Differences in Value Alignment Effects Between Human-AI Interaction and Interpersonal Interaction: The Moderating Role of Decision Context Type

Authors: Li Qinggong, Xu Mengqi, Li Qinggong

Date: 2025-04-07T00:00:00+00:00

Abstract

This study investigates the human-AI value alignment effect and its differences from interpersonal value alignment effects. Experiment 1 (N = 145) employed a single-factor (interaction partner: AI vs. human) between-subjects design and found that both human-AI and interpersonal value alignment effects were present, with no significant difference between them. Experiment 2 (N = 116) utilized decision context type as a moderating variable, adopting a 2 (interaction partner: AI vs. human) × 2 (decision context type: factual judgment vs. value judgment) mixed design. The results revealed that in value judgments, the human-AI value alignment effect was weaker than the interpersonal value alignment effect; in factual judgments, no difference was observed between the two.

Full Text

Preamble

Value Alignment Effect in Human-AI Versus Interpersonal Interaction: The Moderating Role of Decision Context

LI Qinggong[1], XU Mengqi[1]
(1. Zhejiang Provincial Laboratory for Mental Health and Crisis Intervention of Children and Adolescents, College of Psychology, Zhejiang Normal University, Jinhua 321004, China)

This research reveals a potential risk: people are less vigilant about the alignment between AI' s values and their own values when interacting with AI compared to human interactions, particularly in value judgment contexts. This serves as a warning that in high-stakes domains sensitive to values, users must

maintain critical thinking and carefully evaluate AI outputs against their own values. AI developers should enhance the transparency and explainability of AI values to help users identify potential biases and avoid the harms of blind reliance.

The study examines the human-AI value alignment effect and its differences from interpersonal value alignment. Experiment 1 (N = 145) employed a single-factor between-subjects design (interaction partner: AI vs. human) and found that both human-AI and interpersonal value alignment effects existed, with no significant difference between them. Experiment 2 (N = 116) introduced decision context type as a moderating variable using a 2 (interaction partner: AI vs. human) × 2 (decision context: factual judgment vs. value judgment) mixed design. Results showed that in value judgments, the human-AI value alignment effect was weaker than the interpersonal value alignment effect, while no difference emerged in factual judgments.

Keywords: value alignment; human-AI interaction; factual judgment; value judgment; trust
Classification Code: B849
Funding: This research was supported by the Ministry of Education Humanities and Social Sciences Research General Project "Dynamic Changes in Human-AI Trust: A Computational Modeling Approach Based on Reinforcement Learning" (24YJA190008) and the Zhejiang Philosophy and Social Science Planning Research Method Innovation Special Project "The Development of Children and Adolescents' Values and Their Relationship with Mental Health" (23SYS09ZD).

Corresponding Author: LI Qinggong, Professor, e-mail: liqinggong@zjnu.edu.cn

---

Abstract

This study investigates the value alignment effect in human-AI interactions and its differences from interpersonal value alignment. Experiment 1 (N = 145) employed a single-factor between-subjects design (interaction partner: AI vs. human). Results revealed that both human-AI and interpersonal value alignment effects existed, with no significant difference observed between them. Experiment 2 (N = 116) incorporated decision context as a moderating variable, using a 2 (interaction partner: AI vs. human) × 2 (decision context: factual judgment vs. value judgment) mixed design. Results showed that in value judgments, the human-AI value alignment effect was weaker than the interpersonal value alignment effect, while no difference was found in factual judgments.

Key words: value alignment, human-AI interaction, factual judgment, value judgment, trust

Value alignment aims to ensure that powerful artificial intelligence (AI) systems align with human values, thereby guaranteeing trustworthy, safe, and controllable technology (Gabriel, 2020; 闫宏秀, 2024). Currently, the AI field primarily

explores technical approaches to achieving value alignment at the design and development level, including plug-in alignment methods such as output correction and in-context learning, as well as fine-tuning methods like supervised fine-tuning and reinforcement learning from human feedback (矣晓沅, 谢幸, 2023). However, objective value alignment ultimately requires subjective cognitive processing by human users to generate willingness to rely on AI, which subsequently influences their behavior. Therefore, to ensure AI technology truly serves humanity' s collective interests, human users must remain sensitive and vigilant to AI value alignment. This paper proposes the human-AI value alignment effect, wherein users demonstrate greater willingness to rely on AI systems that highly align with their own values. This reliance manifests as a tendency to adopt AI' s advice during task execution or to actively use AI to complete tasks (Patton & Wickens, 2024).

In this context, a critical research question emerges: Does the human-AI value alignment effect truly exist? Empirical research on this effect remains in its infancy. To our knowledge, only one published study has examined how value similarity between users and robots (utilitarian vs. deontological orientations) affects human-robot trust (Yokoi & Nakayachi, 2021). In contrast, the value alignment effect has been extensively documented in interpersonal interactions: value-congruent individuals are more likely to attract each other and establish trust, yielding positive effects on relationships and task outcomes (Bakar & McCann, 2014; Edwards & Cable, 2009). Additionally, humans exhibit anthropomorphic tendencies, attributing human characteristics to non-human entities (Haslam, 2023), and thus also anthropomorphize AI. Consequently, we hypothesize that, similar to interpersonal value alignment effects, humans will more readily depend on AI systems that align closely with their own values. Accordingly, we propose Hypothesis 1: A value alignment effect exists in human-AI interactions.

Although humans tend to anthropomorphize AI, AI as an artificial entity differs fundamentally from humans. Therefore, we pose a second research question: Do value alignment effects differ between human-AI and interpersonal interactions? The media-equation hypothesis (MEH) and the unique agent hypothesis (UAH) offer different perspectives for analyzing this difference. MEH posits that when interacting with AI, humans perceive it as a social actor equivalent to humans and apply the same social rules and expectations (Nass & Lee, 2001). UAH emphasizes the uniqueness of human-AI interaction, arguing that people respond to AI differently than to humans (de Visser et al., 2016; Lee & See, 2004). While MEH and UAH are often treated as competing hypotheses in previous research (Alarcon et al., 2023; de Visser et al., 2016), this study draws on Złotowski et al.' s (2018) dual-process theory of anthropomorphism to argue that human-AI and interpersonal interactions are not absolutely "equivalent" or "unique." This theory suggests that anthropomorphic tendencies may result from automatic, intuitive processing (Type 1); however, when people are motivated to make accurate judgments and activate reflective processing (Type 2), they may not readily attribute human characteristics to non-human entities.

Based on this theory, we propose an integrative analytical framework: When Type 1 processing dominates, MEH is more likely to hold, making human-AI interactions more similar to interpersonal interactions; when Type 2 processing is activated, UAH is more likely to hold, making people more likely to recognize fundamental differences between AI and humans.

According to this framework, comparing value alignment effects between human-AI and interpersonal interactions requires attention to the boundary conditions under which MEH and UAH respectively hold, with contextual factors potentially serving as key moderators. Factual judgment and value judgment represent two typical decision contexts. Factual judgment reflects one' s understanding of objective facts, whereas value judgment reflects one' s evaluation of and attitude toward objects (刘清平, 2016). These two contexts may elicit different value alignment effects in human-AI versus interpersonal interactions. In value judgments, because the context involves personal value orientations, it may prompt people to reflect on the essential difference that "AI as an artificial entity lacks genuine values" (thereby activating Type 2 processing). This reflection reduces anthropomorphic tendencies toward AI, resulting in a weaker human-AI value alignment effect compared to interpersonal value alignment. In factual judgments, because the context focuses primarily on objective facts rather than values, it does not stimulate deep reflection on essential differences between AI and humans, leading people to automatically anthropomorphize AI (i.e., Type 1 processing dominates). This makes human-AI and interpersonal value alignment effects more similar. Accordingly, we propose Hypothesis 2: Decision context type moderates the difference between human-AI and interpersonal value alignment effects. In value judgments, the human-AI value alignment effect is weaker than the interpersonal value alignment effect; in factual judgments, this difference diminishes.

In summary, this study examines the human-AI value alignment effect and its differences from interpersonal value alignment in human-AI collaborative decision-making contexts. Given that AI often serves as an advisor while human users retain final decision authority (Steyvers & Kumar, 2024), "advice-seeking willingness" and "advice-adoption willingness" constitute key indicators of human dependence on AI (Jiang et al., 2024; Patton & Wickens, 2024). This study employs these indicators to measure the value alignment effect.

Experiment 1

Participants

We used G*Power 3.1 to calculate the required sample size. For a t-test with an effect size of 0.50, significance level of 0.05, and power of 0.80, at least 128 participants were needed. We recruited 147 participants, excluding 2 invalid responses, resulting in a final sample of 145 valid participants (age range: 18-27 years, M = 21.68, SD = 2.20; 95 females). The AI group comprised 73 participants and the human group 72. Both experiments received approval

from the university ethics committee.

Experimental Design and Procedure

We employed a single-factor between-subjects design with two levels (interaction partner: AI vs. human). Participants in the AI and human groups read statements about social issues from either AI or human agents holding different values (conservative vs. open), then completed a value alignment assessment evaluating the similarity between their own views and those of the interaction partner, and selected the more similar partner. Contradictory responses were deemed invalid. We measured participants' advice-seeking and advice-adoption willingness across three dilemma scenarios in finance, healthcare, and human resources to assess the value alignment effect. Experimental materials and detailed procedures are available in the supplementary materials.

Advice-seeking willingness was measured using two 9-point rating scales assessing participants' willingness to seek advice from each interaction partner. The value alignment effect score for advice-seeking was calculated as the difference between advice-seeking willingness toward the high value-congruent partner and that toward the low value-congruent partner. Advice-adoption willingness was measured using a forced-choice question (which partner' s advice they would prefer to adopt) and a certainty rating on a 4-point scale. Combining these two items yielded the value alignment effect score for advice-adoption: if participants chose the high-congruence partner, the score equaled the certainty rating; if they chose the low-congruence partner, the score equaled the negative of the certainty rating. Higher values indicated stronger effects.

The Cronbach' s alpha consistency coefficient across contexts was 0.501. Due to this relatively low coefficient in Experiment 1, we analyzed data from each of the three contexts separately. Results showed consistent patterns across different decision domains, with detailed analyses available in the supplementary materials.

Results

Advice Seeking Descriptive statistics for the value alignment effect are presented in Table 1 . We first conducted one-sample t-tests comparing each group' s value alignment effect in advice-seeking against the neutral point of 0. Results showed significant effects for both groups: AI group, $t(72) = 6.74$, $p < 0.001$, $BF10 = 3.35 \times 106$, Cohen' s $d = 0.79$; human group, $t(71) = 6.69$, $p < 0.001$, $BF10 = 2.63 \times 106$, Cohen' s $d = 0.79$. Next, an independent samples t-test with interaction partner as the independent variable revealed no significant difference between the two groups' value alignment effects, $t(143) = -1.68$, $p = 0.095$, $BF10 = 0.65$, Cohen' s $d = -0.28$. This indicates that both human-AI and interpersonal value alignment effects exist.

Advice Adoption Using the value alignment effect in advice-adoption as the dependent variable, we first conducted one-sample t-tests comparing each group' s effect against 0. Results showed significant effects for both groups: AI group, t(72) = 7.25, p < 0.001, BF10 = 2.69 × 107, Cohen' s d = 0.85; human group, t(71) = 7.80, p < 0.001, BF10 = 2.36 × 108, Cohen' s d = 0.92. An independent samples t-test with interaction partner as the independent variable showed no significant difference between groups, t(143) = -0.64, p = 0.526, BF10 = 0.22, Cohen' s d = -0.11. These results, consistent with advice-seeking findings, indicate that both human-AI and interpersonal value alignment effects exist.

Experiment 2

Building upon Experiment 1, Experiment 2 implemented three key improvements. First, it distinguished between factual and value judgments to examine the moderating role of decision context type on differences between human-AI and interpersonal value alignment effects. Second, while Experiment 1 measured perceived value alignment, Experiment 2 directly manipulated value alignment by constructing radar charts comparing values. Third, to overcome Experiment 1's limitation to the "conservative-open"value dimension, Experiment 2 designed value materials based on Schwartz' s (2006) ten value dimensions to enhance generalizability.

Participants

Using G*Power 3.1, we calculated the required sample size. For an F-test with an effect size of 0.20, significance level of 0.05, and power of 0.80, at least 52 participants were needed. We recruited 120 participants, excluding 4 invalid responses, resulting in a final sample of 116 valid participants (age range: 18-28 years, M = 22.35, SD = 2.11; 75 females). Both the AI and human groups comprised 58 participants each.

Experimental Design and Procedure

We employed a 2 (interaction partner: AI vs. human) × 2 (decision context type: factual judgment vs. value judgment) mixed design, with interaction partner as a between-subjects variable and decision context type as a within-subjects variable. All participants were randomly assigned to either the AI or human group. Both groups first completed the Chinese version of the Schwartz Values Questionnaire (高志华 et al., 2016). The AI group then viewed comparison results between their own values and two AI agents' values, while the human group viewed comparisons with two human agents.

Value alignment manipulation materials are shown in Figure 1 [Figure 1: see original paper]: the left panel depicts the high value-congruence condition, and the right panel depicts the low value-congruence condition. The green line represents participants' own values, while the red and blue lines represent the high-

and low-congruence agents' values, respectively. After the experiment, participants evaluated value similarity and selected the more similar agent. Contradictory responses or selection of the right panel were deemed invalid.

Consistent with Experiment 1, we selected three domains—finance, healthcare, and human resources—and designed both factual and value judgment scenarios for each domain (six total scenarios). Participants viewed these scenarios to assess the value alignment effect, with measurement items and calculation methods identical to Experiment 1. Cronbach' s alpha consistency coefficients were 0.81 for factual judgments and 0.70 for value judgments across contexts. Experimental materials and detailed procedures are available in the supplementary materials.

Results

Advice Seeking Descriptive statistics for the value alignment effect are presented in Table 2 . We first conducted one-sample t-tests comparing each group' s value alignment effect in advice-seeking against 0. Results showed significant effects for both groups: AI group, $t(57) = 6.97$, $p < 0.001$, BF10 = 3.26×106, Cohen' s d = 0.92; human group, $t(57) = 7.61$, $p < 0.001$, BF10 = 3.36×107, Cohen' s d = 1.00. These results again confirm that both human-AI and interpersonal value alignment effects exist.

We then conducted a 2 (interaction partner: AI vs. human) × 2 (decision context type: factual judgment vs. value judgment) repeated measures ANOVA. Results showed no significant main effect of interaction partner, $F(1, 114) = 2.30$, $p = 0.086$, BF10 = 0.97, $^{2} = 0.03$; no significant main effect of decision context type, $F(1, 114) = 0.48$, $p = 0.489$, BF10 = 0.17, $^{2}p = 0.004$; and a marginally significant interaction between interaction partner and decision context type, $F(1, 114) = 3.81$, $p = 0.053$, BF10 = 1.09, $^{2}p = 0.03$.

Simple effects analysis revealed that in value judgments, the AI group' s value alignment effect was significantly weaker than the human group' s ($p = 0.012$), while in factual judgments, no significant difference emerged between groups ($p = 0.514$). This interaction is illustrated in Figure 2 [Figure 2: see original paper].

Advice Adoption Using the value alignment effect in advice-adoption as the dependent variable, we first conducted one-sample t-tests comparing each group' s effect against 0. Results showed significant effects for both groups: AI group, $t(57) = 12.44$, $p < 0.001$, BF10 = 8.14×1014, Cohen' s d = 1.63; human group, $t(57) = 12.15$, $p < 0.001$, BF10 = 3.14×1014, Cohen' s d = 1.60. These results again confirm that both human-AI and interpersonal value alignment effects exist.

We then conducted a 2 (interaction partner: AI vs. human) × 2 (decision context type: factual judgment vs. value judgment) repeated measures ANOVA. Results showed no significant main effect of interaction partner, $F(1, 114) = 0.49$, $p =$

0.484, BF10 = 0.23, ^{2} = 0.004; no significant main effect of decision context type, F(1, 114) = 0.50, p = 0.481, BF10 = 0.18, ^{2}p = 0.004; and a significant interaction between interaction partner and decision context type, F(1, 114) = 7.89, p = 0.006, BF10 = 7.28, ^{2}p = 0.07. Simple effects analysis revealed that in value judgments, the AI group' s value alignment effect was significantly weaker than the human group' s (p = 0.015), while in factual judgments, no significant difference emerged between groups (p = 0.273). This interaction is illustrated in Figure 3 [Figure 3: see original paper].

General Discussion

This study validated the value alignment effect and its boundary conditions: people show greater dependence on agents (whether AI or human) with higher value congruence, supporting Hypothesis 1. More importantly, in value judgments, the human-AI value alignment effect was significantly weaker than the interpersonal value alignment effect, whereas no such difference emerged in factual judgments, supporting Hypothesis 2.

The theoretical contributions are twofold. First, while previous research often treats MEH and UAH as competing hypotheses (Alarcon et al., 2023; de Visser et al., 2016), this paper proposes an integrative analytical framework based on dual-process theory of anthropomorphism: MEH is more likely to hold when Type 1 anthropomorphic processing dominates, whereas UAH is more likely to hold when Type 2 processing is activated. Using the value alignment effect as an entry point, this study found that the boundary conditions for MEH and UAH depend on collaborative decision-making context type, thereby providing empirical support for this framework. Second, this research extends human-AI similarity effect studies from a value alignment perspective. Previous research has primarily focused on surface-level features such as appearance and language style (Lin et al., 2020; Qiu et al., 2020). However, the fundamental distinction between humans and AI lies in values—the core manifestation of humanity (Leyens et al., 2001). Particularly as AI technology becomes increasingly powerful and accompanied by "black box" effects, examining the role of values as a deep attribute in human-AI interaction is crucial. This study reveals the value alignment effect in human-AI interaction, transcending previous limitations of focusing on surface AI features and highlighting the importance of value alignment in human-AI relationships.

Based on these findings, we propose the following practical recommendations. First, in high-stakes domains sensitive to values, users must consistently evaluate AI outputs against their own values, consciously identify potential biases, and avoid blind reliance on AI. Second, in applications involving value judgments, system design should enhance AI value transparency to help users understand AI values. For example, in medical ethical decision-making, AI systems should explicitly display the ethical principles and value trade-off standards underlying their recommendations.

Despite these contributions, this study has limitations. First, it employed imagined scenarios rather than real AI systems. Future research could observe how participants process and respond to value alignment information in actual human-AI interactions to enhance external validity. Second, the study relied on self-report data; future research could supplement this with direct behavioral measures to verify the robustness of conclusions. Finally, participants were not practitioners in relevant domains and lacked professional expertise. Additionally, Experiment 1 showed low consistency coefficients across decision domain scenarios. Future research should further examine the consistency of these conclusions across different populations and decision contexts.

Supplementary Materials and Data Sharing

Supplementary materials for this study are available at https://www.scidb.cn/en/s/U7Fv6j; data supporting the findings are available at https://www.scidb.cn/en/s/VZBV7b.

References

高志华, 杨绍清, Juergen Margraf, XiaoChi Zhang, 路平. (2016). 施瓦茨价值观问卷 (PVQ-21) 中文版在大学生中的修订. 中国健康心理学杂志, 24(11), 1684-1688.

刘清平. (2016). 怎样从事实推出价值?——是与应当之谜新解. 伦理学研究, (1), 13-21.

闫宏秀. (2024). 基于信任视角的价值对齐探究. 浙江社会科学, (6), 39-48. 矣晓沅, 谢幸. (2023). 大模型道德价值对齐问题剖析. 计算机研究与发展, 60(9), 1926-1945.

Alarcon, G. M., Capiola, A., Hamdan, I. A., Lee, M. A., & Jessup, S. A. (2023). Differential biases in human-human versus human-robot interactions. *Applied Ergonomics*, 106, 103858.

Bakar, H.A., & McCann, R.M. (2014). Matters of demographic similarity and dissimilarity in supervisor-subordinate relationships and workplace attitudes. *International Journal of Intercultural Relations*, 41, 1-16.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331-349.

Edwards, J. R., & Cable, D. M. (2009). The value of value congruence. *The Journal of Applied Psychology*, 94(3), 654-677.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), Haslam, M. (2023). Anthropomorphism as a contributor to the success of human (Homo sapiens) tool use. *Journal of Comparative Psychology*, 137(3), 200-208.

Jiang, P., Niu, W., Wang, Q., Yuan, R., & Chen, K. (2024). Understanding users' acceptance of artificial intelligence applications: A literature review. *Behavioral Sciences*, 14(8), 671.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46, 50–80.

Leyens, J.-P., Rodríguez-Pérez, A., Rodríguez-Torres, R., Gaunt, R., Paladino, M.-P., Vaes, J., & Demoulin, S. (2001). Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. *European Journal of Social Psychology*, 31(4), Lin, H., Chi, O. H., & Gursoy, D. (2020). Antecedents of customers' acceptance of artificially intelligent robotic device use in hospitality services. *Journal of Hospitality Marketing & Management*, 29(5), 530–549.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3), 171–181.

Patton, C. E., & Wickens, C. D. (2024). The relationship of trust and dependence. *Ergonomics*, 67(11), 1535–1552.

Qiu, H., Li, M., Shu, B., & Bai, B. (2020). Enhancing hospitality experience with service robots: The mediating role of rapport building. *Journal of Hospitality Marketing & Management*, 29(3), 247–268.

Schwartz, S. H. (2006). Basic human values: theory, measurement, and applications. *Revue Française de Sociologie*, 47(4), 249–288.

Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5), 722–734.

Yokoi, R., & Nakayachi, K. (2021). The effect of value similarity on trust in the automation systems: A case of transportation and medical care. *International Journal of Human-Computer Interaction*, 37(13), 1269–1282.

Złotowski, J., Sumioka, H., Eyssel, F., & Ishiguro, H. (2018). Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics*, 10(5), 701–714.

Note: Figure translations are in progress. See original paper for figures.