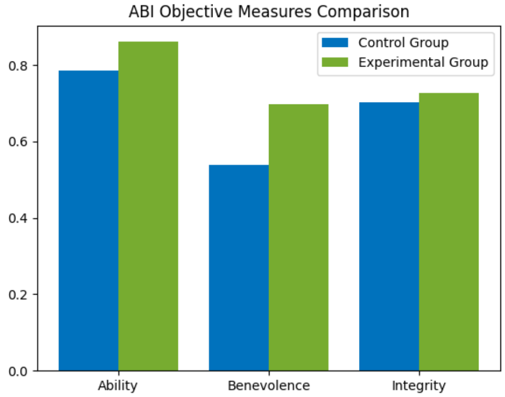

Figure 4: Objective ABI means (control vs experiment)

In tables 4, 5, 6 the distribution values are displayed for the objectively measured ABI scores are given. Figure 4 also gives a side by side bar-chart of the ability, benevolence, integrity distribution means, for both the control and experimental groups. The means of the experimental set are in all cases higher than the control set. How significant this difference is, is explored through statistics significance tests later. The control group also seems to have a higher variance indicating that the data is more scattered. The distributions are negatively skewed for all ABI scores, with the majority of them falling within the 0.5 - 1 range. Distributions of ability and integrity are much more skewed, but there is also some skewness to benevolence as well.

According to the Shapiro-Wilk test results, the ability scores for both the control and experimental group, are not normally distributed. The same can be said for the integrity distribution of the control group. For this reason the Mann-Whitney Test was applied to the data sets to test whether there is significance to the difference in means between the distribution. In the case of **integrity** for control group Mann-Whitney test indicated insignificant results $U(N_{control} = 20, N_{expr} = 20) = 208.5$, $p = 0.829$. **Ability** however yielded different results, $U(N_{control} = 20, N_{expr} = 20) = 109.0$, $p = 0.014$, meaning that the difference between the means ($M_{control} = 0.772$, $SD_{control} = 0.157$) and ($M_{expr.} = 0.861$, $SD_{expr.} = 0.102$) is significant statistically. Data collected on **benevolence** seemed to be normally distributed, as shown by the respective Shapiro-Wilk Test p-values of 0.910 for control group and 0.145 for experimental group. For this, reason the Student's T-test was performed which returned $t(38) = -2.694, p = 0.010$. This means that the difference between the two means of the sets ($M_{control} = 0.542$, $SD_{control} = 0.218$) and ($M_{expr.} = 0.698$, $SD_{expr.} = 0.14$) is significant. These results show that both mean ability scores and mean benevolence scores increase, with benevolence having a higher increase. Integrity scores do not show any significant effect when it comes to objective measurements.

4.2 Subjective Measurement Results For ABI

Subjective measured data of the ABI showed very different statistical significance in results as compared to the objective measures.

	Mean	SD	Var.	Skew.
Ability Distribution Control	0.726	0.145	0.021	0.325
Ability Distribution Experimental	0.743	0.133	0.018	-0.328
Shapiro-Wilk Test	p <0.478		p <0.833	
T-test	t(38) = -0.386, p <0.701 (insignificant)			

Table 7: Subjectively measured ability distribution values (control vs. experimental)

	Mean	SD	Var.	Skew.
Benevolence Distribution Control	0.681	0.268	0.072	-0.432
Benevolence Distribution Experimental	0.70	0.148	0.022	-0.693
Shapiro-Wilk Test	p <0.066		p <0.204	
T-test	t(38) = -0.270, p < 0.789 (insignificant)			

Table 8: Subjectively measured benevolence distribution values (control vs. experimental)

	Mean	SD	Var.	Skew.
Integrity Distribution Control	0.821	0.189	0.036	-1.046
Integrity Distribution Experimental	0.881	0.115	0.013	-0.899
Shapiro-Wilk Test	p <0.008		p <0.009	
Mann-Whitney Test	U($N_{control}$ = 20, $N_{expr.}$ = 20) = 180, p < 0.594 (insignificant)			

Table 9: Subjectively measured integrity distribution values (control vs. experimental)

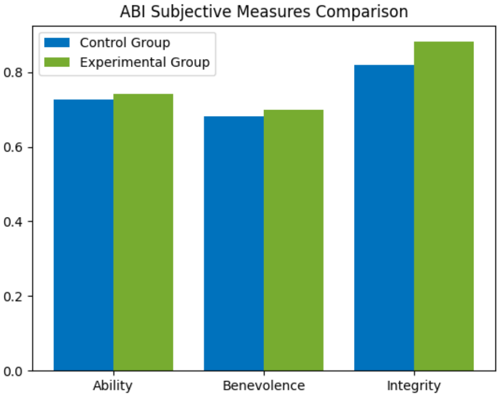

Figure 5: Subjective ABI means (control vs experiment)

The distribution measures of the subjective data of ability, benevolence and integrity are displayed in 7, 8, 9 respectively. The Shapiro-Wilk test concluded that scores of ability and benevolence were normally distributed across both groups. For integrity scores there was no significant evidence to back that. Thus, for the former two a T-test was performed, whereas for the latter a Mann-Whitney test was performed accordingly.

For all three of the ABI components, the significance tests yielded p-values that were way above the 0.05 threshold, as seen from the tables. Ability of the control group ($M_{control} = 0.726$) and experimental group ($M_{expr.} = 0.743$), showed statistically insignificant difference, $t(38) = -0.386, p = 0.701$. Same goes for **benevolence** in which the control group ($M_{control} = 0.681$) and experimental group ($M_{expr.} = 0.70$) showed no significant effect of the offered help, $t(38) = -0.270, p = 0.789$. On the other hand subjectively measured integrity distributions showed lack of normality, thus the control group ($M_{control} = 0.821$, $SD_{control} = 0.189$) and experimental group ($M_{expr.} = 0.881$, $SD_{expr.} = 0.115$) means where put through the Mann-Whitney testm which similarly to ability and benevolence significance tests yielded insignificant results, U($N_{control} = 20$, $N_{expr} = 20$)= 180, p = 0.594. This means that the difference in the means could be a result of random factors and the paper cannot confidently conclude that these differences were caused by the dependent variable of offered help.

4.3 Trustworthiness

As with ABI scores that were previously discussed, the procedure of result processing followed is the same. The distribution data is generated for both subjectively and objectively measured trustworthiness and then the significance tests are performed accordingly.

The trend of negative skewness is also present in the trustworthiness distribution with skewness values of around -0.5, which is towards the lower values of skewness encountered in the results. The mean trustworthiness values for the experimental group are higher than those of the control group, as has been the case

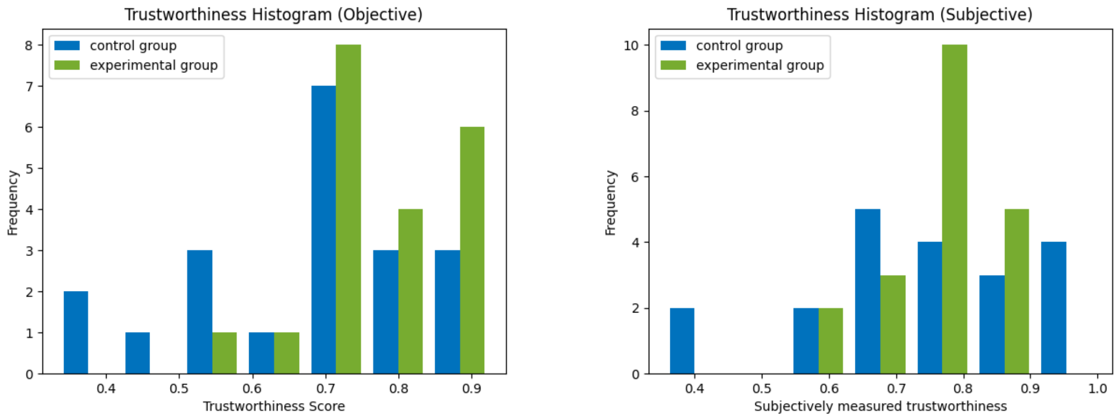

Figure 6: Histograms of objective ABI scores experimental vs. control group.

	Mean	St. Dev.	Var.	Skew.
Trustworthiness Distribution Control	0.667	0.171	0.029	-0.549
Trustworthiness Distribution Experimental	0.762	0.107	0.012	-0.58
Shapiro-Wilk Test	p < 0.523		p < 0.518	
T-test	t(38) = -2.093, p <0.043 (significant)			

	Mean	St. Dev.	Var.	Skew.
Trustworthiness Distribution Control	0.742	0.171	0.029	-0.506
Trustworthiness Distribution Experimental	0.774	0.097	0.009	-0.742
Shapiro-Wilk Test	p < 0.401		p < 0.143	
T-test	t(38) = -0.733, p <0.468 (insignificant)			

Table 10: Trustworthiness distribution data: objective (left) vs. subjective (right)

also when looking at the ABI components individually. However, in order to determine whether there is any statistical significance to this difference in means a Shapiro-Wilk test is performed to determine normality and then consequently a T-test is performed.

The Shapiro-Wilk test resulted in p-values that resident significantly above 0.05, thus we cannot reject the null hypothesis of normality, which means the distribution was considered to be normal, for objective and subjective trustworthiness scores. Subsequently, Student's T-test was performed on the control and experimental data sets. Interestingly, the T-test results differed between the objective and subjective measures. The 20 participants in the experimental group ($\mathrm{M}_{control} = 0.762, \mathrm{SD}_{control} = 0.107$) compared to the 20 participants of the control group ($\mathrm{M}_{expr} = 0.667, \mathrm{SD}_{expr} = 0.171$) demonstrated significantly better trustworthiness scores, $t(38) = -2.093, p = 0.043$. A p-value of 0.043 means that the alternative hypothesis is accepted. This result validates the research hypothesis that was phrased in section 1, which in short, states that offering help will improve human trustworthiness, in this case by roughly by 0.1. This is however, only partially supported by the data as the subjective perspective did not yield the same result. When performing the T-test on the subjective data the 20 participants in the experimental group ($\mathrm{M}_{control} = 0.774, \mathrm{SD}_{control} = 0.097$) compared to the 20 participants of the control group ($\mathrm{M}_{expr} = 0.742$, $\mathrm{SD}_{expr} = 0.171$) did not demonstrate a significant difference in the distribution means, $t(38) = -0.733, p = 0.468$. A p-value of 0.468 was obtained, which means that the data does not provide enough evidence to reject the null hypothesis. The results indicate that although the human trustworthiness as perceived by the AI-agent seems to be affected positively by the help, there is not enough evidence to prove the same for the human's perception of their own trustworthiness.

4.4 Correlation Between Help Frequency and Trustworthiness

To further explore the dynamics between the agent actions and the effect in human trustworthiness Pearson's Correlation Coefficient (Freedman, Pisani, & Purves, 2007) was calculated to test if there is a correlation between the number of times that the agents helps the human and the latter's trustworthiness score.

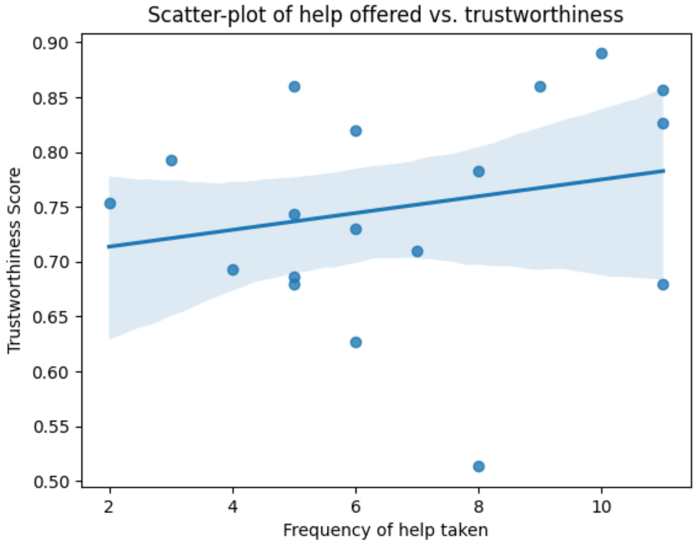

Figure 7: Scatter Plot of help frequency vs. trustworthiness.

As seen from Figure 7 there is a positive correlation as noted by the regression line of the scatter points. However, this seems to not be a very strong correlation. The Pearson's matrix that was obtained from the data was: $\begin{bmatrix} 1 & 0.195 \\ 0.195 & 1 \end{bmatrix}$. A correlation score of 0.195 indicates that although it can arguably be a positive correlation between the number of times that the human agent relies on the offered help and their trustworthiness score, the correlation is quite weak and therefore, negligible. This means that although offering help does increase human trustworthiness, it does not always mean that this increase is proportional to the increase in the frequency of help offered.

5 Responsible Research

The research abides by the ethical guidelines of TU Delft for responsible research. TU Delft's HREC guideline form is approved by the respective authority within the institution and a few ethical considerations were taken into account throughout the process of experimenting. The collected data from the experiment were not manipulated or altered.

The participants that took part in the research were made aware through a consent form of the risks involved in the experiment. The consent form was signed and approved by each participant. They were made aware beforehand of the personal information that was recorded, which was their age, nationality and some information on their language command and gaming experience. Through the consent form they were also informed about how this information about them was stored within the university and what was made public for further research.

The challenge itself does involve putting the human in a difficult setting given that the task revolves around rescuing injured people/animals. The concept behind the game may be unpleasant and evoke feelings of empathy guilt and stress, however, this is countered by the very simplistic and playful animation style of the game. Moreover, the participants were given the chance to withdraw from the experiment at any time in case they preferred to do so.

In terms of reproducibility of results the paper ensures a clear explanation of the methodology used as well as the result-processing methods. The experiment was conducted through an open source MATRX scenario, which can be accessed through GitHub. The implementation of the AI-agents are also described in detail. The repository with the respective implementation as well as the collected data is part of an open source repository and accessible via GitHub as well. (Delia, Obame, Dinu, Rademaker, & Lindhorst, 2022) A clear description of population profiles and the data collected from the participants is also discussed in the paper. Section 3, has the necessary information to enable the reproduction of the results.

6 Discussion and Limitations

Throughout this research the concept of trustworthiness was looked at from two different perspectives. It was modelled through the perception of AI and the human. The results of the paper demonstrated that the positive effect help has on human trustworthiness is clear through the objective metrics. Subjective metrics did not show the same significance in results. Practically speaking this shows that the agent's perceived and measured trustworthiness of a human was positively affected. On the other hand, the same claim cannot be made for the human's perception of their own trustworthiness. A reason for that could be the self-bias that the human has towards themselves. Research has shown that humans tend to have self-serving bias, meaning that they often try to escape responsibility when failing or performing badly and take more credit than they are due (Wang et al., 2017). At the same time, the questions included in the questionnaire arguably empower the self-enhancing bias. For example, the questions that aimed to measure integrity, mention keywords such as 'honour' and 'truth', which normally evoke such bias. This could also be the reason why participants scored the highest in the subjective measure of their integrity. Because of these observations, it seems that the actions of the agent in the context of the game have not had an effect at overcoming this bias and thus the subjective trustworthiness did not show a significant improvement.

The objective measures on the other hand seemed to back the hypothesis of the research that being offered help improves human trustworthiness. However, even in objective measures integrity did not show significant results, as opposed to ability and benevolence. This could be because of a few reasons. One being that in terms of implementable metrics, ability and benevolence had a few more dimensions upon which they were calculated, as opposed to integrity. Unlike in the questionnaire, in which each component of ABI was assigned 5 questions, the implementation of the game had a limited range of things that could be measured in the context of integrity, with truthfulness being chosen as the key sub-component of the latter. This is because there was not a lot of opportunities within the USAR scenario for the human to lie/tell the truth. Another reason, which is also backed by the fact that in both subjective and objective measures the mean integrity scores were high, could be that the participants trusted their partner. Both control and experimental agents were programmed to have a high trustworthiness in the sense that they were devoted to the task completion, they were honest and communicative. This behaviour might have instilled the human agents in both groups to do the same. That, combined with the desire to save the victims and complete the game, might have incentivized the participants to be truthful regardless of the presence or absence of the help factor.

Although the objective metrics provided significant evidence to support the hypothesis, it fell short in proving a strong correlation between the frequency of help and the trustworthiness. Thus, it can be concluded that it is not guaranteed that trustworthiness always increases with an increase in help frequency. The lack of a strong correlation here could also be related to the nature of what help was defined as in the context of the experiment, as a lot of the help buttons were one-use. Being offered and taking help more often can also be an indicator of lack of awareness, thus high reliance on help could mean the human agent is in some cases not capable to operate without the help received. Also, offered help does not ensure that the help given is absorbed by the human agent. This could explain the cases in which participants are offered a substantial amount of help but still their trustworthiness score remained subpar.

Another observation worth highlighting, was the negative skew that every distribution had. The overwhelming majority of the participants scored well above the 0.5 threshold within the MATRX game and very few of them scored below it. This can be a consequence of one of the confounding factors that were identified, namely the fact that the majority of the participants were young people and with considerable gaming experience. That, matched with the simplicity of the game, might have been the cause for the skewness that was observed. However, because the time frame of the research did not allow for a longer experimenting period, the demographics of the experiment turned out to be quite homogeneous. Given that each member of the research team was responsible with recruiting 24 participants, from which 4 would contribute to the control group population, the control group was more diverse than the experimental one in terms of age. This could have had an effect on the performance of the two groups as different demographics offer different skill-sets, which affects performance overall. Thus, a shortcoming of the research could arguably be the slight differences between the two populations. Future work could examine this hypotheses by producing more results on how exactly each age-group performed individually.

7 Future Work

One suggestion regarding future work is overcoming the participant population limitations discussed in section 6. A larger set of people, with a more diverse background as well as skill set will improve chances of obtaining normal distribution across experiments. That matched with a more complex task that offers more opportunities for exploration could provide more evidence as to how exactly the agent's actions affect human trustworthiness.

Another interesting perspective that could be explored in the future is to test how different trust models opt against the same research question. The SWIFT model was a trust model that was thoroughly considered, simply because of the fact that it is designed for settings where collaboration happens with no prior knowledge or experience between the two team-mates. In emergency rescue teams this is often the case, therefore, the SWIFT model seem like a reasonable choice to test within the USAR environment.

Furthermore, when using the ABI model there is also more to explore, especially when it comes to the objective definitions that are used to model the scores for the ABI components. For the limited scope of the research a simplistic approach was chosen to merely calculate the average of the different metrics to calculate the scores, however, this may not be the best approach. In order to determine exactly which measurements impact ability, benevolence and integrity scores and at what weights a Neural Network algorithm can be helpful. In absence of ground truth of human trustworthiness, the Neural Network could be trained to predict for artificial agents with a predefined trustworthiness value that determines their behavior within the environment. This could enable future research to better identify which agent actions have the most impact in defining the trustworthiness scores, and the same model could then be applied to human participants.

Another insightful and challenging experiment to further deepen understatement of help and the effect it has on trustworthiness could be to test for bigger and more diverse teams. For example research could be conducted in collaborative teams of humans, where the AI agent is merely active in providing the help as opposed to solving the tasks themselves. This could help to further understand how help affects trustworthiness of the human agents when more variables are involved and not every teammate has the same qualities when it comes to ability, benevolence and integrity.

8 Conclusion

In general the research was designed to simulate team work between an AI agent and a human agent. This was done through a search and rescue scenario in which humans teamed up with an AI system in order to complete the search and rescue scenario. The actions that the agents undertook during the performance of the task as well as communication between them was logged. As described, trustworthiness was calculated through in-game metrics which represented objective metrics. A questionnaire was also used to measure these metrics subjectively. In both cases trustworthiness was measured using the ABI trust model. The ability, benevolence and integrity of the human agent were measured and discussed individually. They were also aggregated into a score that represented trustworthiness. The participation of 2 populations, the control group and experimental group, both with a size of 20, allowed to test the same scenario on two different AI agents. A baseline agent as well as a helper agent which implements the dependent variable of offering help, were used to explore the research question.

This research contributed in discovering the effect that AI agent's actions have on the trustworthiness of their human teammate. The results demonstrated that trustworthiness when measured through the AI agent's perspective (objective) increased when the human agent was offered help. Ability and benevolence of the human agents were the qualities most affected by the help, as they were more willing to communicate back with the agent and got a better understanding of the task altogether. The trustworthiness showed a weak positive correlation between the amount of help and trustworthiness, which indicates that more help leads to higher trustworthiness, however, the correlation value was low and therefore it is not a significant result. When measured through their own subjective opinion, no significant change in trustworthiness was detected. Being offered help by their AI teammate does not affect the human's perception of their trustworthiness.

These findings are important because it helps teams that work on development of AI to better understand how these systems should be build so that they inspire improvement for the humans that collaborate with them. It also fills the literature gap of exploring how the teamwork dynamics influence trustworthiness of the human. Literature is mostly written around how to make artificial intelligence systems more trustworthy, but it is also important for these systems to understand their team-mate as well. This way AI could be trained to adjust their behavior and task delegation strategies to account for their human teammate's qualities. In conclusion, offering help tends to have positive impact on the human teammate when the trustworthiness is measured objectively, but does not significantly impact subjectively measured trustworthiness.

References

Bradshaw, J., Dignum, V., Jonker, C., & Sierhuis, M. (2012, 03). Human-agent-robot teamwork. In (Vol. 27, p. 487-487). doi: 10.1109/MIS.2012.37

Delia, A., Obame, C., Obiang, Dinu, I., Rademaker, J., & Lindhorst, P. (2022). *Towards trust.* https://github.com/plindhorst/Towards-Trust. GitHub.

Falcone, R., & Castelfranchi, C. (2004, 02). Trust dynamics: How trust is influenced by direct experiences and by trust itself. In (Vol. 2, p. 740- 747). doi: 10.1109/AAMAS.2004.286

Freedman, D., Pisani, R., & Purves, R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Hardin, R. (2002). Trustworthiness. In *Trust and trustworthiness* (pp. 28–53). Russell Sage Foundation. Retrieved 2022-05-30, from http://www.jstor.org/stable/10.7758/9781610442718.7

Haring, K., Phillips, E., Lazzara, E., Ullman, D., Baker, A., & Keebler, J. (2021, 01). Applying the swift trust model to human-robot teaming. In (p. 407-427). doi: 10.1016/B978-0-12-819472-0.00017-4

Heineman, R. A. (1984). The logic and limits of trust. by bernard barber. (new brunswick, n.j.: Rutgers university press, 1983. pp. 190. 27.50,*cloth*;9.95, paper.). *American Political Science Review*, *78*(1), 209â210. doi: 10.2307/1961263

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (p. 624â635). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3442188.3445923 doi: 10.1145/3442188.3445923

Johnson, M., & Bradshaw, J. M. (2021). How interdependence explains the world of teamwork. In W. F. Lawless, J. Llinas, D. A. Sofge, & R. Mittu (Eds.), *Engineering artificially intelligent systems: A systems engineering approach to realizing synergistic capabilities* (pp. 122–146). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-89385-9_8 doi: 10.1007/978-3-030-89385-9_8

Kiesler, S., & Goetz, J. (2002, 01). Mental models of robotic assistants. In (p. 576-577). doi: 10.1145/506443.506491

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of trusted autonomy* (pp. 135–159). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-64816-3_8 doi: 10.1007/978-3-319-64816-3_8

Mann–whitney test. (2008). In *The concise encyclopedia of statistics* (pp. 327–329). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-0-387-32833-1_243 doi: 10.1007/978-0-387-32833-1_243

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. Retrieved 2022-04-22, from http://www.jstor.org/stable/258792

Razmerita, L., Brun, A., & Nabeth, T. (2021, 10). Collaboration in the machine age: Trustworthy human-ai collaboration.. doi: 10.1007/978-3-030-93052-3_14

Salas, E., Sims, D., & Burke, S. (2005, 10). Is there a âbig fiveâ in teamwork? *Small Group Research*, *36*, 555 -599. doi: 10.1177/1046496405277134

Shapiro, S. S., & Wilk, M. B. (1965, dec). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. Retrieved from https://doi.org/10.1093/biomet/52.3-4.591 doi: 10.1093/biomet/52.3-4.591

Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

van der Waa, J., & Haije, T. (2022). *Matrx software.* https://github.com/matrx-software/matrx. GitHub.

Verhagen, R. (2022). *Usar-hat.* https://github.com/rsverhagen94/TUD-Research-Project-2022. GitHub.

Wang, X., Zheng, L., Li, L., Zheng, Y., Sun, P., Zhou, F. A., & Guo, X. (2017). Immune to situation: The self-serving bias in unambiguous contexts. *Frontiers in Psychology*, *8*. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2017.00822 doi: 10.3389/fpsyg.2017.00822