Proceedings of the 2nd International Conference on Social Psychology and Humanity Studies
DOI: 10.54254/2753-7048/41/20240805

Who Should I Trust: Human-AI Trust Model in AI Assisted Decision-Making

Yechen Yang[1,a,*]

[1]*Hong Kong University of Science and Technology, Hong Kong SAR, 999077, China*
a. yyangfj@connect.ust.hk
**corresponding author*

Abstract: AI technology, relying on its extraordinary data searching and calculating capability, has been widely applied in assisting human decision-makers in various industries: healthcare, business management, public policies, etc. As a crucial factor influencing the performance of human and AI interaction, trust has come to be valued more in the research area in recent years. Previous studies have emphasized multiple factors that have significant impacts on the trust between human decision-makers and AI assistants. Yet, more attention needs to be paid to building up a systematic model for trust in the human-AI collaboration context. Therefore, to construct a systematic model of trust for the AI decision-making area, this paper reviews the recently conducted research, analyzes and synthesizes the significant factors of trust in the AI-assisted decision-making process and establishes a theoretical ternary interaction model from three major aspects: human decision-maker-related, AI-related, and scenario-related. Factors from the three aspects construct the three major elements of trust, which can eventually evaluate trust in the assisted decision-making process. This systematic trust model fills the theoretical gap in the current studies of trust in human-AI interaction and provides implications for further research studies in studying AI trust-related topics.

Keywords: AI Trust, Artificial Intelligence, Human AI interaction, Decision-making process, Trust modeling

1. Introduction

With the fast development of computer science and information technology in the past decades, AI techniques have been vastly developed and applied in various areas of society. The collaboration between humans and AI in the decision-making process is the research hotspot at the moment, recent study conducted by Steyvers has found three major challenges within the human-AI interaction: Human-AI complementarity, Human mental models of AI, and effective methods of interaction with AI [1], indicating the ultimate goal of the improvements in AI assistance under human decision-making context is to elevate the collaboration performance. Moreover, another study conducted earlier by Ueno pointed out a key factor that seriously impacts human-AI interactions - trust, from its significance in interpersonal interactions [2]. However, there are still limited systematical theories that can explain the mechanisms behind this process, nor are there any models and measures of trust under the human-AI collaboration contexts. Therefore, this study is trying to bridge the theoretical gap of trust in the AI-assisted decision-making process, and is going to focus on the following

DOI: 10.54254/2753-7048/41/20240805

research questions: (RQ1) What are the major elements and properties of trust within the human-AI interaction context? (RQ2) How do the major elements and properties of trust structure and operate within the human-AI interaction context? To answer the above questions, the current study will adopt the method of literature reviewing, by analyzing and synthesizing the existing works and theories, and constructing a potential theoretical model of trust for the human-AI collaboration context. Through a thorough analysis of the constructs and elements of trust, following the establishment of a systematical model of trust within such context, this study seeks a deeper understanding of trust's definition and mechanism within the AI-assisting decision-making process, that can provide guidance and implications for further researches on this topic.

2. Human decision-maker related factors

Trust is mostly valued as a critical factor in the context of interpersonal relationships, with its potential influence on human-machine interaction [2]. Correspondingly, the human-related factors should also be discussed as key factors since their importance in the decision-making process. Based on the arguments and perspectives of previous research studies, this study mainly discusses the human related factors from three aspects: cognitive involvement, individual preference, and confidence.

Human decision maker's cognitive involvement is one of the key factors that directly influences the decision outcomes and performance. According to the dual-process theories of higher cognition [3], decision maker's behaviors would be influenced by an autonomous processing system which yields to default decision outcome, in this case would lead to over-reliance on AI assistance, leading to serious consequences of mistrust in AI even it is wrong. However, such processing mechanism would be largely intervened by human cognitive involvement, which is a higher-order analytical processing system that allows an individual to think analytically [3]. Another recent experimental study conducted by Bucinca found that cognitive involvement can significantly reduce over-reliance on AI assistance; moreover, the results also indicated that such benefit of reducing AI reliance is positively correlated with the level of cognition needs of the decision maker [4].

As a factor of cognitive involvement in AI assistance, individual preference itself is one of the most critical influences on decision-making processes. An experimental study conducted by Mayer revealed the relationship between AI forecasts, trust in AI technology and individual preferences, that the trust in AI technology has a moderating effect on AI forecast's influence on the decision-making process, and such effect is moderated by the congruency between the forecast source and individual preference [5]. This finding has shown the distinct importance of individual decision preference, that individual would inclined to trust more on the AI suggestions if the forecasts provided by the AI fit in their decision preferences, indicating the tendency of self-confirmation. Another recently conducted mock test by Selton further verifies this finding. This experiment was examining on the Dutch police officers' trust in AI assistance in street-level decision-making contexts, and the results indicate that the participants are more likely to take AI suggestions while the recommendation confirms their professional judgement [6].

Self-confidence, which in this case would be related to an individual's belief in its capability or willingness to adopt AI recommendations while making decisions [7]. A recent study by Chong found that there is a vicious cycle, that individual's misattribution to poor AI's responsibility could lead to over-reliance on poor-performing AI due to low self-confidence levels [7]. This conclusion suggests a bidirectional relationship between human self-confidence and their trust in AI performance, and further emphasizes the significant impact of self-confidence on the decision-making process. Another work in the same year also indicated that self-confidence level is correlated with whether to take AI advice or not by assessing the correctness likelihood of human and AI recommendations [8]. The same study also points out the difference between confidence and an individual's actual capability, and suggests that by reducing the gap between the perceived and actual capability, the human-AI

DOI: 10.54254/2753-7048/41/20240805

collaboration performance can be elevated [8]. Furthermore, the concept of online self-efficacy was introduced in a previous study by Araujo et al, being defined as the feeling of an individual's control over online information [9]. This study has found that there is a positive association between online self-efficacy and the perceived usefulness and fairness of automated decision-making [9], further demonstrating the impact and effect of human self-confidence on the outcomes and performance of AI-assisted decision-making.

3. AI-related factors

Besides the decision make-related factors, AI-related factors should not be neglected in the case of assisting the decision-making process. According to previous research studies, three constructs of AI-related factors is discussed in this section: accuracy, explanation, and fairness.

A study conducted years ago revealed the significance of AI accuracy from 2 citizen juries in the UK, and found that jurors would value more on AI accuracy more specifically in healthcare-related scenarios, due to its positive influence on the final decisions [10], meaning that AI suggestions with higher accuracy would be favored more, and thus more likely to be considered by the decision-makers. This finding indicates that AI accuracy is positively correlated with trust in AI assistance and AI preference, especially in the healthcare-related contexts.

The research study mentioned above has also investigated the explainability of AI advices, and points out that in non-healthcare-related scenarios, explainability is usually valued equally or even more than accuracy, since its ability to allow decision-makers to understand more about the logistic process and thus avoid potential biases [10]. Two studies in the later year also have the results agreed with the above argument. Tuncer indicated in the article that the interpretability and explainability of AI recommendations would positively influent human decision-maker's trust [11]. In the same year, Becker pointed out that the appearance of procedural instructions can enhance the explainability of AI suggestions, and then improve the human-AI collaboration performance [12]. These findings have provided more evidence to the assumption that the explainability of AI advice is closely related to the preference on AI advice, and then affect the performance of human-AI interaction. Beyond that, a study has found that the emergence of explanation in AI advice is also associated with decision maker's cognitive involvement [13], meaning that an analytical reasoning system would be activated for interpreting AI suggestions. Therefore, the automation bias would be reduced as a higher level of cognitive system is involved in the decision-making process; but the likelihood of confirmation bias would increase accordingly [13].

Last but not least, the fairness of AI is also an important factor in the decision-making process. A recent research has found fairness of AI decisions is interrelated with accuracy, explainability, and is subject to specific scenarios, that for healthcare and recruitment-related cases, people would believe AI to be fairer if AI decisions express more accuracy; and for criminal-related cases, critical explanations would have more impact on the perceive of fairness [10]. Moreover, Angerschmid had indicated in an experiment that introduced AI fairness can affect the user's trust, but only under the low fairness circumstance, meaning that low introduced AI fairness is associated with low user trust, but the results are not statistically significant under high fairness levels [14].

4. Scenario-based factors

Although scenario-related factors are another crucial factor, there is only limited research focusing on them. According to the existing studies, the scenario-related factors should be discussed in two aspects: decision time scales, and scenario-based decision types.

The decision time scale was operationally defined as time horizon (long-term or short-term) of the decision by Mayer in the experiment conducted in 2023 [5]. And the study's result has shown a

DOI: 10.54254/2753-7048/41/20240805

significant positive association between decision time horizon and advice-taking rate, meaning that individuals would prefer AI advices and with higher intention of advice adoption while facing a long-term decision, and such inclination would even go beyond the preference for advices made by humans [5]. This finding further supports the argument that decision time scale is an important scenario-related factor which is positively correlated with trust in AI decisions and human-AI collaboration performance.

The other key factor influencing trust in AI decision-making is scenario-based decision types. A recent study has indicated that there is a difference in the AI-assisted decision-making logistics for different scenarios, that in healthcare-related scenes individuals would tend to value accuracy over explainability of AI suggestions, but in non-healthcare-related scenes they would value explainability as well as accuracy; moreover, in criminal-related scenes, AI fairness would be considered more than the other two factors [10]. This finding suggests that specific scenes could affect an individual's decision-making logistics on AI assistance, and then there could be indirect impact on trust and preference for AI suggestions. Scholar Tuncer also incorporated the factor of decision types in the study conducted in 2022, and the results showed that human decision-makers would prefer AI assistance in operational decisions rather than strategical decisions, since the decision makers would believe AI to have strong data analysis skills instead of managerial capabilities that involve with direct impacts on humans [11]. This conclusion agrees with the assumption that decision types are correlated with trust in AI advices, while AI decisions would be trusted more on the data analytical aspects, decision makers would not trust AI's suggestions while considering problems straight related to humans.

5. Analysis

As one of the most popular research topics, AI-assisted decision-making has been discussed many times in recent years. Most previous studies have paid more attention to the factors or variables of trust in AI decisions, only limited works have been devoted to the construction of a systematical model to illustrate the structure and logistics of trust for AI assistance in the decision-making process.

A research study conducted in 2021 by Vereschak came up with a model of trust, as shown in Figure 1, demonstrating the constructs of trust, including vulnerability, positive expectations, and attitude; within the model, the major elements of trust are also expressed in the cross-sections of the constructs: distrust, confidence, and reliance etc. [15]. This model presents not only the basic constructs of trusts, but also the interchangeability and potential interactions between the factors. Based on this well-constructed model, a new model could then be established specifically for trust in human-AI collaborative decision-making.

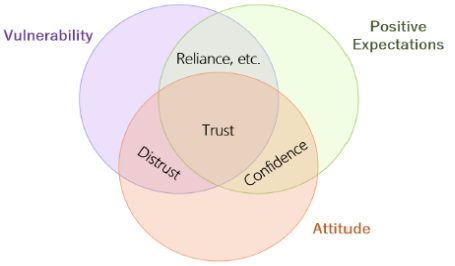

Figure 1: Constructs and key elements of trust [15].

Therefore, based on the findings of existing studies, this article establishes a theoretical model to demonstrate trust in AI decision-making. With the foundation of the trust model by Vereschak [15], and as shown in Figure 2, the new ternary interaction model is consisted of three clusters of basic

DOI: 10.54254/2753-7048/41/20240805

factors: human decision-maker-related, AI-related, and scenario-based. Each of the cluster includes several constructs that were discussed in the previous sections of this article. The interactions of the three major clusters of factors form three elements of trust in human-AI interaction. While focusing only on the AI-related and scenario-based factors, the emphasis would be on the autonomous systems and the over-reliance on AI suggestions. And when ruling out AI related factors and only considering the other two clusters of factors, the study would be more inclined to discuss about the confidence and analytical skills of the decision maker over the environmental conditions of decision-making. Moreover, the sole discussion about AI and human-related factors would be talking more about mistrust, which focuses on how human decision-makers interpret and adopt AI suggestions. The just mentioned three elements of reliance, confidence, and mistrust then constitute trust in human-AI collaborative decision-making.

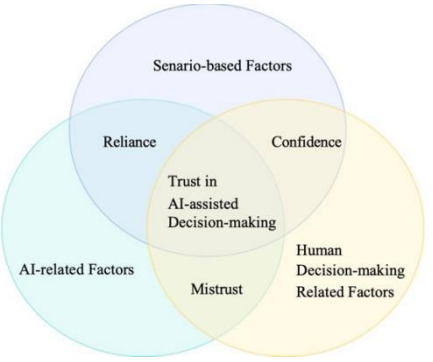

Figure 2: A simple illustration of factors and elements of trust in human-AI collaborative decision-making process.

6. Conclusion

In conclusion, with the development and popularization of artificial intelligence, increasing attention has been paid to the collaboration between humans and AI. Many researchers have focused on studying trust as a crucial factor influencing human-AI interaction, but limited effort has been devoted to the construction of a model of trust in AI-assisted decision-making. Based on the review and analysis of the evidences and perspectives of existing studies on AI decision-making and trust in human-AI interaction, the current study is categorizing the previously discussed factors into three clusters of human decision-maker-related, AI-related and scenario-based factors, to have a more explicit and comprehensive understanding on the topic of trust in human-AI collaborative decision-making. This article then proposed a ternary model of trust in the AI-assisted decision-making, which also includes three major elements of trust expressed by the interactions of the key factors: reliance, confidence, and mistrust. Meanwhile, the constructs and factors mentioned in the model are closely interrelated, meaning that the assessment of human-AI trust should be a dynamic process by considering various factors and conditions. By introducing this model, this study hopes to fill the research gaps of trust study in human-AI interaction.

Yet, the proposed model of trust in the human-AI collaborative decision-making process is a theoretical-based model, which still lacks statistical proof for credibility and validity. Through the discussion about the interactions between constructs of the model, a more detailed correlational study could be conducted in order to have a more profound understanding of the constructs and elements in the human-AI trust topics.

DOI: 10.54254/2753-7048/41/20240805

References

[1] Steyvers, M., & Kumar, A. (2023) Three Challenges for AI-Assisted Decision-Making [J] Perspectives on Psychological Science, pp.17456916231181102–17456916231181102.

[2] Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022) Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods [J] ArXiv.Org.

[3] St. B. T. Evans, J., & Stanovich, K. E. (2013) Dual-Process Theories of Higher Cognition: Advancing the Debate [J] Perspectives on Psychological Science, 8(3), pp.223–241.

[4] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021) To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making [J] Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), pp.1–21.

[5] Mayer, B., Fuchs, F., & Lingnau, V. (2023) Decision-Making in the Era of AI Support—How Decision Environment and Individual Decision Preferences Affect Advice-Taking in Forecasts [J] Journal of Neuroscience, Psychology, and Economics, 16(1), pp.1–11.

[6] Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023) "Just like I thought": Street-level bureaucrats trust AI recommendations if they confirm their professional judgment [J] Public Administration Review, 83(2), pp.263–278.

[7] Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022) Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice [J] Computers in Human Behavior, vol. 127, pp.107018.

[8] Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023) Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making [C] Conference on Human Factors in Computing Systems - Proceedings.

[9] Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence [J] AI & Society, 35(3), pp.611–623.

[10] Van Der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., et al. (2021) Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries [J] Journal of the American Medical Informatics Association: JAMIA, 28(10), pp.2128–2138.

[11] Tuncer, S., & Ramirez, A. (n.d.) (2022) Exploring the Role of Trust During Human-AI Collaboration in Managerial Decision-Making Processes [J] HCI International 2022 – Late Breaking Papers: Interacting with EXtended Reality and Artificial Intelligence, pp.541–557.

[12] Becker, F., Skirzyński, J., van Opheusden, B., & Lieder, F. (2022) Boosting Human Decision-making with AI-Generated Decision Aids [J] Computational Brain & Behavior, 5(4), pp.467–490.

[13] Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2022) Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making [J] ArXiv.Org.

[14] Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022) Fairness and Explanation in AI-Informed Decision Making [J] Machine Learning and Knowledge Extraction, 4(2), pp.556–579.

[15] Vereschak, O., Bailly, G., & Caramiaux, B. (2021) How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies [J] Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), pp.1–39.