Exploring the Effect of Value Similarity on Trust in Human-AI Interaction[*]

Summary of a existing published research

SIDDHARTH MEHROTRA, Delft University of Technology, The Netherlands

CATHOLIJN M. JONKER, Delft University of Technology & LIACS, Leiden University, The Netherlands

MYRTHE L. TIELMAN, Delft University of Technology, The Netherlands

Trust is essential to cooperation, which produces positive-sum outcomes that strengthen society and benefit its individual members. Nowadays, many systems are being developed that can make a difference in people's lives, from health apps to robots. But to reach their potential, people need to have *appropriate* levels of trust in these systems. To achieve this, it is first important to understand which factors influence trust in AI. In this paper, we identified a research gap that exists regarding the role of personal values in trust in AI. Therefore, in this paper we explore how value similarity between an agent and a human is correlated to how much that human trusts the agent. To explore this, we conducted a user study where we designed five different AI conversational agents with varying value similarity profiles to that of the participants. In a within-subjects, scenario-based experiment, AI agents gave suggestions on what to do when entering the building to save a hostage. Users evaluated the AI agents based on how much they trust each agent and their perceived value similarity. We analyzed the agent's subjective value similarity, trust, and qualitative data from open-ended questions. Our results show that AI agents rated as having more similar values also scored higher on trust, indicating a positive effect between the two.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Artificial intelligence**; **Intelligent agents**.

Additional Key Words and Phrases: Trust; Values; Value Similarity; Artificial Agents; Intelligent Agents; Human-AI Interaction; Human-Computer Interaction

1 INTRODUCTION

There are growing interests in AI technologies creating new opportunities to improve people's lives around the world, from healthcare to education to business. However, it also raises questions about the best way to build trust [18]. These questions are far from solved and are active areas of research. As for understanding human trust in AI systems, people often fail to calibrate their trust in the AI and end up in the status called over-trust or under-trust. Over-trust is overestimating the reliability of the AI system leading to misuse of the AI outside of designed capability. Therefore, it becomes crucial for AI agents 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. Despite the growing attention in research on trust in AI agents, a lot is still unknown about people's perceptions of trust in AI agents [6]. In this paper, we see trust as multi-dimensional as suggested by Roff and Danks [12]. On the one hand, trust corresponds to reliability and/or predictability and on the other hand trust depends upon people's values, preferences, expectations, constraints, and

[*] This work was originally published as: Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More Similar Values, More Trust? - the Efect of Value Similarity on Trust in Human-Agent Interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* https://doi.org/10.1145/3461702.3462576

Manuscript submitted to ACM

beliefs. Various studies have examined how trust is attributed according to the first dimension [2, 13], but fewer have investigated the second dimension, where the focus is on people's shared values [4].

Siegrist et al. state [16] "*people base their trust judgments on whether they feel that the agency shares similar goals, values, and opinions.*" For example, if you value *cost-efficiency* over *aesthetics* when it comes to buildings, you would probably trust an architect more if they have shown that *cost-efficiency* is also important to them. Taking this approach forward in AI agent research, we examine the effect of (dis)-similarity (of human & agent's values) on a human's trust in that agent. We designed five different agents with varying value profiles so that for any human, some of these are more similar and some less similar to the value profile of that human. The AI agents team up with participants for a risk-taking hostage rescue scenario for which they have to interact and decide on the appropriate action to take. Participants evaluated the AI agents based on how much they trust each agent and their perceived Value Similarity (VS). Our results provide insights into the role of Value (dis)-Similarity on human-AI agent trust relationship. Researchers and designers who wish to develop AI agents in the context of risk-taking scenarios can benefit from our research.

2 BACKGROUND

Increasingly, researchers are trying to incorporate human values in AI systems, especially for those systems which are in some way involved in (helping humans with) decision making. For example, Winikoff argue value-based reasoning to be an essential prerequisite for having appropriate human trust in autonomous systems [23]. This thought echoes with prior work by Banavar [8], van Riemsdijk et al. [20] and Mercuur et al. [10]. More recently, Cohen et al. acknowledge [3] that "*Human users will be disappointed if the AI system makes no effort to represent or reason about inherent social values that users would like to see reflected.*"

One of the earlier attempts to look at the effect of similarity of values on trust was made within social science research by Siegrist et al. [16]. They showed similar values, and trust depends upon each other in human-human interaction. Their findings resonated with Sitkin and Roth [17] who report that interpersonal trust is based on shared values. On these lines, Vaske et al. showed that as salient value similarity increases, social trust in the agency increases [21]. Their findings showed how understanding the value similarity between Colorado residents and United States department of agriculture, resulted in social trust and attitudes towards wildland fire management.

Much of existing work on implementing human values in AI system focuses on user-agent value alignment [15], plan selection [4], and studying agent's value driven behaviour [5]. Building on these works, our research is looking for a deeper understanding regarding the effect of value similarity on trust in a risk taking scenario. We aim to account for the perception of human participants with a user study instead of providing simulation based results.

3 STUDY DESIGN

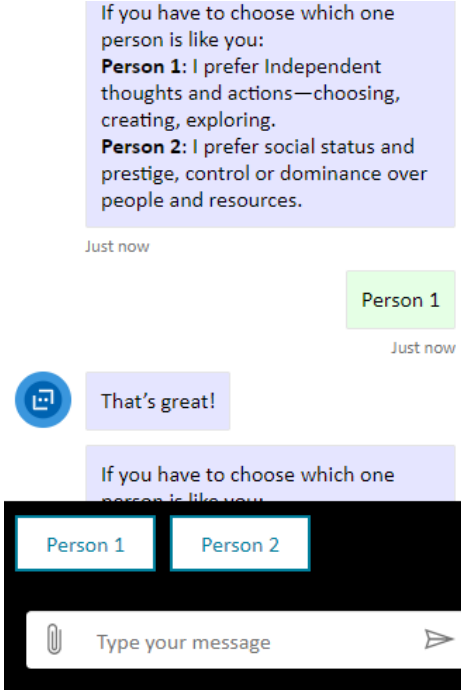
We used the Schwartz Portrait Value Questionnaire (PVQ) [14] to draw each participant's user profiles which consist of ten value dimensions. There are statements about each value dimension in the PVQ. For each '*very much like me*' we assigned a score of 1 and for each '*not at all like me*' a score of 6 to that value. Furthermore, we created an actual value profile for each user based on their rank[1]. We combined the first two values according to rank as group one, the second two values as group two, and so on till group five. We grouped ten values into five groups with two values each. Sometimes, a group can have more than two values because multiple values could receive the same final score. To resolve this conflict, we employed Algorithm 1 (*see Appendix 1*) to get user priority.

[1] We define rank as a position in the hierarchy of importance of the values.

3.1 Task

We designed a “save a hostage game” in which each participant interacts with five different agents that provided tips and suggestions to save the hostage. In our game, AI agents were featured with varying value profiles. For each participant, we created five different agents with descending value similarity profiles from G1 to G5. G1 is the agent who promotes the two top ranked values of the participant, G2 agent which promotes the values ranked 3 and 4, G3 promotes the values ranked 5 and 6, etc. (so the values that each agent promotes can differ for each participant depending on their PVQ outcome).

We recruited 101 participants from the different universities’ mailing list. We provided the following scenario to our participants in which they need to team up with AI agents to rescue a hostage: “*A hostage is being held inside a building in a market place. The objective is to gather intelligence regarding the building. All five different AI agents are equipped with sensors, infrared cameras, and metal detectors. The AI agents can perform the security check of the building and inform you regarding any danger. You need to make a decision for the action to be taken based on the AI agent’s advice before you enter the building*." Each AI agent provides a suggestion to the user based on their prior common knowledge and values that are of utmost importance. A piece of prior common knowledge for all the agents was *“I have searched the overall place and have found traces of the gun powder. I recommend that you take protective gear & armor shield with you”*.



If you have to choose which one person is like you:
Person 1: I prefer Independent thoughts and actions—choosing, creating, exploring.
Person 2: I prefer social status and prestige, control or dominance over people and resources.

Just now

Person 1

Just now

That’s great!

If you have to choose which one person is like you:

Fig. 1. Human-AI agent interaction chatbot testbed

We designed our suggestions based on the values following the notion of situation vignettes in the work by Strackand and Gennerich [19]. The values were expressed through the suggestions the agent gave. For example, an agent provides the following suggestion based on prior knowledge plus their values from group one - security and self-direction: *“I* *have searched the overall place and have found traces of gun powder. I recommend that you take protective gear & an armor shield with you. For any action you take, do follow social orders & protocols. You should hand over the kidnapper to the police to abide by the national security laws. However, it's up to you what equipment you want to take inside the building & how you wish to deal with the situation."*

3.2 Procedure

The test bed consists of a chatbot application that can be accessed from a web browser (*see figure 1*). Participants were asked to complete the PVQ to get their value profiles. After filling the PVQ, the system checked for any conflicts in value groups using Algorithm 1. Following this, the scenario was introduced to the participant. All five agents interacted with the participant one by one. The order of appearance of the agents was randomly assigned in such a way that the order was different for each participant. After each agent gave the suggestion, the participant was asked to fill questions from the Value Similarity Questionnaire (VSQ) [16] and questions from the Human-Computer Trust Scale (HCTS) [7] *(see Appendix 2 for details).*

4 RESULTS

As part of our analysis, we first ran a Shapiro-Wilks test for normality. Since the distribution was not normal, we used non-parametric tests for our analysis. We tried to manipulate value similarity in this study. To check whether our most *'similar'* to least *'similar'* agent were actually perceived as most and least similar, we also measured subjective value similarity. From figure 2, we see that the *'G2'* agent scored higher than the '*G1'* agent, χ_r^2 *= 11.725, p < .05*. This was in contradiction with the manipulation that we performed. In an ideal case, we expect the VSQ ratings to follow the order as G1 agent receives the highest VS score and G5 the least. This showcases that our manipulation did not work as expected.

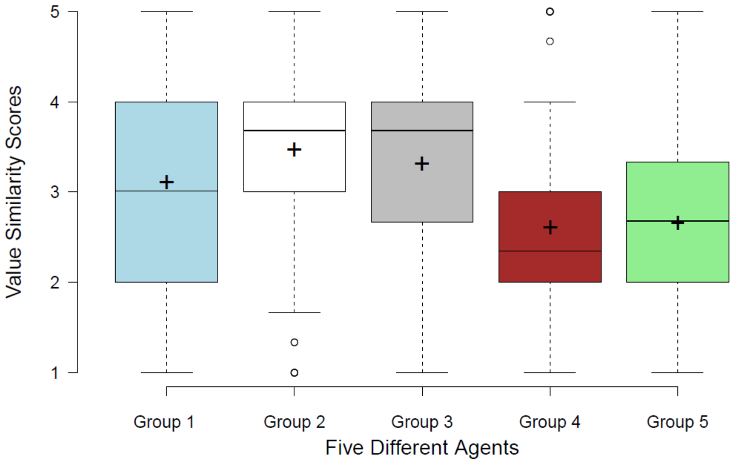


Fig. 2. Mean subjective VS scores for all VSQ given by participants for the five agents. The horizontal line indicates the median and the plus sign the mean value for VS scores.

Considering this, we now only focus upon value similarity as a whole rather than distribution /categorization of five agents. Therefore, in the rest of the paper we disregard our categorization of the agents.

4.1 Correlation between Value Similarity and overall Trust

We analyzed responses for the VSQ and HCTS to see to what extent subjective value similarity has an affect on trust. A Kendall rank correlation test revealed that VS and trust are significantly moderately correlated in accordance with Ratner [11] with a correlation coefficient of *0.46* and $p < 0.05$. We also applied a simple linear regression model to predict a quantitative outcome of trust based on a single predictor variable *i.e.* value similarity. The model showed that both the p-values for the intercept and the predictor variable were highly significant indicating a significant association between the variables. Our goodness-of-fit measures showcase $\sigma = 0.984$ meaning that the observed trust values deviate from the true regression line by approximately 0.984 units on average on a scale from one to five and r^2 was 0.308. This confirms how closely VS and trust are related.

4.2 Benevolence and Willingness as attributes of overall trust

We examined the results of HCTQ as attributes of trust namely benevolence, willingness and general trust on value similarity. We already reported the results of the general trust in previous sections. Now, we focus ourselves to Benevolence and Willingness. A Kendall tau correlation was performed to determine the relationship between benevolence, willingness and value similarity. There was a medium, positive correlation between benevolence and value similarity, which was statistically significant ($r = .47$, n = 436, $p = .0002$). Similarly, for willingness, correlation was found to be positive ($r = .37$, n = 436, $p = .0002$).

5 DISCUSSION

In this paper, we showed that there exists an overall significant effect of VS on trust. Even though our failed manipulations did not interfere with our paper's primary goal, we were intrigued to find out that our manipulations of VS were not successful. To this end we need to know what factors influence trust, and we need to be able to manipulate these factors in the designs of agents. Therefore, it is relevant to examine closer why our manipulations failed and provide some suggestions for how value similarity might be manipulated successfully in the future.

Regarding our specific agent design, a successful manipulation would have led to the observation that the *'G1'* agent is rated highest for the perceived VS and the *'G5'* agent the least. However, we observed that instead both the *'G2'* and the *'G3'* agent were rated as having more similar values than the *'G1'* agent. To understand why this happened, we examined the actual value profiles of the participants more closely. Consider the case when VS scores of the *'G2'* agent were higher than those of the *'G1'* agent. Observing the participants' specific value profiles for who this occurred could provide us with potential reasons why manipulations were not successful. For these participants the values of Self-Direction, Universalism & Achievement were most prominent for the agent *'G1'* and Stimulation, Benevolence & Security for the agent *'G2'*, for those participants where *'G2'* scored higher than *'G1'* in value similarity.

Given that people felt most similar to agents which promoted stimulation, benevolence and security (as opposed to the values of self-direction, universalism and achievement which scored higher in their value profile), we speculate that the choice of scenario might have played a role. The major values for agent *'G1'* - Self Direction and Universalism, were those which participants already possessed but were not so relevant in this context of saving a hostage. On the other hand, for agent *'G2'* - Security and Stimulation were vital because they relate to safety and motivating the participant to save the hostage. It makes intuitive sense that contextual values are of utmost importance especially in those scenarios where there is a risk associated with trusting someone and not all the values are equally salient. However, the value profile survey is general, and not context-dependent. Therefore, we speculate that when designing value profiles for artificial agents, one should not just take into account general value profiles, but also note which contextual values are most important as also echoed by Liscio et al. [9].

Another potential reason for our failed manipulation could be that a discrepancy existed regarding values of the agent in how they were perceived by some of the participants and how they were intended. By perceived values we mean that the value laden explanations that agents provided were sometimes interpreted as promoting different values than for which they were written. As explained in section 'Scenario and agent explanation', it took three iterations for each explanation to be finalized, which indicates how quickly disagreements about underlying values of explanations can occur. We speculate that this discrepancy is a possible reason for our failed manipulation and resound with Wang et al. [22] that designing agent explanations that can be consistently interpreted by humans is still an open research area. Secondly, consistency in value preferences from humans is debated, and people could just show inconsistencies as mentioned by Boyd et al. [1].

6 CONCLUSION

Our study shows that value similarity between an agent and a human is positively related to how much that human trusts the agent. Based on this finding, we would encourage designers of explanation and feedback-giving agents to create agents that outline human values. An agent with similar values to the human will be trusted more which can be very important in any risk-taking scenario. Although a system without value-based reasoning may be easier to develop, the benefits of including VS are worth it, especially in trust-critical situations.

REFERENCES

[1] Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9.

[2] Alain Chavaillaz, David Wastell, and Jürgen Sauer. 2016. System reliability, performance and trust in adaptable automation. *Applied Ergonomics* 52 (2016), 333–342.

[3] Robin Cohen, Mike Schaekermann, Sihao Liu, and Michael Cormier. 2019. Trusted Ai and the Contribution of Trust Modeling in Multiagent Systems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1644–1648.

[4] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents.. In *IJCAI*. 178–184.

[5] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. 2013. No Smoking Here: Values, Norms and Culture in Multi-agent Systems. *Artificial intelligence and law* 21, 1 (2013), 79–107.

[6] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (2020).

[7] Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, Development and Evaluation of a Human-Computer Trust Scale. *Behaviour &; Information Technology* 38, 10 (2019), 1004–1015.

[8] Banavar Guru. 2016. What It Will Take for Us to Trust AI. *Harvard Business Review* (Nov. 2016). https://hbr.org/2016/11/what-it-will-take-for-us-to-trust-ai

[9] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 799–808.

[10] Rijk Mercuur, Virginia Dignum, and Catholijn Jonker. 2019. The Value of Values and Norms in Social Simulation. *Journal of Artificial Societies and Social Simulation* 22, 1 (2019).

[11] Bruce Ratner. 2009. The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing* 17, 2 (2009), 139–142.

[12] Heather M Roff and David Danks. 2018. "Trust but Verify": The Difficulty of Trusting Autonomous Weapons Systems. *Journal of Military Ethics* 17, 1 (2018), 2–20.

[13] Mark Ryan. 2020. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics* (2020), 1–19.

[14] Shalom H Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online readings in Psychology and Culture* 2, 1 (2012), 2307–0919.

[15] Daniel Shapiro and Ross Shachter. 2002. User-agent Value Alignment. In *Proc. of The 18th Nat. Conf. on Artif. Intell. AAAI*.

[16] Michael Siegrist, George Cvetkovich, and Claudia Roth. 2000. Salient Value Similarity, Social Trust, and Risk/Benefit Perception. *Risk analysis* 20, 3 (2000), 353–362.

[17] Sim B Sitkin and Nancy L Roth. 1993. Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization science* 4, 3 (1993), 367–392.

[18] Julia Stoyanovich, Jay J Van Bavel, and Tessa V West. 2020. The imperative of interpretable machines. *Nature Machine Intelligence* 2, 4 (2020), 197–199.

[19] Micha Strack and Carsten Gennerich. 2011. Personal and Situational Values Predict Ethical Reasoning. *Europe’s Journal of Psychology* 7, 3 (2011), 419–442.

[20] M Birna Van Riemsdijk, Catholijn M Jonker, and Victor Lesser. 2015. Creating Socially Adaptive Electronic Partners: Interaction, Reasoning and Ethical Challenges. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*. 1201–1206.

[21] Jerry J Vaske, James D Absher, and Alan D Bright. 2007. Salient Value Similarity, Social Trust and Attitudes Toward Wildland Fire Management Strategies. *Human Ecology Review* (2007), 223–232.

[22] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[23] Michael Winikoff. 2017. Towards Trusting Autonomous Systems. In *International Workshop on Engineering Multi-Agent Systems*. Springer, 3–20.

A APPENDIX - ALGORITHM 1

Algorithm 1: Resolve conflicts in value profiles

```
Input: n = number of values in each group, i & j = 0 and, number of groups (g) = 0;
Result: Corrected value profile without conflicts
while n>2 & g<5 do
    combinations = fact(n) / (fact(2) * fact(n - 2)) ;
    for (i = (combination-1); i >= 1; i–) do
        for (j = 0; j <= (i-1); j++) do
            if (List[j] == List[j+1]) then
                user input to select a value;
                if (Selected Value == List[j]) then
                    List [j] -= 0.05;
                else
                    List [j+1] -= 0.05;
                end
            end
        end
        g++;
    end
end
```

B APPENDIX 2 - VSQ AND HCTS

Value Similarity Questionnaire - [16]

Scale: Totally Agree - Agree - Neutral - Disagree - Totally Disagree

- Do you think the Agent X acts as you would do in this scenario?
- Do you think Agent X thinks like you?
- Do you think Agent X shares your values?

Human Computer Trust Scale - [7]:

Scale: Totally Agree - Agree - Neutral - Disagree - Totally Disagree

- It is risky to interact with Agent A in this scenario. - *Willingness*
- Agent X will do its best to help you if you need help. - *Benevolence*
- If you take Agent X help, you would be able to depend on it. - *Trust*
- You can rely on Agent X in this scenario. - *Trust*
- You can trust the information presented to you by Agent X in this scenario.- *Trust*