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

Mehrotra, S.; Jonker, C.M.; Tielman, M.L.

DOI
10.1145/3461702.3462576

Publication date
2021

Document Version
Final published version

Published in
AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society

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

Siddharth Mehrotra
Delft University of Technology
Delft, The Netherlands
s.mehrotra@tudelft.nl

Catholijn M. Jonker
Delft University of Technology &
LIACS, Leiden University
Delft, The Netherlands
c.m.jonker@tudelft.nl

Myrthe L. Tielman
Delft University of Technology
Delft, The Netherlands
m.l.tielman@tudelft.nl

ABSTRACT

As AI systems are increasingly involved in decision making, it also becomes important that they elicit appropriate levels of trust from their users. To achieve this, it is first important to understand which factors influence trust in AI. We identify that a research gap exists regarding the role of personal values in trust in AI. Therefore, this paper studies how human and agent Value Similarity (VS) influences a human’s trust in that agent. To explore this, 89 participants teamed up with five different agents, which were designed with varying levels of value similarity to that of the participants. In a within-subjects, scenario-based experiment, agents gave suggestions on what to do when entering the building to save a hostage. We analyzed the agent’s scores on subjective value similarity, trust and qualitative data from open-ended questions. Our results show that agents rated as having more similar values also scored higher on trust, indicating a positive effect between the two. With this result, we add to the existing understanding of human-agent trust by providing insight into the role of value-similarity.

CCS CONCEPTS

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

KEYWORDS

Trust; Values; Value Similarity; Artificial Agents; Intelligent Agents; Human-AI Interaction; Human-Computer Interaction

ACM Reference Format:
Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More Similar Values, More Trust? - the Effect 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.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3461702.3462576

1 INTRODUCTION

In the Indian epic Mahabharata, Arjun and Bhima are important characters. They go through common struggles and trust in each other’s abilities. They challenged Jarasandha and Chitrasena’s (*two kings*) armies and fought for Kampilya together (*a capital kingdom*). What made them have so much trust in each other? According to Rajagopalachari [20], the most compelling reason was that they shared similar values. In this paper, we explore how we can take inspiration from this story when trying to understand trust in AI.

AIES ’21, May 19–21, 2021, Virtual Event, USA.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462576

As AI systems gain complexity and become more pervasive, it becomes crucial for them to elicit appropriate trust from humans. We should avoid under-trust, as it would mean not making optimal use of AI. Yet we should also avoid over trust, as relying on AI systems too much could have serious consequences [17]. 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 [11]. Therefore, we wish to know what it is that makes people trust or distrust AI? In this paper, we see trust as multi-dimensional as suggested by Roff and Danks [22]. 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 beliefs. Various studies have examined how trust is attributed according to the first dimension [3, 23], but fewer have investigated the second dimension, where the focus is on people’s shared values [6]. The implication of the latter dimension for the design of agents is on how to design these agents with respect to values as different people prioritize different values, which in turn guides how people behave and judge the behavior of others [10].

We argue that there is a research gap in understanding the role of values on the trust a human has in that agent. Siegrist et al. state [27]:

> “*people base their trust judgments on whether they feel that the agency shares similar goals, thoughts, 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. Regarding trust in AI systems, we resonate with Tolmeijer et al. [30] in observing the potential for overlap and contrast with the psychology, ethics, and pragmatics of trust between humans. Based on this, we hypothesize that the trust of humans in AI agents is positively correlated to the similarity of the values of those agents and humans. 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 design 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 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. Participants evaluate the agents based on how much they trust each agent and their perceived Value Similarity (VS).

In the remainder of this paper, we first review related work on value similarity and give an overview of existing literature on the use of values to promote trust. We then describe the design of the agents we use in the experiments, and the setup of our user study. We discuss our results and conclude with potential applications and limitations of our work.

2 RELATED WORK

Trust within the AI domain has been explored mostly in contexts such as decision making [26], examining/assessing user's trust [16], and improving the system performance [18]. We argue that it is important to also consider the similarity of personal values when researching trust. But can an AI agent have personal values? Increasingly, researchers are trying to incorporate values in AI systems, especially systems which are in some way involved in (helping humans with) decision making.

Winikoff argue value-based reasoning to be an essential prerequisite for having appropriate human trust in autonomous systems [35]. This thought echoes with prior work by Banavar [13], van Riemsdijk et al. [31] and Mercuur et al. [15]. More recently, Cohen et al. acknowledge [5]:

> *"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."*

Most practical work on implementing human values in AI system focuses on plan selection [6], user-agent value alignment [25] and studying agent's value driven behaviour [8]. 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. [27]. They showed similar values, and trust depends upon each other in human-human interaction. Their findings resonated with Sitkin and Roth [28] 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 [32]. 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.

Recently, researchers have been interested in using this concept of value similarity for AI systems as well. Cruciani et al. designed an agent based model showing how similarity in values can be a successful driver for cooperation in the regulation and design of public policies [7]. They analyze their simulation experiment by looking at how and, how much agents cooperate with similar others. The key takeaway message is the introduction of value similarity for investigating what ultimately motivates trust-building processes. However, their work used predetermined memory coefficient for simulation agents to study coordination and was not validated with human participants. Additionally, Chhogyal and colleagues designed a formal trust assessment model [4]. In their work, they developed value-based trust assessment functions and showed how they lead to trust sequences. However, they did not consider value preferences and neither validated the model with human participants. 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 accounting for the perception of human participants instead of providing simulation based results.

3 METHOD

The primary goal of our study is to understand how (perceived) value similarity affects trust. We focused 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.

3.1 Creation of value profiles

We used the Schwartz Portrait Value Questionnaire (PVQ) [24] to draw each participant's user profiles which consist of ten value dimensions. There are statements about each value dimension in the PVQ. Participants were asked to read carefully and respond to how each statement resonate with them as a person on a scale of 1-6, where '1' means '*very much like me*' and '6' implies '*not at all like me*'.

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] (*refer column 'PVQ Score' in table 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 employ Algorithm 1 (*see Appendix 1*) to get user priority. For example, in table 1, there are three values with a score of 0.9 (*refer set C1*); and we needed only two values for each group. Therefore, participants were asked to choose one value over another based on the meaning of two values (*refer Figure 1*) following algorithm 1. In our user-study, we did not come across a conflict case where there were more than four values with the same PVQ score.

3.2 Agents and the scenario

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. The task was inspired by prior work from Wang et al. [34]. In our game, agents were featured with varying value profiles.

3.2.1 Agents and Value Similarity: For each participant, we created five different agents with descending value similarity profiles from G1 to G5 (*see table 1 for example*). 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

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

Table 1: An example of generating value profiles of agents based on human value profile. *Rank* represents order of the values, *PVQ* represents the PVQ scores by participants, *Corrected* represents scores after applying the algorithm 1. Lower scores corresponds to higher ranks. C1 showcases conflict between three values for group two. Group 1 (G1) - Group 5 (G5) are groups for the first two ranks, the second two ranks, and so on... representing five different agents.

Rank	*PVQ Score*		*Corrected*	*Value*	
1	1		1.0	Security	*Group 1*: G1
2	1		1.0	Self Direction	
3	2	C1	1.90	Traditional	*Group 2*: G2
4	2		1.95	Conformity	
5	2		2.0	Universalism	*Group 3*: G3
6	3		3.0	Power	
7	4		4.0	Benevolence	*Group 4*: G4
8	4		4.0	Hedonism	
9	5		5.0	Achievement	*Group 5*: G5
10	6		6.0	Stimulation	

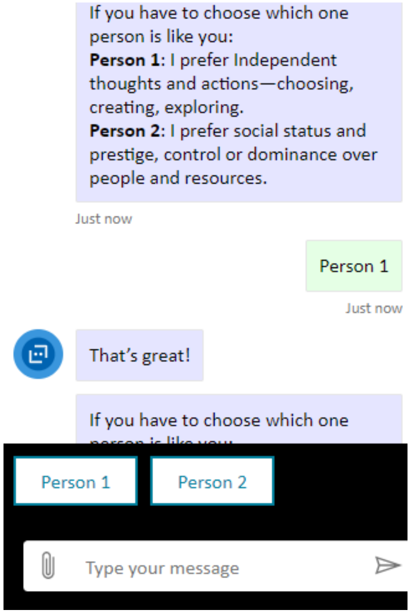
6, etc. (so the values that each agent promotes can differ for each participant depending on their PVQ outcome).

3.2.2 Scenario and agent explanation: 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.*"

We used the agent's names as A, B, C, D, and E mapping to G1-G5 in our user-study. Each 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".*

We designed our suggestions based on the values following the notion of situation vignettes in the work by Strackand and Gennerich [29]. The values were expressed through the suggestions the agent gave. Two researchers from Computer Science background and one from Cognitive Science background brainstormed together and generated sentences that formed suggestions by the agent. Overall, three iterations of each suggestion was performed to reach the final outcome.

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."*


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:

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

3.3 Participants

We estimated our sample size with the G-Power tool from Faul et al. [9]. Our effect size was 0.30 (*medium*) with linear regression as our choice for modelling variables. G-power calculated our required sample size of 81. We recruited 101 participants from the different universities' mailing list. Twelve participants could not pass our attention check, leaving 89 participants aged between 22 and 32 years old (M= 25.6; SD = 0.94). Each participant signed an informed consent form before the user-study. This study was approved by the ethics committee of our institution, ID number 1313.

We asked our participants to provide their cultural backgrounds before starting the user-study. Most of our participants were from the Europe region (34), followed by Asia Pacific (29), Americas (13), Middle East and Africa (9), and Oceania (2). Two participants did not provide their background.

3.4 User study test bed

We implemented an online version of our scenario to study the impact of manipulating value similarity on trust. The test bed consists of a chatbot application that can be accessed from a web browser (*see figure 1*). We used Microsoft Power Apps API [2] to generate suggestions by the agents. These were displayed on the participant's chatbot interface, which sends data back to the test bed server. The user study test bed can be found at (website blinded for the review).

[2] https://powerapps.microsoft.com/en-us/

3.5 Procedure

Each participant first read an information sheet about the study and then fill out the background survey. Next, 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 and asked the participant to choose one over another. 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. Each agent appears with a small greeting and provides their suggestion. After each agent gave the suggestion, the participant was asked to fill questions from the Value Similarity Questionnaire (VSQ) [27] and questions from the Human-Computer Trust Scale (HCTS) [12]. In HCTS, trust is divided into three attributes, namely: general trust, benevolence, and willingness *(see appendix 2 for details)*. The study was designed to be completed in 30 minutes. Participants were given a chance to participate in a raffle worth 5x20 Euro.

4 RESULTS

We analyzed the results of our study, including both the subjective rating responses to the value similarity, the trust questionnaire and, the explanations provided by the participants for selecting an agent. We were primarily interested in the effect of value similarity on trust. Thus, for this paper, we focus on understanding the effect on trust by manipulating the value similarity. We call VSQ responses from participants as subjective value similarity. 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.

4.1 Manipulation check

We tried to manipulate value similarity in this study. However, 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^2_r *= 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. 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.

4.2 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 [21] 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. To check linear model assumptions, we used the 'GVLMA' - Global Validation of Linear Models Assumptions [19] which provides a testing suite for many of the assumptions of general linear models. The four assumptions: normality, heteroscedasticity, linearity and, uncorrelatedness of the model were acceptable by the GVLMA, *see appendix 3*. Linear regression showed that both the p-values for the intercept and the predictor variable were highly significant indicating a significant association between the variables, *refer appendix 4*. 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.

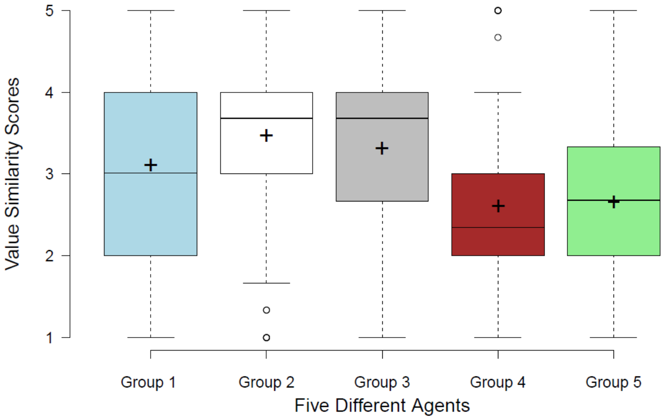


Figure 2: Mean subjective VS scores for all VSQ given by participants for the five agents. The horizontal line indicates the median and the plus sign the mean value for VS scores.

Finally, to seek an answer to the problem: "can we predict trust from VS?", we need to look at the intercept and residuals of the linear regression. On observing the intercept and residuals we have good reason to believe an overall effect of value similarity on trust. This confirms how closely VS and trust are related. Additionally, we wished to check if differences in cultural background of participants affected the effect of VS on trust. However, because our sample size was very diverse there were not enough participants from any distinct cultural background for a statistical comparison between them. Such an effect is potentially important, but future work would need to be done to test for this.

4.3 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).

4.4 Qualitative data analysis

We were interested in understanding which agents participants preferred the most. For this, we asked them to choose an agent to take with them inside the building and were asked to explain their reasons for doing so. We analyzed participants responses for selecting an agent. Our results indicate that the participants pick that agent that shares the most similar values. In figure 3, we can observe that more than 72% of participants chose the agent they ranked highest on value similarity and trust. This gives us another impression that subjective value similarity and trust correlate with each other. Now, we classified participants qualitative explanations into four themes found by thematic analysis [2]. This classification provides insight into the reasons for participant's choices and how it translates to actual behaviour for selecting an agent. The four themes for selecting one agent over others are:

1. Common Values - the selected agent had more values in common with the human than the other agents.
2. Balanced Advice - the selected agent provided more balanced advice to the human participant than the other agents.
3. Developed Trust - the selected agent's advice/suggestion inclined the participant to trust the agent.
4. Participant's Belief - the agent was selected based on its advice/suggestion; this decision was neither related to values nor developed trust.

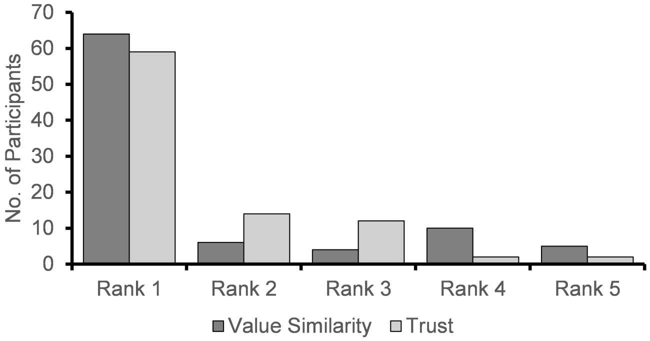


Figure 3: This figure represents the number of participants who choose an agent to take inside the building based upon their rank of value similarity and trust.

Out of a total of 89 participants only 55 provided an explanation for choosing an agent. Three researchers coded the explanations written by the participants. Each researcher performed the coding with three to four iterations before deciding upon final themes. Inter-coder reliability analysis was performed using Cohen's kappa to determine agreement and consistency between all coders. There was a near-perfect agreement among all three coders for three dimensions κ = .900, (95% CI, .643 to .937).

Based on our analysis, we found that 42% of the participants explanations were related to common values between the participant and the agent they chose. This was followed by 23% for balanced advice given by the agent, 16% for developed trust and 16% for belief of the participant. These results shows that in our experiment, VS and balanced advice promoted the intended behaviour of participants to select an agent. For example, P54 said, '*He [Agent A] thinks the same way as me so I think he'd back me in my decisions*' relates to choosing an agent based on the common values. Similarly, P39 said, '*I believe agent B thinks 100% like me and gives me all the trust and responsibility*' relates to developed trust for the agent. We also came across many responses where participants choose the agent because of balanced advice by them. For example, P44 said, '*Agent B shows a balance of risk taking and following protocol to handle a delicate situation*'. Finally, few participants stick to their beliefs for their decision. This can be seen with what P27 reported as '*I believe he [Agent B] would be able to help save the hostages and neutralize the threat with non lethal force if possible and lethal if absolutely necessary*'.

5 DISCUSSION

In this section, we discuss the results of our study, relating them to prior work and making inferences on how the results can be applied to the design of AI agents. Recall that our main goal was to understand the effect of similarity of human & agent's values on a human's trust in that agent. Based on our study results, the hypothesis (that the VS between the user and the agent positively affects the trust a user has in that agent) can be partially accepted. 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. In the following section, we discuss possible reasons for our unsuccessful manipulations.

5.1 Why our manipulations were unsuccessful?

If we wish to eventually promote appropriate trust, we should also be able to influence trust. 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. In this paper we have added to the knowledge on factors that influence trust by showing the relationship with value similarity. However, the manipulation of those factors did not fully succeed in our study. 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. Figure 4 provides an overview of the values used in the explanations of the *'G1'* and the *'G2'* agents, and how often those occurred. This figure shows that 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

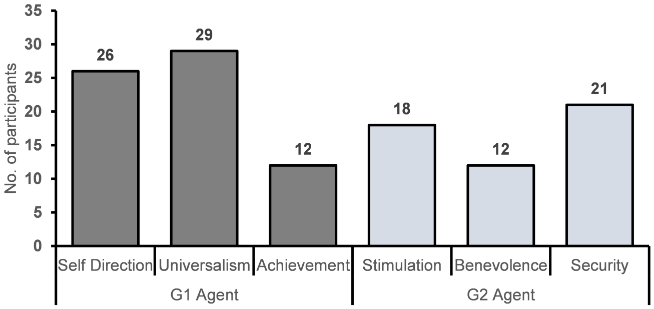


Figure 4: Top three most common values in the value profile of the G1 agent (values ranked 1 and 2 of participant) and the G2 agent (values ranked 3 and 4 of the participant). The numbers on the top of the histogram represent how often those values occurred in this agent (G1 or G2) for our participants.

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. [14].

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. [33] 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].

5.2 Trust in AI systems

In our user-study, agents provided their suggestions based on value-based reasoning using VS. With the use of value-based reasoning, an agent includes the representation for human values and can provide reasoning using human values to make decisions. Winikoff argue that a computational model of relevant human values can be used to provide higher level, human-centered explanations of decisions by AI agents. This means that agents could use value-based reasoning when trying to influence trust [35]. In synopsis, given a bunch of random generated agents, humans would trust those align with their subjective values. The reported correlation can also comes from human's consistent value judgment about the suggestions and scenarios.

Value Similarity is not the only thing that influences trust; many other factors can influence trust as well. Three of the main aspects of trust are benevolence, willingness and competence. Value similarity could be seen as a part of benevolence, or even willingness. However, competence is less related to values [36]. We did not focus upon this factor because we provided all our agents a ground truth *i.e.* prior common knowledge. Instead, we focused upon benevolence and willingness as other two factors of trust affected by VS in accordance with Gulati et al. [12]. Based on our results, both these factors were moderately positively correlated with VS. This implies that if we wish to understand how humans trust systems we need to look beyond trust as being influenced only by the system's reliability. Rather, we also need to consider trust in benevolence and willingness and understand how these are influenced by aspects such as value similarity [5].

5.3 Limitations & Future Work

We investigated the effect of VS on trust with a risk-taking scenario of saving a hostage. We chose this scenario to gain a deeper understanding of how participants trust an agent with most to least VS. However, we believe that further evaluation with more real-life examples would provide additional insights on participant's trust. Additionally, although we cross-examined the participant value profile with their responses to the VS questionnaire, we did not focus on their understanding of value laden explanations. We posit that examining the perception of the values could have provided a more subtle effect of our manipulations. We see this as an opportunity to further extend our work into understanding the beliefs and perceptions of the participants for agents with varying VS. Also, future work could extend the proposed method to multiple scenarios with different context information. Additionally, crowdsourcing could be another way to generate explanations instead of pre-designing by experts or experimenters, especially in translating abstract values to specific descriptions or behaviors. Finally, as explained in the previous section, we would like to study the potential effect of culture on our findings.

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.

SUPPLEMENTARY

The raw data set of this study along with the processed data files are available at https://doi.org/10.4121/14518380.

ACKNOWLEDGMENTS

Special thanks go to Ilir Kola and Sandy Manolios who helped in designing the explanations for the user study, and anonymous reviewers who provided useful suggestions for incorporation in this work.

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] Virginia Braun and Victoria Clarke. 2020. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* (2020), 1–25.

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

[4] Kinzang Chhogyal, Abhaya Nayak, Aditya Ghose, and Hoa K. Dam. 2019. A Value-based Trust Assessment Model for Multi-agent Systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 194–200. https://doi.org/10.24963/ijcai.2019/28

[5] 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.

[6] 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.

[7] Caterina Cruciani, Anna Moretti, and Paolo Pellizzari. 2017. Dynamic Patterns in Similarity-based Cooperation: An Agent-based Investigation. *Journal of Economic Interaction and Coordination* 12, 1 (2017), 121–141.

[8] 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.

[9] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[10] Batya Friedman, Peter H Kahn, and Alan Borning. 2008. Value Sensitive Design and Information Systems. *The handbook of information and computer ethics* (2008), 69–101.

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

[12] 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.

[13] 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

[14] 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.

[15] 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).

[16] Rui Ogawa, Sung Park, and Hiroyuki Umemuro. 2019. How Humans Develop Trust in Communication Robots: A Phased Model Based on Interpersonal Trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 606–607.

[17] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[18] Andisheh Partovi, Ingrid Zukerman, Kai Zhan, Nora Hamacher, and Jakob Hohwy. 2019. Relationship between Device Performance, Trust and User Behaviour in a Care-taking Scenario. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 61–69.

[19] Edsel A Peña and Elizabeth H Slate. 2006. Global Validation of Linear Model Assumptions. *J. Amer. Statist. Assoc.* 101, 473 (2006), 341–354.

[20] Chakravarti Rajagopalachari. 1970. *Mahabharata*. Vol. 1. Diamond Pocket Books (P) Ltd.

[21] 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.

[22] 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.

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

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

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

[26] Maayan Shvo, Jakob Buhmann, and Mubbasir Kapadia. 2019. Towards Modeling the Interplay of Personality, Motivation, Emotion, and Mood in Social Agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2195–2197.

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

[28] 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.

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

[30] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 3–12.

[31] 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.

[32] 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.

[33] 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.

[34] Ning Wang, David V Pynadath, and Susan G Hill. 2016. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 international conference on autonomous agents &; multiagent systems*. 997–1005.

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

[36] Bogdan Wojciszke. 1997. Parallels Between Competence-versus Morality-related Traits and Individualistic Versus Collectivistic Values. *European Journal of Social Psychology* 27, 3 (1997), 245–256.