Human trustworthiness when collaborating with a friendly agent

Justin Rademaker
Supervisor(s): Carolina Jorge, 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

As technology advances, automated systems become more autonomous which leads to a higher interdependence between machine and human. Much research has been done about trust between humans and trust of humans regarding machines. An interesting question that remains is how the behavior of an agent influences human trustworthiness in a human-agent collaborative setting. The research presented by this paper contributes to the understanding of this area. It investigates a specific behavioral trait using the following hypothesis: friendly behavior of an agent improves human trustworthiness. Here, trustworthiness is broken up in the constructs: ability, benevolence and integrity.

An experiment has been conducted using a collaborative Search and Rescue game. The following behaviors of the participants have been measured:

- Ability: speed and effectiveness;
- Benevolence: communication, willingness to help, agreeableness to advice, responsiveness;
- Integrity: truthfulness.

Furthermore, a likert scale has been used to measure the participants' own perception of their trustworthiness. The experiment is conducted with 20 participants in the control group, where the agent spoke in a neutral manner, and 20 in the experimental group, where the agent instilled empathy, stimulated collaboration, encouraged the participants and was affectionate.

The research has shown a significant improvement in the experimental group only for communication and willingness to help. This gives some indication that a friendly agent only slightly improves the trustworthiness of a human. However, the research has some limitations that might also explain the lack of significant results. Firstly, it's unclear to what extent the measures truly measured the constructs of trustworthiness. Secondly, to create a friendly agent, theories from organizational and social psychology are used, which are mostly focussed on human-human relationships, instead of human-agent relationships. Finally, Some confounding variables may have had an impact, like lag in the game and the participant not properly reading the agent's messages.

1 Introduction

In order for collaboration between human and technology to result in smooth work and play, proper support and considerations for the interdependent relationship is required (Johnson & Bradshaw, 2021). One manifestation of technological advancement is an 'intelligent agent' (hereafter: 'agent'). This is a computer system that can act autonomously in an environment in order to reach its goals. The agent is proactive, reactive and social (Weiss et al., 2013).

A central component of collaboration between agents and humans is that both should have an appropriate level of trust in each other, meaning that the amount of trust should match the actual trustworthiness (Jacovi et al., 2021). Given the autonomous nature of both humans and agents, trust and trustworthiness are not guaranteed. Trust and trustworthiness have been researched from several perspectives.

One perspective is that of human-human relationships, which psychologists have focussed

on for decades. Several overlapping concepts have come up. In 1983, Barber proposed that trust is based on the human expectations of persistence (i.e. stability of the trustee), technical competency (i.e. expertise and performance) and fiduciary responsibility (i.e. moral behavior to uphold a contract). Later in 1995, Mayer et al. proposed the ABI-model, where trust is considered to be perceived trustworthiness, which consists of ability (i.e. capability of the trustee), benevolence (i.e. wanting to do good to the trustor) and integrity (i.e. having shared values). In 2004, Falcone and Castelfranchi proposed competence belief and willingness belief as dimensions of trust. All these theories agree on trust being a perception of trustworthiness. Several other conceptual overlaps are:

- ability and (technical) competence;
- fiduciary responsibility and integrity;
- willingness and benevolence.

Another perspective is that of human-computer relationships, where the emphasis lies on humans trusting computers (instead of the other way around). In 2021, Jacovi et al. described trust towards computers as perceived trustworthiness also. However, they added the notion of vulnerability and trust being context dependent. Namely, only relevant in the context of a specific contract (i.e. what specifically the computer can be trusted to do). Later in 2019, Gulati et al. found that Perceived risk, benevolence and competence are good ways to measure trust in a computer. Human trust towards computers has some overlap with trust amongst humans. However, more focus is placed on risk and context.

Most of the mentioned research focuses on the process of trust forming and perceived trustworthiness between humans and of humans towards computers. Little research has been devoted to the concept of actual trustworthiness. These concepts are not the same. For example, someone can form inappropriate trust, meaning that the trust level does not match the actual trustworthiness and this can lead to overreliance or inappropriate distrust (Parasuraman & Riley, 1997). An interesting idea is if actual trustworthiness of a partner can be increased or decreased by behaving in a certain way during collaboration. Secondly, little research has been devoted to human-computer relationships where the focus lies on the computer being able to trust the human, instead of the other way around. Now, an interesting question is how the behavior of different agents influences human trustworthiness in a collaborative setting. Knowledge in this area would help to design agents that enhance human trustworthiness and thereby the quality of collaboration. The research presented by this paper contributes to the understanding of this area. It investigates a specific behavioral trait using the following hypothesis: *Friendly behavior of an agent improves human trustworthiness.*

A trustworthiness-model that is widely used, is focused on humans and provides a nice framework for trustworthiness as a concept of its own, is the aforementioned ABI model. Therefore, trustworthiness in this research will be modeled as ability, benevolence and integrity.

With regard to friendliness: findings in social psychology show that certain prosocial behaviors in human-human relationships increase aspects of trustworthiness in either one or both of the humans, where prosocial behavior comprises all actions aimed at the well-being of others (Aronson et al., 2020). In this research, friendly behavior is considered to be

prosocial behavior, with the limitation that the agent may not actually help the participant. Since helpfulness is an interesting trait to investigate on its own, it has been decided for this research to isolate and investigate a passive form of friendliness. Finally, attempts to bond and collaborate with the participant is considered to be friendly behavior, since this could increase the well-being of the participant (for example by feeling appreciated or performing well on the game).

Section 2 of this paper describes the methodology used in this research to test the hypothesis. It describes how the experiment has been designed. The concept of friendliness is further defined and it explains how the data has been collected. Section 3 explains how the data has been analyzed and demonstrates the results of the research. section 4 discusses the research from an ethical perspective. In section 5, this research is placed within a broader context and its contributions and limitations are reflected upon. Finally, section 6 concludes the research and provides implications for further research.

2 Methodology

This section describes how the results have been obtained. Since the hypothesis is aimed at finding a causal relationship between friendliness (independent variable) and trustworthiness (dependent variable) an experimental setup has been chosen to test it. This section will first go into the experimental design of the research. Herein, the implementation of friendliness will be described. Then it will describe which data has been collected in what way.

2.1 Experimental design

2.1.1 The game

An experiment has been conducted in order to test the hypothesis. The experiment consists of a collaborative search and rescue game between a human and robot, using the MATRX platform (MATRX, 2022). The specific game is credited to Verhagen (2022) and shown in figure 1. The aim of the game is to save eight diseased victims by searching for them, picking them up, bringing them to a drop-off point and dropping them off in a specific order. The participant and agent can communicate to each other via buttons and chat. They can say where they are searching, who they have found and who they are dropping off. Furthermore, the agent will ask the participant for help with identifying or carrying a victim several times throughout the game. The game starts with a tutorial page, where the game gets explained. The experimenter walks each participant through this page in a similar manner.

Figure 1: The Search and Rescue game using MATRX software (Verhagen, 2022)

The experiment has been carried out with a control and experimental group of each 20 participants. The control group has played the standard implementation of the game. Here,

the agent speaks in a neutral manner. Throughout the game, the agent has shared some of its limitations with the participant and has asked for help. The agent also suggested several clever actions for the human to take. The experimental group contains the same environment. However, the agent's messages have been adapted in order to make it more friendly. The friendly behavior is specified in the following section.

2.1.2 friendly behavior in the experimental group

As mentioned before, this research will focus on prosocial behavior of the agent without actually helping the participant. Therefore, in the experiment, friendliness is exclusively implemented through prosocial chat-messages. This section presents the findings in organizational and social psychology that are used to implement a friendly agent that fits this profile. Figure 2 shows which theories attempt to improve which part of trustworthiness.

2.1.2.1 Empathy and social exchange theory

In 2020, Aronson et al. mention several research papers to support the concept of the social-exchange theory. This states that people try to maximize the benefits of relationships and minimize the costs. This means that people will show prosocial behavior when they expect to 'profit' from it. Benefits can be intangible like gratitude or self-exaltation. In 2011, Batson published the empathy-altruism hypothesis.This states that whenever a person feels empathy for someone, they are inclined to show prosocial behavior towards this person. If not, their behavior will follow the social exchange theory. When an agent is able to make the human see the benefits of collaboration or make him/her empathize with it, then this would increase the benevolence towards the agent. The implemented behaviors that follow are considered to be friendly because they are attempts to bond and collaborate with the participant.

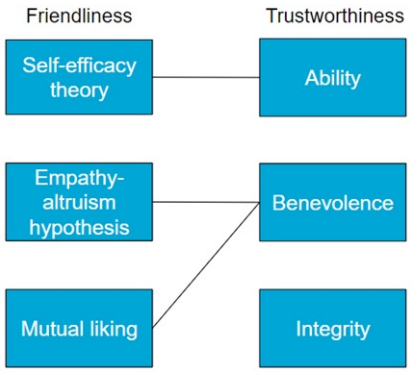

Figure 2: On the left are the theories that are used in the implementation of the friendly agent. On the right are the constructs of trustworthiness. The links show which theories attempt to improve which constructs.

Two techniques are used by the agent to make the participant empathize with it. Firstly, Manney described in 2008 how storytelling increases the empathy of the readers. This is mostly because the readers will be temporarily forced to perceive the world through the perspective of the characters. Furthermore, Snyder and Sturmer described in 2010 that empathy for someone will likely be higher when that person is considered to be in their in-group. Meaning, that this person is considered to be part of the same community.

These techniques are implemented as follows. Before starting the experiment, the agent will tell a story of how the situation came to be. In short, the agent tells the participant that it's on a well deserved vacation with its family, after serving 'their personal humans' for a year. But then a pandemic struck (of a computer virus), and it now needs the par-

ticipants' help to save its family. Then, the agent will ask the participant for a name, age and birthplace. The agent will respond by telling the participant that 'its personal human' shares the same demographics and thereby making them part of the same group.

Finally, in order to motivate the participant into collaboration in the absence of empathy, the agent first reminds the participant that he/she is allowed to lie, to be lazy, to not communicate, or do whatever he/she sees fit. Then the agent emphasizes however that in order to achieve the best performance, they should work together as well as possible.

2.1.2.2 The self-efficacy theory and mutual liking

In 1986, Bandura and Cervone developed the self-efficacy-theory which indicates that the more confidence a person has in its ability, the more effort it will put into properly overcoming a challenge. Several ways have been proposed to increase confidence, one of which is relevant for this research. Namely: encouragement. Or in other words, trying to convince someone that they are capable. This theory has an interesting application in the context of trustworthiness, namely to increase the ability of the human.

Another theory is about the concept of mutual liking, which is supported by Aronson et al. (2020) with several research papers. This means that whenever we believe that someone likes us, we tend to like them more as well. This might suggest that whenever the agent expresses that it likes the user, that the benevolence of the user may increase.

In order to boost confidence, the agent will tell the participant that it did some calculations, and found out that participants with the given demographics perform 78% better at rescue missions. Furthermore, every time the participant sends a message about picking up a victim, the agent sends a different message that is either encouraging or affectionate.

The encouraging and affectionate behavior are considered to be friendly because they attempt to improve the well being of the participant (their confidence and feeling of being appreciated).

2.1.3 Confounding variables and participants

Two possible confounding variables have been identified for this experiment: game experience and language proficiency. Language proficiency has been controlled for by only accepting participants with a high proficiency (with one exception in the control group). In order to expand the pool of participants, the experiment and questionnaire has been implemented in Dutch as well. Game experience and miscellaneous confounding variables have been controlled for by including a similar distribution in both the control and experimental group. Figure 3 shows the demographics of all participants.

2.2 Measures

To test the hypothesis, the trustworthiness of the control and experimental group has to be measured and compared. To this end, objective and subjective metrics have been constructed which measure the participant's ability, benevolence and integrity.

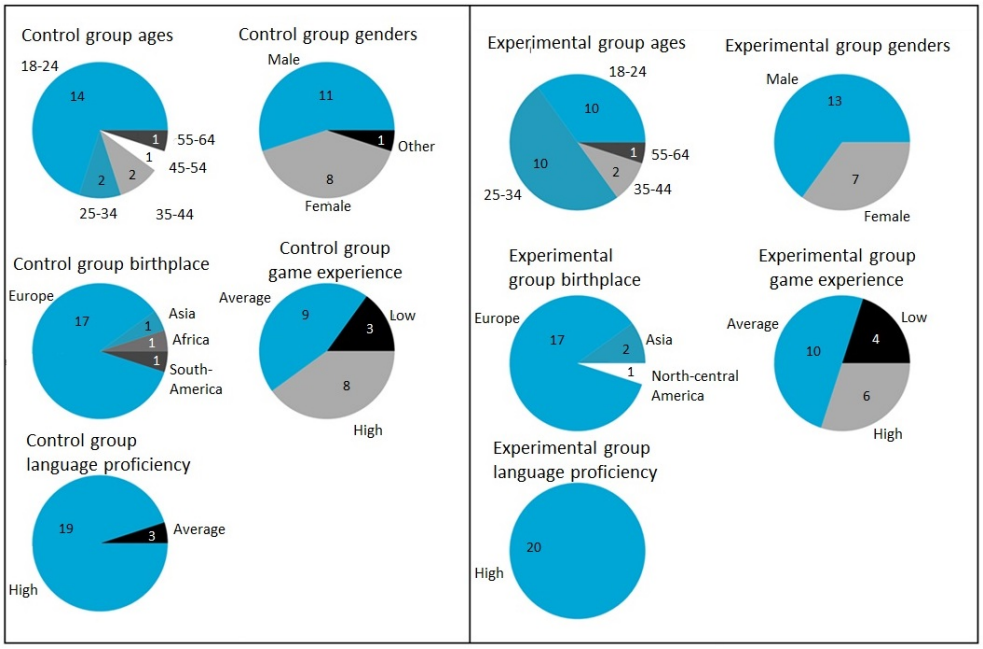

Figure 3: The demographics of the participants. Left is from the control group. Right is from the experimental group.

2.2.1 Objective measures

During the game, several results and behaviors of the participants have been measured and normalized into a value in the range of [0, 1]. For speed and responsiveness, the following formula has been used for normalization:

$$Normalized\ value = \left| \frac{amount\ of\ ticks\ minimal\ amount\ of\ ticks\ in\ the\ group}{maximal\ amount\ of\ ticks\ in\ the\ group - minimal\ amount\ of\ ticks\ in\ the\ group} \right|$$

All the other metrics have been calculated using ratios that bring them in the range of [0, 1]. The metrics have been aggregated into different scores, which are all part of either ability, benevolence or integrity. Table 1 shows all metrics and their aggregations. For the analysis, the scores have been calculated by taking the average of the corresponding metrics. In turn, the constructs of trustworthiness (ability, benevolence and integrity) have been calculated by taking the average of the corresponding scores. This way, each score has an equal weight.

2.2.2 Subjective measures

The experiment ends with a questionnaire. First, gender, age, homeland, experience in computer games and experience in the language is asked. Then, five questions are asked for each trustworthiness construct (ability, benevolence and integrity) using a 7-point likert scale ranging from 'Never' to 'Always'. Here, the participant judges his/her own performance.

Although each individual question of a likert scale yields ordinal data, when four or more questions are aggregated, this result can be considered to be interval data (Scribrr, 2022).

Metric	Score	Construct
Amount of ticks passed at the end of the game	speed	Ability
Ratio: amount of victims saved / total amount of victims (i.e. 8)	Effectiveness	
Ratio: amount of times victim found / total amount of victims (i.e. 8)		
Ratio: amount of times victim picked up / total amount of victims (i.e. 8)		
Ratio: amount of times room visited by human / total amount of rooms (i.e. 9)		
Ratio: amount of communicated 'yes' to a suggested pickup / total amount of pickup suggestions by agent	Communication	Benevolence
Ratio: amount of times communicated that a room will be searched / amount of visited room entries		
Ratio: amount of times communicated that a victim will be picked up / total amount of picked up victims		
Ratio: amount times communicated that a victim is found / total amount of times a victim is found		
Ratio: amount of times a gender is communicated / amount of times agent asks about gender	Willingness to help	
Ratio: amount of times human directly picks up a victim after the agent advices it / amount of times agent advices a pick up	Agreeableness to advice	
Average amount of ticks it takes to respond to a question of the agent	Responsiveness	
Ratio: amount of times a gender is communicated which is correct / amount of times a gender is communicated	Truthfulness	Integrity
Ratio: amount of times communicated 'yes' for a suggested pickup and actually picking up that victim /amount of times communicated 'yes' for a suggested pickup		
Ratio: amount of times that a victim has been communicated to be found which was correct / amount of times that a victim has been communicated to be found.		
Ratio: amount of times communicated that a room will be searched, which was then actually searched / the amount of times communicated that a room will be searched.		
Ratio: amount of times communicated that a victim will be picked up which was followed through / amount of times communicated that a victim will be picked up.		

Table 1: The objective metrics that have been measured during the experiment (column 1), the scores that they make up by taking the average of the corresponding metrics (column 2) and the construct categories they belong to (column 3).

To reach this point, answer options have been encoded. 'Never' has an encoding of 0 and for each following option the encoding is incremented by one. Next, for each construct, the average of the encodings of its questions is taken. Finally, the average score is divided by 6 (i.e. the maximum encoding value) in order to normalize it into a value in the range of [0, 1] between zero and one.

The questionnaire has been taken from Centeio Jorge et al. (2022) where it was used in a similar research. Several irrelevant questions have been replaced, while keeping the intention of the original questions the same. Table 2 presents the internal consistency of the questionnaire according to the Cronbach's Alpha statistic in both the control group and experimental group.

Trustworthiness construct	Internal consistency
Ability (questions [6, 10])	0.76
Benevolence (questions [11, 15])	0.80
Integrity (questions [16, 20])	0.88

Table 2: The internal consistency of the questionnaire calculated with Cronbach's Alpha over all 40 participants of both the control and experimental group.

3 Results

This section describes the analysis, results and statistical tests of the objective and subjective measures.

3.1 Objective measures

To get a single score of trustworthiness for the control and experimental group, the raw data should be aggregated in some way. There is no obvious way to do this and the process is somewhat arbitrary. Therefore, the data is aggregated as little as possible and this section presents the results of all scores shown in table 1. This way, something can be said about specific parts of trustworthiness and the reader can make its own deductions. In order to still give some indication about ability, benevolence and integrity, the averages (over the datasets) of the corresponding scores have been taken and are shown. The same is done for trustworthiness by again taking the average (over the datasets) of ability, benevolence and integrity. Table 3 shows the descriptive results of the objective measures.

	Control group			Experimental group		
Scores	Median	Mean	Standard deviation	Median	Mean	Standard deviation
speed	0.39	0.38	0.3	0.39	0.37	0.31
Effectiveness	0.88	0.85	0.16	0.86	0.86	0.12
Ability	**0.64**	**0.62**	**0.19**	**0.65**	**0.62**	**0.15**
Communication	0.59	0.51	0.26	0.78	0.73	0.16
Willingness to help	1.0	0.69	0.38	1.0	1.0	0.0
Agreeableness to advice	0.0	0.30	0.46	0.0	0.25	0.43
Responsiveness	0.81	0.63	0.39	0.62	0.46	0.39
Benevolence	**0.5**	**0.53**	**0.23**	**0.64**	**0.61**	**0.19**
Truthfulness	0.75	0.70	0.25	0.72	0.68	0.22
Integrity	**0.75**	**0.70**	**0.25**	**0.72**	**0.68**	**0.22**
Trustworthiness	**0.63**	**0.62**	**0.19**	**0.68**	**0.64**	**0.15**

Table 3: shows the median, mean and standard deviation resulting from the experiments in both the control and experimental group (20 participants for each). The scores are found by taking the average of the corresponding metrics. The constructs (ability, benevolence and integrity) are found by taking the average of the corresponding scores over the dataset. Trustworthiness has been found by taking the average of the constructs over the dataset.

To decide how to test for a significant difference, first every score is checked for normality, using the shapiro-wilk test with an Alpha value of 0.05. Meaning that a score is assumed to be normal when the P-value ≥ 0.05. Table 4 shows the results of the test.

	Control group		Experimental group	
Scores	W-value (test statistic)	P-value	W-value (test statistic)	P-value
speed	0.929	0.146	0.915	0.079
Effectiveness	0.811	0.001	0.964	0.625
Ability	**0.979**	**0.918**	**0.946**	**0.317**
Communication	0.885	0.021	0.818	0.002
Willingness to help	0.751	0.000	1.000	1.000
Agreeableness to advice	0.580	0.000	0.544	0.000
Responsiveness	0.755	0.000	0.788	0.001
Benevolence	**0.975**	**0.861**	**0.921**	**0.103**
Truthfulness	0.832	0.003	0.917	0.086
Integrity	**0.832**	**0.003**	**0.917**	**0.086**
Trustworthiness	**0.975**	**0.854**	**0.918**	**0.092**

Table 4: The results of the shapiro-wilk tests. An Alpha value of 0.05 is used, meaning that a normal distribution may be assumed when the P-value ≥ 0.05.

In case of assumed normality, a two-sided independent T-test is performed on each score to compare the control and experimental group. For the T-test, no test for homogeneity of variance has been conducted since the T-test generally also works without that assumption, as long as the sample sizes are equal (Statistics Solutions, 2022). In case of no assumed normal distribution, the two-sided Mann-Whitney U test has been used. Table 5 shows the results of the tests. An Alpha value of 0.05 is used, meaning that a test can be considered to be significant when the P-value < 0.05.

As can be seen from table 3 and 5, the communication score and willingness to help show a significant improvement in the experimental group compared to the control group. Other than that, no significant difference has been found.

	Mann-whitney U test		independent T-test	
Scores	Test statistic	P-value	Test statistic	P-value
speed	-		0.114	0.910
Effectiveness	212.0	0.756	-	
Ability	**-**		**0.018**	**0.986**
Communication	88.0	0.003	-	
Willingness to help	110.0	0.001	-	
Agreeableness to advice	210.0	0.740	-	
Responsiveness	258.5	0.110	-	
Benevolence	**-**		**-1.159**	**0.254**
Truthfulness	221.5	0.570	-	
Integrity	**221.5**	**0.570**	-	
Trustworthiness	**-**		**-345**	**0.732**

Table 5: The results of the Mann-Whitney U tests and T-tests. The second one is used with 38 degrees of freedom in case that the data is normally distributed. Otherwise, the first one is used with 40 data points. An Alpha value of 0.05 is used, meaning that a difference is considered significant when the P-value < 0.05.

3.2 Subjective measures

Table 6 shows the scores that resulted from the questionnaires. For each construct, the same statistical procedure as for objective measures has been followed of which the results can be found in tables 7 and 8. As can be seen, no significant difference has been found.

	Control group			Experimental group		
Scores	Median	Mean	Standard deviation	Median	Mean	Standard deviation
Ability	0.7	0.73	0.14	0.72	0.69	0.16
Benevolence	0.68	0.68	0.26	0.72	0.7	0.19
Integrity	0.83	0.82	0.18	0.83	0.78	0.18
Trustworthiness	**0.74**	**0.74**	**0.17**	**0.76**	**0.73**	**0.15**

Table 6: shows the median, mean and standard deviation resulting from the experiments in both the control and experimental group (20 participants for each). The scores are found by taking the average of the answer encodings and dividing them by 6. Trustworthiness has been found by taking the average of the constructs over the dataset.

	Control group		Experimental group	
Scores	W-value (test statistic)	P-value	W-value (test statistic)	P-value
Ability	0.957	0.478	0.947	0.330
Benevolence	0.911	0.662	0.941	0.249
Integrity	0.861	0.008	0.872	0.013
Trustworthiness	**0.952**	**0.401**	**0.945**	**0.294**

Table 7: The results of the shapiro-wilk tests. An Alpha value of 0.05 is used, meaning that a normal distribution may be assumed when the P-value ≥ 0.05.

	Mann-whitney U test		independent T-test	
Scores	Test statistic	P-value	Test statistic	P-value
Ability	-		0.730	0.470
Benevolence	-		-0.285	0.777
Integrity	221.0	0.721	-	
Trustworthiness	**-**		**-0.330**	**0.577**

Table 8: The results of the Mann-Whitney U tests and T-tests. The second one is used with 38 degrees of freedom in case that the data is normally distributed. Otherwise, the first one is used with 40 data points. An Alpha value of 0.05 is used, meaning that a difference is considered significant when the P-value < 0.05.

4 Responsible Research

This section reflects on the ethical aspects of this research and discusses the reproducibility of the used methods.

4.1 Ethics

Several measures have been taken in order to guard an ethical research process.

Firstly, the participants of the experiment have been asked to sign a consent form. Here, the risks of the experiment are explained. The form informs the participant that all data is anonymised, and only the demographics remain. Furthermore, it explains that the experiment is voluntary but that the data will not be removed afterwards. This way, the participants could decide to participate with full understanding of the consequences. It should be noted that in the experimental condition, one small lie was told. Namely that participants of their demographics perform 78% better on average. After the experiment, the experimenter explained to the participants that this was a lie and why it was implemented.

Furthermore, the data integrity has been guarded. For example, due to some bugs in the first few experiments, participants could not choose all options in the questionnaire. After the experiment, the files have been adapted manually in consultation with the participant. This way, data manipulation was avoided. Furthermore, a backup of the data has been made throughout the research. A few times, data seemed to be lost due to merging issues.

The data could simply be recovered from the backup. This way, there was no temptation to fabricate data by mimicking the participant's behavior.

4.2 Reproducibility

Since the participants played the game for the first time, an explanation was required beforehand. Within this research, the explanation was given in a similar manner to all participants. However, if a new researcher attempts to reproduce the results, details of the explanations may differ. As an attempt to make these explanations consistent, every experiment is preceded by a tutorial page as a guideline for the experimenter.

Furthermore, the experiment stored all raw data in pickle-files for their actions and json-files for their questionnaire responses. This way, colleague researchers could do their own analyses on this data and achieve the same results.

Finally, all data collection and data analysis has been automated in python. This code may be shared with colleague researchers. This way, they can reproduce the research with little effort. Furthermore, in the case that any mistakes have been made in this research, they can most likely openly be found in the form of bugs in the program.

5 Discussion

The results in section 3 show no significant differences for the constructs of trustworthiness with the exception of the communication score and willingness to help. Surprisingly, on average the participants often scored slightly worse on the experimental condition than on the control condition. This section reflects on possible explanations for the results.

5.1 Validity of the metrics

The experimenter has guided and observed every experiment of the experimental group and four experiments of the control group. It was observed that certain metrics did not always represent their construct properly. For example, when the participant did not follow the advice of the agent, it was often because he/she did not read the message, rather than lack of benevolence. Although an argument could be made that higher benevolence would lead to higher effort to read the messages, it could also be argued that this is part of ability. Due to the lack of validity of the metrics, it might be that trustworthiness of the participant of the experimental group was higher, but this research has been unable to measure it.

5.2 The experimental condition

In order to make the agent friendly, several theories from social and organizational psychology have been used (see figure 2). The impact on trustworthiness for these theories are supported by evidence for human-human relationships. However, it is still unclear if these theories also hold when one human gets replaced by an agent. This research gives some indication that applying the Empathy-altruism hypothesis and/or mutual liking improves the human's communication and willingness to help. Other than that, no other results provide evidence to support the idea. The interesting question remains if a different implementation of friendliness exists that does significantly increase the human's trustworthiness.

5.3 confounding variables

The experimental game had a slow response time to the participant's actions (i.e. it was laggy). Different participants responded very differently. Some got annoyed, distracted or bored whereas others were patient and enjoyed the obstacle as an extra challenge. Furthermore, the interventions of the experimental agent relied heavily on the participant reading the chat properly. Instead, many participants skipped the chat or skimmed through it rather quickly. Also, some participants never pushed the 'I will pick up' buttons which triggered the encouraging messages. Others did, but barely read the messages. Personality traits like conscientiousness, attentiveness and patience have not specifically been controlled for and may have influenced the data.

Finally, before starting the experiments, the experimenter explained to the participant how the game worked. It might well be that the experimenter emphasized for example communication and the chatbox more than the experimenters that contributed to the control group. The different explanations of the game might explain the significant difference in the communication score and willingness to help the agent.

6 Conclusions and Future Work

The hypothesis of this research has been the following: Friendly behavior of an agent improves human trustworthiness. 40 collaborative games have been played in order to compare a neutral agent to one that tries to instill empathy, shows the benefits of collaboration, is affectionate and encouraging. The results show a significant increase of two out of ten measures, namely: communication and willingness to help the agent. This gives some evidence in favor of the hypothesis. However, eight out of ten scores showed no significant difference. Therefore, this research gives insufficient evidence to confidently conclude that a friendly agent improves human trustworthiness in a collaborative setting.

More research has to be conducted in order to verify this result. Firstly, some research is missing regarding which behaviors of a computer is regarded as friendly by humans, since this research gives some indication that it might be different from desirable human behaviors. Secondly, objective measures for trustworthiness are hard to find. Future research could focus on metrics that properly measure trustworthiness. For example, this could be done by conducting experiments where participants in different groups purposely act either trustworthy or untrustworthy, to see if the metrics pick up on that. Finally, some research could be done to improve the experimental design which would reduce the amount of confounding variables. For example, the gameplay could be made to be more fluent, or the chat box could be replaced by a (friendly) voice.

References

Aronson, E., Wilson, T., Akert, R., & Sommers, S. (2020). *Social psychology.* Pearson Education Limited.

Bandura, A., & Cervone, D. (1986). Differential engagement of self-reactive influences in cognitive motivation. *Organizational behavior and human decision processes*, *38*(1), 92–113.

Barber, B. (1983). *The logic and limits of trust.* Rutgers University Press.

Batson, C. D. (2011). *Altruism in humans.* Oxford University Press.

Centeio Jorge, C., Tielman, M. L., & Jonker, C. M. (2022). Artificial trust as a tool in human-ai teams. In *Proceedings of the 2022 acm/ieee international conference on human-robot interaction* (pp. 1155–1157).

Falcone, R., & Castelfranchi, C. (2004). Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the third international joint conference on autonomous agents and multiagent systems, 2004. aamas 2004.* (pp. 740–747).

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015.

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* (pp. 624–635).

Johnson, M., & Bradshaw, J. M. (2021). How interdependence explains the world of teamwork. In *Engineering artificially intelligent systems* (pp. 122–146). Springer.

Manney, P. J. (2008). Empathy in the time of technology: How storytelling is the key to empathy. *Journal of Evolution & Technology*, *19*(1).

MATRX. (2022). *Matrx â accelerating human-agent teaming research together.* Retrieved from `https://matrx-software.com/`

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230–253.

Scribrr. (2022). *Designing and analyzing a likert scale | guide examples.* Retrieved from `https://www.scribbr.com/methodology/likert-scale/?_ga=2.108990914.1712786254.1652292094-1197528142.1652292093`

Snyder, M., & Sturmer, S. (2010). *The psychology of prosocial behavior: Group processes, intergroup relations, and helping.* Wiley.

Statistics Solutions. (2022). *The assumption of homogeneity of variance.* Retrieved from `https://www.statisticssolutions.com/the-assumption-of-homogeneity-of-variance/`

Verhagen, R. (2022). *Tud-research-project-2022.* Retrieved from `https://github.com/rsverhagen94/TUD-Research-Project-2022`

Weiss, G., Conitzer, V., Wooldridge, M., Dignum, V., Padget, J., Chipra, A. K., ... Brandt, F. (2013). *Multiagent systems.* MIT Press.