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

Siddharth Mehrotra,† Catholijn M. Jonker1,2 and Myrthe L. Tielman†

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 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

![Image of bar chart showing value similarity scores for different groups of agents]

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.

![Image of scatter plot showing correlation between value similarity score and trust score]

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.

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
