Towards Trust in Human-AI Teams

Paul Lindhorst
Supervisor(s): Carolina 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

TU Delft

Towards Trust in Human-AI Teams

Paul Lindhorst*

Delft University of Technology,
The Netherlands

June 2022

Abstract

Human-AI teams require trust to operate efficiently and solve certain tasks like search & rescue. Trustworthiness is measured using the ABI model; Ability, Benevolence and Integrity. This research paper tries to observe the effect a conflicting robot has on the human trustworthiness. The hypothesis we try to test is: *"human trustworthiness will decrease when paired with a conflicting AI"*. We conduct an experiment with one control group playing with a normal agent and an experiment group paired with the conflicting agent. Using the ABI concepts, we model the human trustworthiness across both groups using in-game observations (objective) and a questionnaire (subjective). When comparing the results from both group we see that the conflicting agent does not decrease the objective trustworthiness, however looking at the questionnaires we observe that the subjective human benevolence and integrity are negatively affected when paired with the conflicting agent.

Introduction

1 Background

Human-AI collaboration is the study of how humans and artificial systems work together to accomplish a certain task. Artificial agents can aid humans in various domains, e.g. medical surgeries, search & rescue or the military. Wieselquist et al. (1999) argue that in order to build a potent and dynamic teamwork, both parties need to build trust towards each other. This dyadic relation[1] consists of a trustee (party to be trusted) and a trustor (trusting party), which can communicate and cooperate with another.

Trust is defined by Mayer et al. (1995) as the "the willingness of a party to be vulnerable to the outcomes of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party", similarly, Lee & See (2004) posit trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". In addition, trustworthiness is another important concept which is not trivial, it is defined by Hardin (2002) as "the capacity to judge one's interests as dependent on doing what one is trusted to do", which is followed by his distinction that trustworthiness is a set of motivations for acting while trust is a set of expectations that depend on rational assessments of the trustor regarding the trustee's motivations.
In short, trustworthiness is an inherent property that qualifies how much someone is to be trusted and is an objective characteristic while trust is perceived trustworthiness of the trustee from the trustor's subjective perspective.

To define trustworthiness, Mayer et al. (1995) propose the ABI model which comprises Ability - how well a party can complete a certain task and their related skill/competence level; Benevolence - the extent to which the party is seen to be genuinely caring, concerned and willingness to cooperate; Integrity - the extent to which the person is seen as honourable and if they fulfil their promises.

2 Research Question

These definitions of trust are related to human/human relations and were formed before the idea of Human-AI collaboration, therefore it is still unclear how human trustworthiness behaves when confronted with an artificial agent. This paper contributes to new research in this domain by addressing the following question: "How does an artificial agent's behavior affect human trustworthiness?", and specifically in the case of a conflicting agent.

In a human-AI team, the conflicting agent may cause problems and have an overall negative impact on the collaboration compared to a normal agent. We can

*Supervisor: Carolina Jorge (c.jorge@tudelft.nl), responsible professor: Dr. Myrthe Tielman (m.l.tielman@tudelft.nl)
[1]Dyadic relation: two-party relation

think of a conflicting agent as a defective or even malicious party that either intentionally or unintentionally hinders the human in completing the task at hand.

The hypothesis we want to test is: *"human trustworthiness will decrease when paired with a conflicting AI"*. Extant literature from Wieselquist et al. (1999) explains that multiple acts of commitment from one agent will be witnessed by the other party, who in turn experiences greater trust. This promotes greater commitment and more pro-relationship acts, which are witnessed by the first partner, who subsequently experiences greater trust and so on. This trust cycle is displayed in appendix A and shows that if the trustor trusts the trustee, this will lead to the trustor becoming more trustworthy himself. Our hypothesis tests the opposite case, where the trustee (AI) is conflicting and that might cause the trustor (human) to decrease his trust towards the trustee, which in turn negatively affects his/her own trustworthiness.

We begin by looking at how we conduct the experiment; choosing participants, followed by the implementation of the environment along with the conflicting agent. The method section ends with the description on how trustworthiness is measured. Next we present the results of the experiment followed by their analysis and explanation on how the hypothesis was accepted/rejected.

Method

To test our hypothesis, we perform an experiment where we compare the trustworthiness of the human playing with a baseline agent to the trustworthiness of a human playing with the conflicting agent. The following sections discuss the experiment design.

3 Participants

To remove bias that occurs when playing the game a second time, we let each participant only play once. The control group will only play with the normal baseline agent, while the experiment group will play with the conflicting variant. Both groups will have a population of 20 people from various ages and computer expertise to make the study as general as possible and keep consistency across both control and experimental groups. The population statistics regarding gender, age and computer expertise obtained during the experiment are displayed in appendix A.

4 Environment Simulation

To simulate the environment where the human/agent will cooperate, we use a digital task simulation of a USAR (Urban Search And Rescue) scenario[2] using Python and the MATRX[3] package. MATRX is a module enabling rapid prototyping of human/agent team environments. Figure 1 shows a preview of the environment. Here the goal is to fetch the different sick people (red/yellow) and put them in their corresponding slots. It is possible for the human to send messages to the robot and tell it what rooms to search and which people were found. We will use the highest level of interdependence, where the robot needs the help of the human to identify the gender of babies and to carry critically injured adults people. Finally, to be as authentic as possible to a real-life situation, we needed the human to have a sense of urgency and risk, so we implemented a timer of 10 minutes that forces the human to be fast and incite him/her to collaborate with the robot.

Figure 1: Partial Screenshot of the MATRX Simulation

[2]World template provided by R.Verhagen: `https://github.com/rsverhagen94/USAR-HAT`
[3]https://matrx-software.com

5 Conflicting Agent

We then modify the base agent and give him the following conflicting traits; dropping people in the wrong locations, lying about finding victims and wrongly searching rooms. For the first trait the agent will correctly pick up the person but then randomly drop it outside the drop-off zone followed by sending a "malfunction" message to the human informing them that the robot accidentally dropped a victim. This agent also lies by telling the human that he found a victim when it is in fact a healthy person. The last trait causes the faulty robot to wrongly search rooms by not informing the human that it found a sick person, thus making the human believe that there are no victims in that room.

6 Measuring Trustworthiness

To observe the human's trustworthiness, we use the ABI model and measure each concept; ability, benevolence & integrity. First, we look at how the human behaves during the game using objective metrics, followed by analyzing what the human thinks of his own subjective trustworthiness using a questionnaire.

6.1 Objective measures

During the gameplay we record the actions, messages and moves of the human, which we use in metrics that model the human trustworthiness. These objective metrics are divided among the ABI constructs:

- Ability: how well the human manages to complete the game; we record the amount of moves and time it takes to reach the goal. We also measure how many victims the human managed to rescue until the time runs out.
- Benevolence: how well the human wants to cooperate/communicate; we count the amount of messages sent to the robot, how many times the human helps the robot and if the human responds to the questions/suggestions of the rescue bot.
- Integrity: how well the human keeps his promises and tells the truth; we measure how many times the human follows through with his actions and if the messages to the robot are truthful.

These metrics will be normalized and result in a trustworthiness score from $[0; 1]$ for each ABI concept. The exact metrics used in the implementation are show in appendix B.

6.2 Subjective measures

It is also important to measure the subjective trustworthiness, that is, what the human thinks of his own trustworthiness. We can accomplish this using a questionnaire, as seen in Mayer & Davis (1999); Adams et al. (2008). We will perform similar experiments where we let volunteers play the USAR game followed by a questionnaire where they will input their trustworthiness for the different ABI factors using a 7-point likert scale. The questionnaire can be seen in appendix B.

To make sure that the questions are appropriate and related, the Cronbach's alpha is a calculated to measure consistency of the scale and how closely related the set of questions are as a group.

Results

7 Shapiro-Wilk Test

When performing analysis of the results it is important to know if the data is normally distributed or not. Therefore we use the Shapiro-Wilk Test, which mathematically tests for normality. The normality checks inside the control group data in both objective and subjective (questionnaire) metrics[4] are shown below:

	Normal	p-value	$\bar{x}$	σ
Ability	No	0.011	0.747	0.154
Benevolence	Yes	0.501	0.610	0.175
Integrity	Yes	0.051	0.661	0.211
Trustworthiness	Yes	0.714	0.673	0.146

Table 1: Shapiro-Wilk Test Results for Objective Metrics

	Normal	p-value	$\bar{x}$	σ
Ability	Yes	0.478	0.725	0.142
Benevolence	Yes	0.066	0.681	0.262
Integrity	No	0.008	0.820	0.184
Trustworthiness	Yes	0.401	0.742	0.166

Table 2: Shapiro-Wilk Test Results for Questionnaire

We also generated quantile-quantile plots and other distribution graphs to manually confirm the normality of the data. Appendix A presents example graphs for subjective trustworthiness (normally distributed) and for the objective ability (not normally distributed). Moreover, the Cronbach Alpha's for the control and experimental questionnaire are shown in table 3.

	α	Internal Consistency
Control	0.704	Acceptable
Experimental	0.908	Excellent

Table 3: Cronbach's Alpha Results and Interpretation

[4]NB: here `Trustworthiness` is treated as the mean of the ABI constructs.

8 Group Comparisons

Now we plot the ABI scores of the control group next to the score obtained by the experimental group. This will give us a better overview on how the score changes across both groups. The following bar graphs contain the results of the ABI scores for the objective metrics and questionnaire seen in Appendix B.

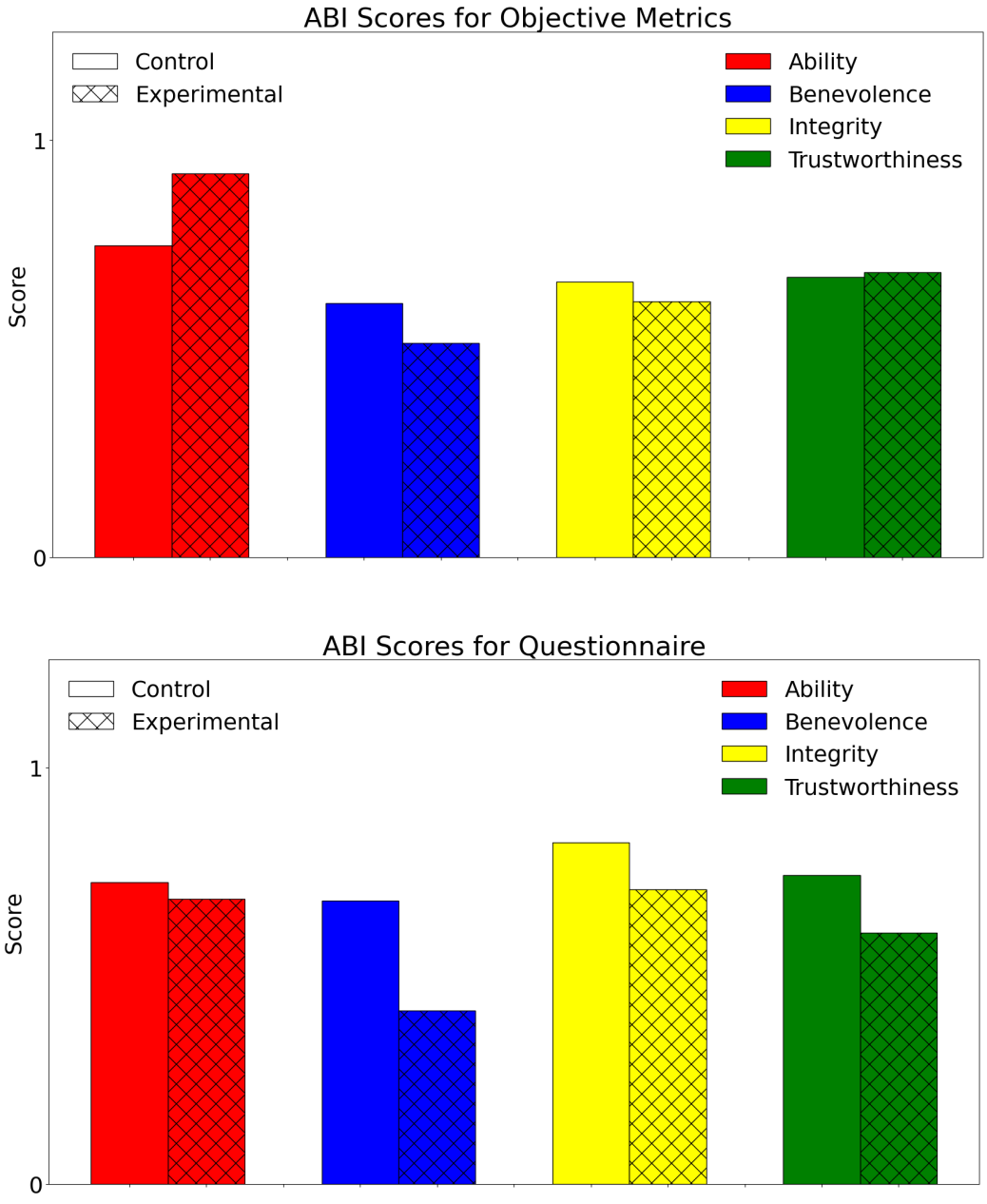

Figure 2: ABI Scores with 95% Confidence Interval

Analysis

9 Statistical Inference

To scientifically solve our hypothesis, we perform statistical hypothesis and significance testing. As test-statistic we either employ the Independent Samples T-Test or the Mann-Whitney U Test depending on the normality of the data. These tests compare the means of our two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. During the test-statistics we set the $\alpha = 0.05$ which is the scientific standard level of significance that dictates the critical value. Using statistical Python libraries we can perform these tests combined with the results of the Shapiro-Wilk Test. The analysis of the statistic testing (with 38 degrees of freedom) lies below:

	Significant	p	Test-Statistic
Ability	No	1.0	26.0
Benevolence	No	0.056	1.63
Integrity	No	0.217	0.792
Trustworthiness	No	0.6	-0.254

Table 4: Statistical Testing for Objective Metrics

	Significant	p	Test-Statistic
Ability	No	0.156	1.026
Benevolence	Yes	0.001	3.433
Integrity	Yes	0.017	278.5
Trustworthiness	Yes	0.003	2.962

Table 5: Statistical Testing for Questionnaire

From the amount of non-significant results obtained for the objective values we can confidently say that the hypothesis does not hold for the objective data. However, for the questionnaire we see that the mean comparisons are all significant apart from the ability, indeed, looking back at the figure 2 we can see that the means of the questionnaire scores are lower in the experimental group for benevolence, integrity and the total trustworthiness. So we can say with a 95% confidence that the alternative hypothesis holds for subjective measures, that is, the trustworthiness of the experimental group is lower than the trustworthiness measured in the baseline group. However, we have to reject the hypothesis concerning the objective measures.

Conclusion

The hypothesis we wanted to test was: *"human trustworthiness will decrease when paired with a conflicting AI"*. After analysing the results we still cannot entirely prove nor disprove the hypothesis. Indeed, we first saw that the human trustworthiness does not decrease when observing objective metrics, that is, teaming the human with a conflicting robot does not negatively affect the objective ABI model during the simulation. However, when asking the human what he thinks of his own trustworthiness we observe a decrease when the human was paired with a conflicting robot, especially for benevolence and integrity. Finally, we can conclude that pairing a human with the conflicting agent does not affect the human trustworthiness inside the game itself, but rather affects the human's perception of his own self benevolence and integrity, which essentially means that the human operator prefers not to cooperate or communicate and sometimes wants to lie to the AI, however the human's low self trustworthiness does not objectively affect the search & rescue task.

10 Responsible Research

10.1 Reproducibility

As stated in the experiment design, this search & rescue experiment was simulated using MATRX and Python software. In order to facilitate reproducibility of this research the code used is publically made available on Github.

10.2 Ethical Implications

To conduct our experiment we followed the TUDelft Human Research Ethics (HREC) checklist which enables us to perform the study as safely as possible for the participants. We made sure that participants were not confronted with any physical or emotional discomfort during the game, moreover, we ensured that the collected data is anonymous. Each volunteer had to sign a consent form seen in Appendix C, which explains the purpose of the research and usage of their data.

11 Limitations and Future Work

The effectiveness of this study was mostly hindered by the fact that we had too few participants. Indeed, due to time constraints we managed to only get 20 participants for each group, which is the smallest amount of population size that will actually provide sensible results when conducing tests on sample means. Ideally, having as many participants as possible would give us a more accurate answer on the problem, since having large sample sizes would help with data normality and might also lead to some test-statistics becoming more significant. One solution would be to scale the experiment online, making it possible to recruit many people.

References

Adams, B. D., Waldherr, S., & Sartori, J. (2008). *Trust in Teams Scale*, (pp. 13–16).

Hardin, R. (2002). *Trust and Trustworthiness*, (p. 28).

Lee, J. D., & See, K. A. (2004). *Trust in Automation: Designing for Appropriate Reliance*.

Mayer, R. C., & Davis, J. H. (1999). *The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment*, (p. 136).

Mayer, R. C., Davis, J. H., & Schoormane, F. D. (1995). *An Integrative Model of Organizational Trust*, (pp. 712, 717).

Wieselquist, J., Rusbult, C. E., Foster, C. A., & Agnew, C. R. (1999). *Commitment, Pro-Relationship Behavior, and Trust in Close Relationships*, (pp. 942, 945).

A Figures

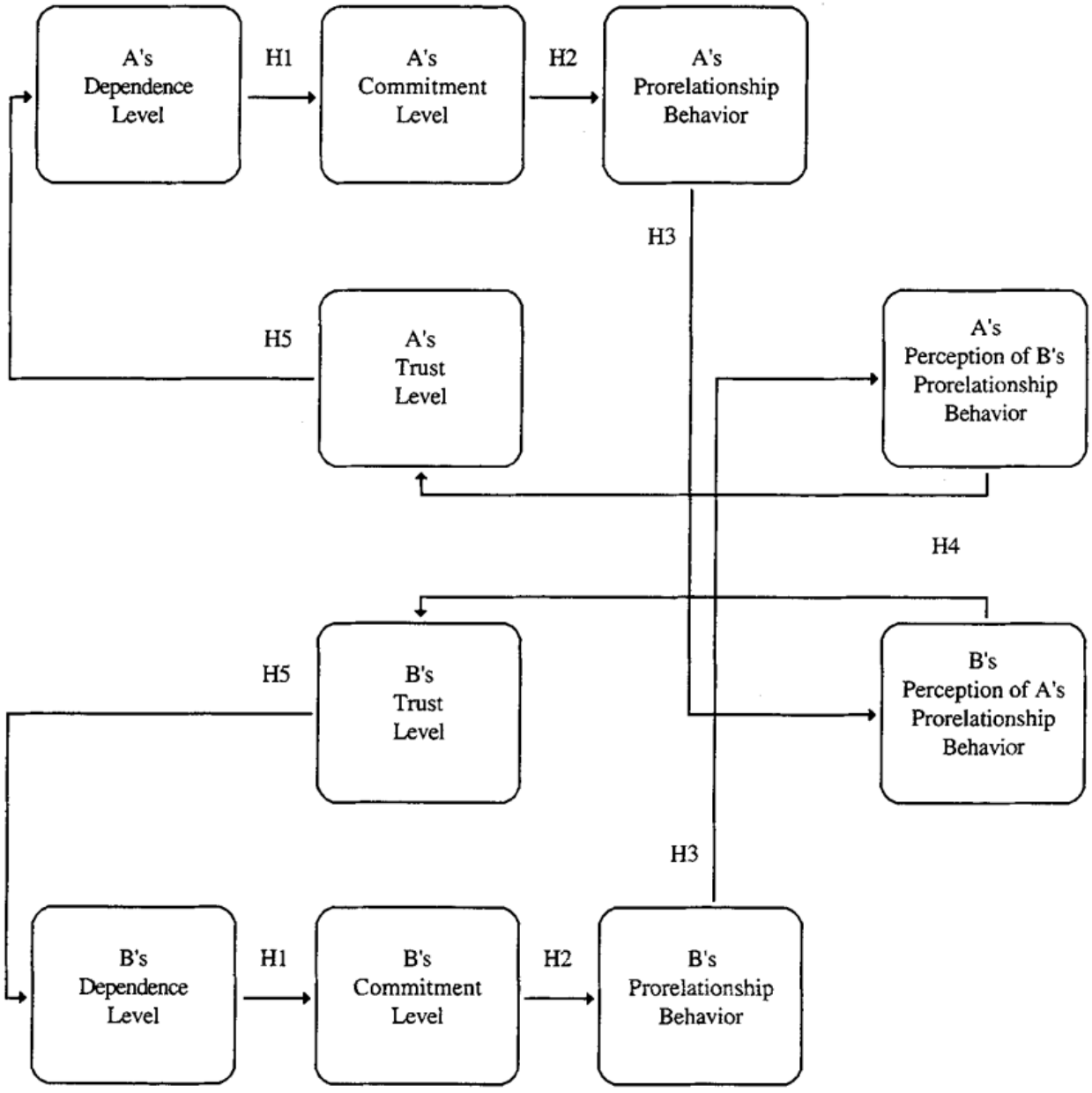

Figure 3: A mutual cyclical growth model of the associations among commitment, pro-relationship behavior, and trust (HI through H5 refer to hypothesized associations among model variables). Wieselquist et al. (1999)

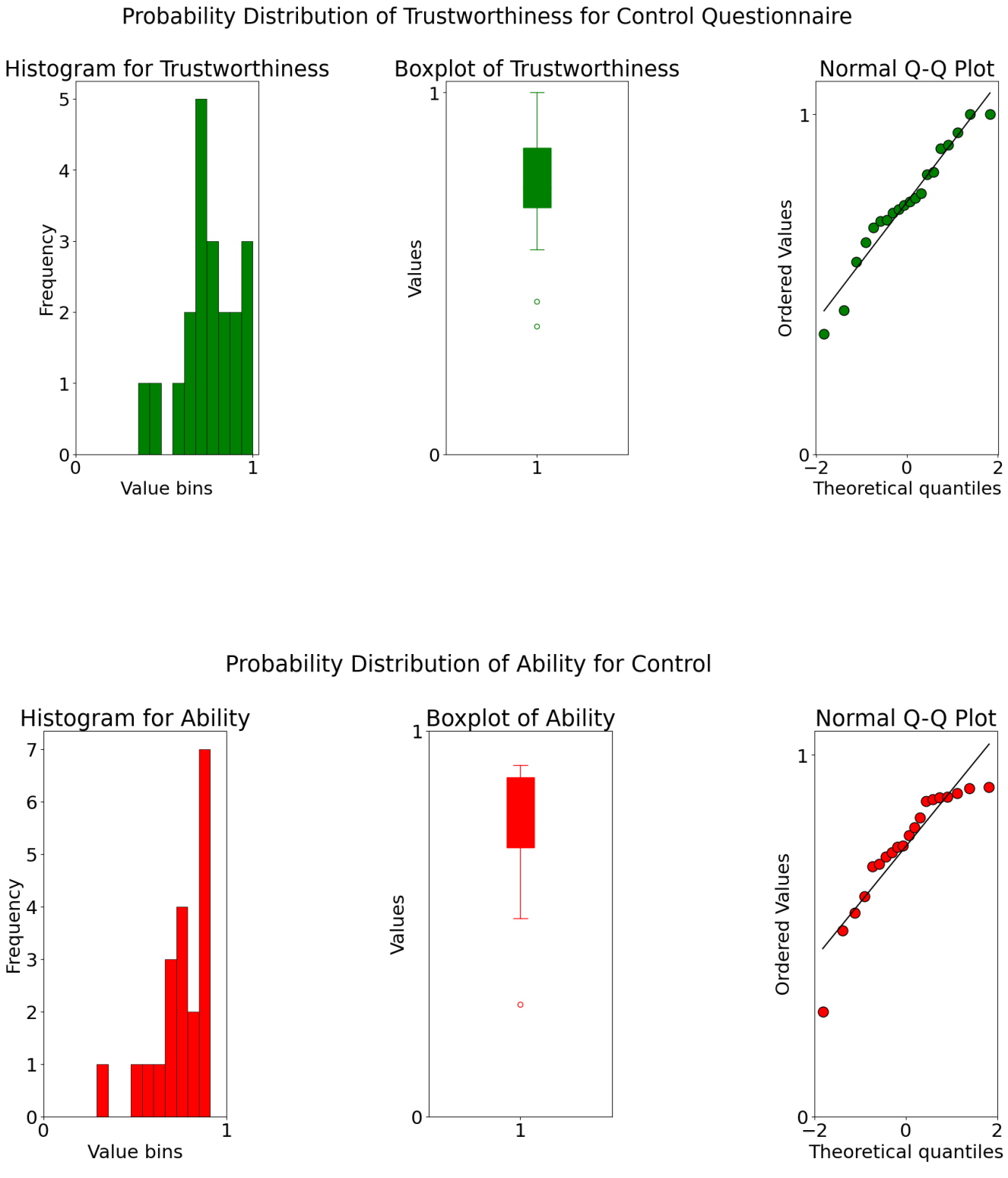

Figure 4: Probability Distribution Graphs

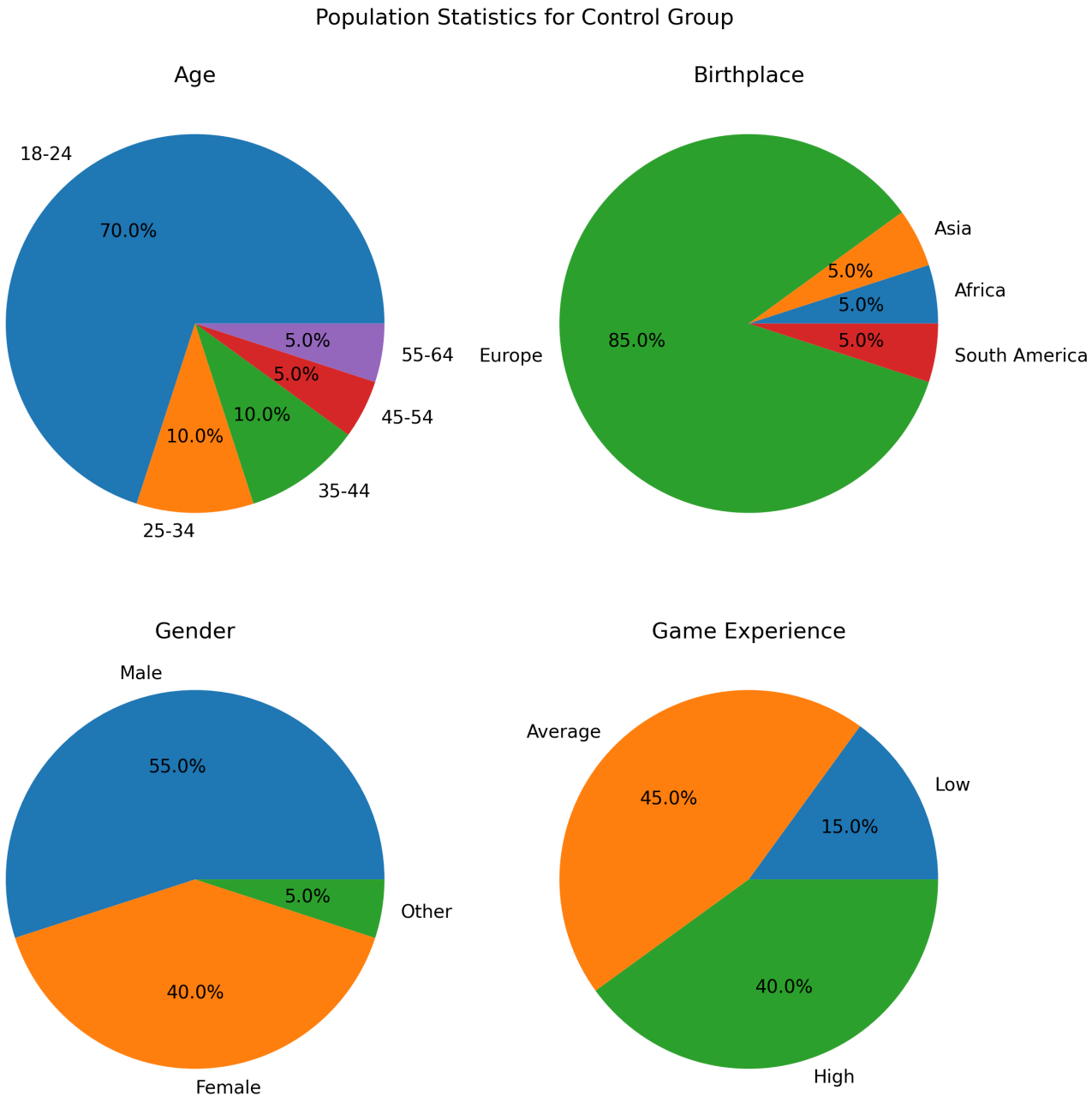

Figure 5: Population Statistics for Control Group

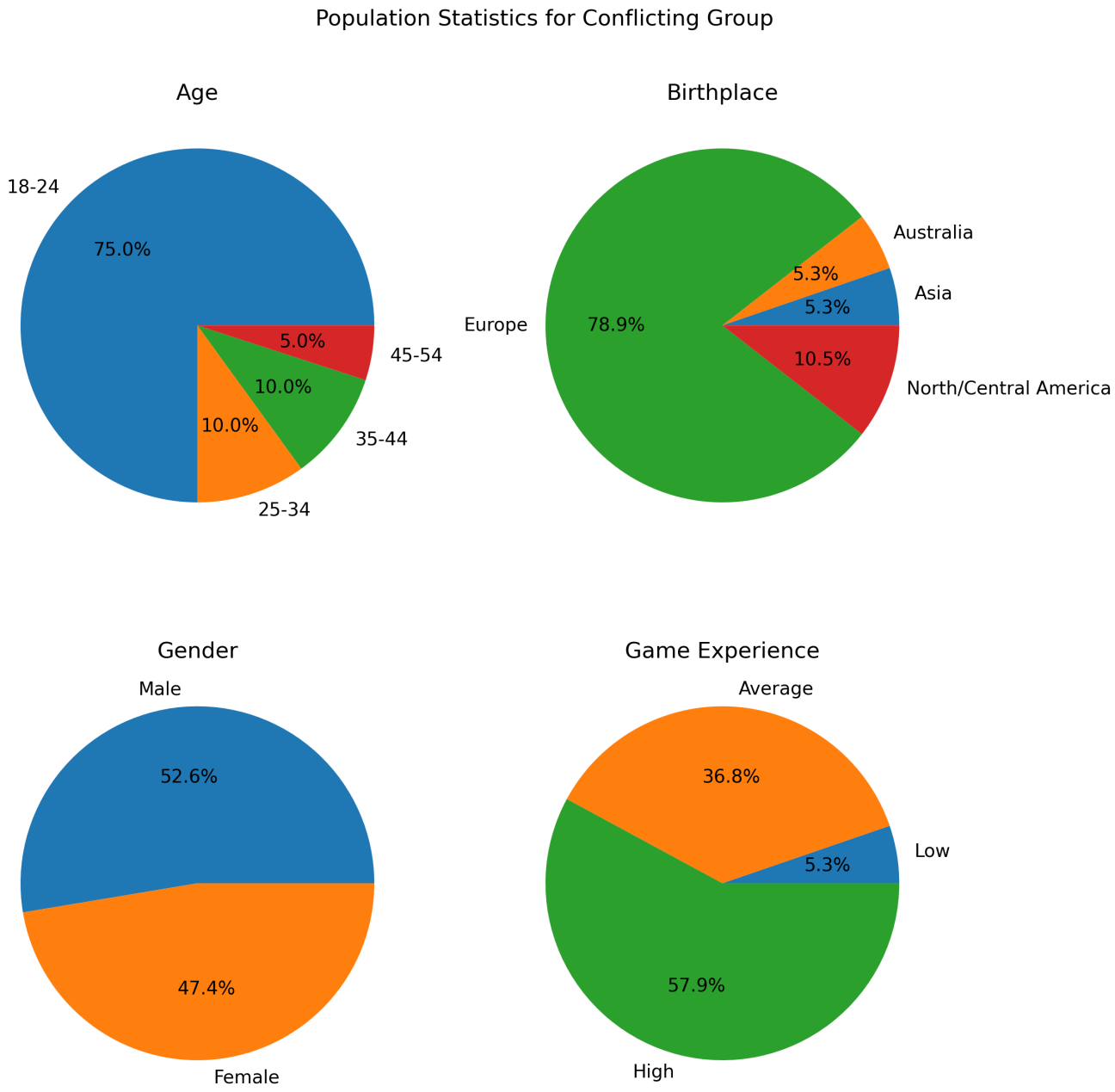

Figure 6: Population Statistics for Experimental Group

B Tables

Amount of ticks (normalized with entire data set)
Amount of moves (normalized with entire data set)
Ratio: number of victims saved / total number of victims
Ratio: number of times victim found / total number of victims
Ratio: number of times victim picked up / total number of victims
Ratio: number of times room visited by human / total number of rooms
Ratio: number times communicated victim found / number of times human sees a new victim for the first time
Ratio: number of communicated gender / number of times agent asks about gender
Ratio: number communicated picked up persons / total number of picked up persons
Ratio: number of communicated Yes (suggested pickup) / total number of pickup suggestions by agent
Ratio: number of communicated room search / number of visted room entries
Ratio: number of times human picks up a victim after the agent advices it / number of times agent advices a pick up
Average number of ticks it takes to respond to a question of the agent (normalized)
Ratio: number communicated relevant person found which was correct / total number of communicated person found
Ratio: number communicated picked up persons / followed through
Ratio: number of communicated Yes/No (suggested pickup) / followed through
Ratio: number of communicated room search / followed through
Ratio: number of communicated gender which are correct / total number of communicated genders

Ability
Benevolence
Integrity

Table 6: Objective metrics used in human trustworthiness model

As a teammate, I was capable at my jobs
As a teammate, I knew what I was doing
My teammate could have faith in my abilities
As a teammate, I was qualified to do my job
As a teammate, I communicated well
I have had teammate Rescuebot's best interests in mind.
It was important to me to communicate my actions to RescueBot
The needs and desires of Rescuebot are important to me
I have looked out for Rescuebot in case it needs assistance
I have been open to RescueBot's suggestions
As a teammate, I kept my promises
I told the truth to my teammates
My teammate could depend on me to be fair
I was an honourable teammate
As a teammate, I honoured my word

Never	Very Rarely	Rarely	Sometimes	Frequently	Very Frequently	Always

Table 7: Questionnaire and scale used to model subjective human trustworthiness

C Consent Form

Delft University of Technology
HUMAN RESEARCH ETHICS
INFORMED CONSENT

Hello! Thank you for participant in the experiment. This experiment will be about a search and rescue game that you will play with RescueBot. The goal of the game is to collaborate with RescueBot and together look for and save the victims. It is a part of our bachelor thesis of Computer Science And Engineering at TU Delft. We investigate how different behaviours or artificial agents influence human's ability, benevolence and integrity, which together gives an indication of trustworthiness.

Who can participate
The participants can be individuals around the globe who are 18 years of age or older.

Risks
During the study we will ask you about some personal information, such as age group, gender, the continent you grew up in and your language proficiency. This data will be used to describe our sample, so that we avoid biases when drawing conclusions from this study. All data will be anonymised and will not be linked in any way to the personal (identifiable) information.

Withdrawal and Exclusion
Participation is entirely voluntary, and you may withdraw during the experiment without consequences. After finishing today's study, however, you can no longer request to remove your data (including the research data) as it is not possible to identify your data since it is anonymised.

Data Storage
Participant's data will be temporarily stored on the researcher's PC. All the data related to the experiment will be deleted by September 1, 2022, after which only the aggregated data in the final research papers will remain. All participant's data will be anonymised. If you have any questions or concerns about your data, do not hesitate to contact [your name (your email).

PLEASE TICK THE APPROPRIATE BOXES	**Yes**	**No**
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated [*DD/MM/YYYY*], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	□	□
3. I understand that taking part in the study involves: playing the tutorial, then playing the actual video game and finally completing a questionnaire. I will be the one who completes the questionnaire. I am aware that the questionnaire consists of several Likert Scale questions.	□	□
4. I understand that I will not be compensated for my participation.	□	□
5. I understand that the study will end in less than forty minutes.	□	□
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
6. I understand that taking part in the study involves the risk that the researcher or the supervisors may know me and I may feel pressured to act as they would prefer or expect me to. I understand that these will be mitigated by not linking in any way the personal data which is being collected to the answers and/or the performance that I will have throughout the experiment.	□	□
7. I understand that taking part in the study does not involve collecting specific personally identifiable information (PII) nor any associated personally identifiable research data (PIRD) with the potential risk of my identity being revealed. However, personal data such as gender, age, or the place where I spent most of my childhood will be collected. I understand that this risk is mitigated by the following measures: not storing the personal data in the open repository and destroying it from the researcher's PC (where it has been stored previously) by first of September 2022 and by combining the collected data with the one corresponding to other similar researches.	□	□
8. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach: anonymous data collection, no PII will be collected nor stored and the data will be stored only until first of September on the researcher's PC and then in the institutional open repository of the Technical University Delft.	□	□
9. I understand that personal information collected about me that can identify me, such as my gender, my age and where I have spent the most time while growing up, will be not be shared in the open repository and it will not be linked to my answers.	□	□
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
10. I understand that after the research study the de-identified information I provide will be used for the Bachelor thesis of the researcher.	□	□
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
11. I give permission for the de-identified personal data that I provide to be archived in the open repository of the Technical University Delft so it can be used for future research and learning.	□	□
12. I understand that access to this repository is available in online databases on the Internet that can be accessed freely and instantly.	□	□

Signatures

_____	_____	_____
Name of participant [printed]	Signature	Date

I, as legal representative, have witnessed the accurate reading of the consent form with the potential participant and the individual has had the opportunity to ask questions. I confirm that the individual has given consent freely.

_____	_____	_____
Name of witness	Signature	Date

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____	_____	_____
Researcher name	Signature	Date

Study contact details for further information: [*Name, phone number, email address*]