RQ: How does an artificial agent offering help affect human trustworthiness?

Hypothesis: Being offered help will allow for a better collaboration which will lead to a better task performance and therefore ultimately higher human trustworthiness.

Literature

Human Trustworthiness - is the property of the human agent in behaving in a way that you stays true to what it has promised or what its expected actions are. [Hardin, 2002]

ABI Trust Model

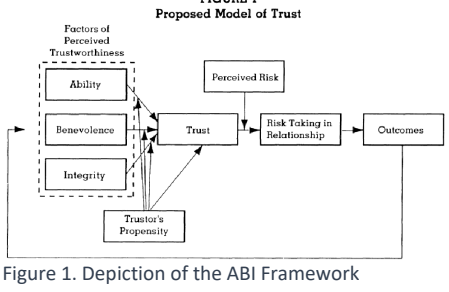


Figure 1. Depiction of the ABI Framework

1) Ability - is that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain.

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.

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]

1) Hardin, R. (2002). *Trust and trustworthiness.* Russell Sage Foundation.
2) 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.

Experiment Environment

Urban Search and Rescue MATRX

- **Team consists of 1 human and 1 AI agent**
- **The task: Search around for injured people/animals in the map and bring them to rescue box.**
- **High Interdependency setup to incentivize collaboration.**
- **Experimental Group and Control Group to test for completing task with helper agent and without helper agent.**

Agent Design

"Ask a question?"

- What are the commands of the game?
- What are the severity colors?
- Who are you able/unable to carry ?

"Need Assistance"

- Found X at Y, could you please come pick them up?

"Ask for progress status"

- How much time do we have left?
- Who have we found so far and where are they located?

Data Processing

Objective Metrics – Agent Measurements

- **Total ticks**
- **Total number of moves**
- **Injured patients saved**
- **Number of messages sent**
- **Number of times human lies/tells truth**
- **How many times the user abides to the advice**
- **Amount of times human identifies objects correctly**
- **etc...**

- **Ability score**
- **Benevolence score**
- **Integrity score**

→ **Human Trustworthiness**

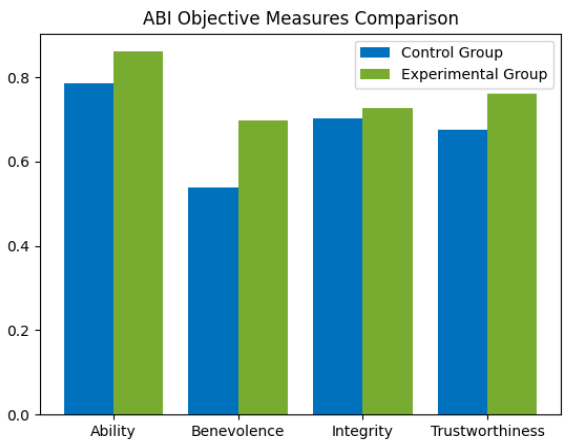
Normalized Values in the range [0, 1]

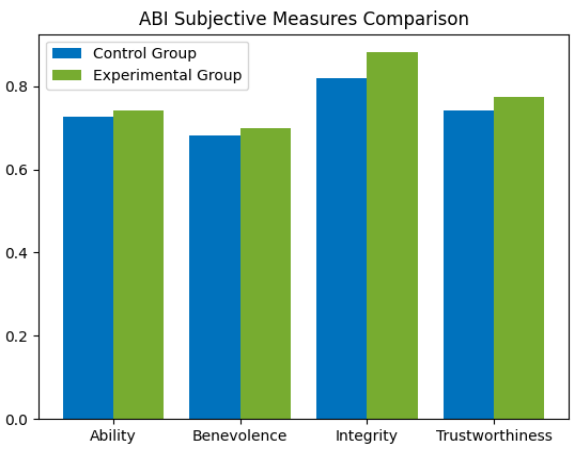
Questionnaire – Human Evaluation

- **Ability, Integrity and Benevolence that the human projects on themselves after the completion of the experiment**
- **Questions divided into the ABI categories with 7 possible answers**
- **Proccessed using the Likert Rating Scale**

Results

- **Ability and Benevolence showed statistically significant increase in objective measures.**
  - **Ability Mann-Whitney Test -> $U(N_{control} = 20, N_{expr.} = 20) = 109.0, p < 0.014$**
  - **Benevolence T-test -> $t(38) = -2.694, p < 0.010$**
- **Trustworthiness showed a significant increase as well.**
  - **T-test -> $t(38) = -2.093, p < 0.043$**
- **Subjective measures showed no significance difference between the control and experimental group**





Discussion & Limitations

- **Human trustworthiness's increase when measured subjectively is caused by a significant increase in ability and benevolence.**
- **Change in objective integrity and subjective integrity is not signficant. Could be because the participant's high trust in their AI partner is not affected by the help offering action alone.**
- **Subjective measures were not significant, arguably due to self-bias not being affected by the dependent variable.**
- **Distribution of values were generally negatively skewed, which was arguably caused by the simplicity of the game.**
- **Age differences of the control group and experimental group was different which might have affected results.**

Future work

- **Different trust model like the SWIFT model**
- **Improvement in terms of sample size and sample diversity.**
- **Testing for help effect on human trustworthiness in more dynamic team settings.**
- **Training the agent to determine action weights to measure trustworthiness**

Conclusion

- **Trustworthiness when measured through the AI agent's perspective (objective) increased when the human agent was offered help, however their perceived trustworthiness was not.**
- **The finding help 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**

Alto Delia
a.delia@student.tudelft.nl
Responsible Lecturer:
Dr. Myrthe Tielman
Supervisor:
Carolina Jorge

Research Project – CSE
June 2022

TU Delft