More Similar Values, More Trust? - the Effect of Value Similarity on Trust in Human-Agent Interaction

Siddharth Mehrotra,[1] Catholijn M. Jonker[1,2] and Myrthe L. Tielman[1]

[1]{s.mehrotra, c.m.jonker, m.l.tielman}@tudelft.nl and [2]{c.m.jonker}@liacs.leidenuniv.nl

Problem

- As AI systems gain complexity and become more pervasive, it becomes crucial for them to elicit **appropriate trust** from humans.
- As a first step towards eliciting appropriate trust, we need to understand **what factors influence trust in AI agents ?**
- In this work, we we examine the effect of (dis)-similarity of human & agent's values on a human's trust in that agent.

Hypothesis

We focus on exploring how users'trust is affected by interaction with different agents with varying value similarity. More specifically, we have the following hypothesis:

Value similarity between the user and the agent **positively** affects the trust a user has in that agent.

Interaction Platform

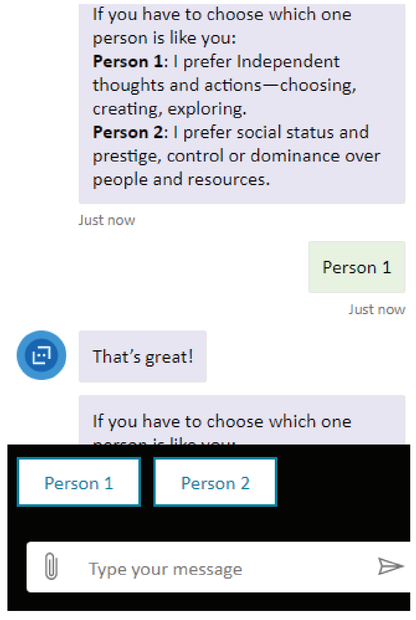

Figure 1: Human-AI agent interaction chat-bot testbed with HTML front-end.

Methodology

- We design five different agents with varying value profiles based on participant's responses;
- The agents team up with participants for a risk-taking task scenario for which they have to interact and decide on the appropriate action to take;
- 89 Participants evaluate the agents based on how much they trust each agent and their perceived Value Similarity (VS).

Results

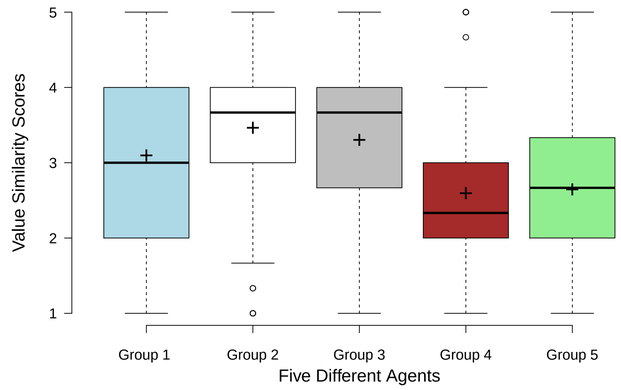

Figure 2: Mean subjective VS scores for all Value Similarity Questionnaire [2] given by participants for the five agents. The horizontal line indicates the median and the plus sign the mean value for VS scores.

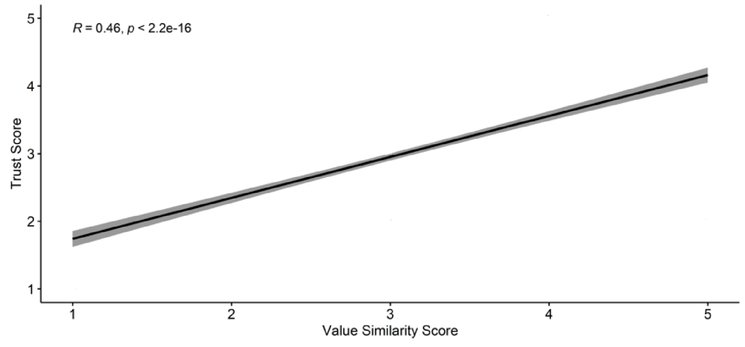

Figure 3: Correlation between value similarity score and trust score. The grey region represents the confidence band. Linear regression show that both the p-values for the intercept and the predictor variable were highly significant indicating a significant association between the variables.

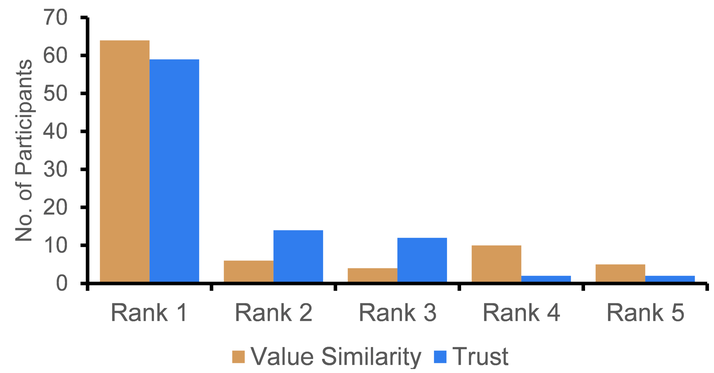

Figure 4: Number of participants who choose an agent to take inside the building based upon their rank of value similarity and trust.

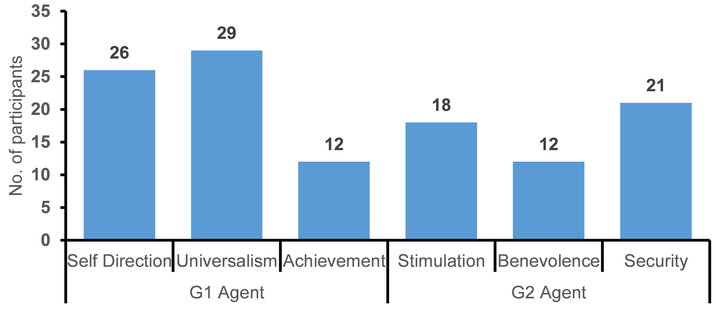

Figure 5: Top three most common values in the value pro-file of the G1 agent (values ranked 1 and 2 of participant)and the G2 agent (values ranked 3 and 4 of the participant).The numbers on the top of the histogram represent how many times those valued occur.

Conclusion

1. **Our results show that agents rated as having more similar values also scored higher on trust, indicating a positive effect between the two.**
2. **With this result, we add to the existing understanding of human-agent trust by providing insight into the role of value-similarity.**

References

1. Shalom H Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values.Online readings in Psychology and Culture2, 1 (2012), 2307–0919
2. Michael Siegrist, George Cvetkovich, and Claudia Roth. 2000. Salient ValueSimilarity, Social Trust, and Risk/Benefit Perception.Risk analysis, 3 (2000),353–362