Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Rethinking Trust in Human-AI Collaboration in the Generative AI Era

Yijie LU[a] & Bo JIANG[b*]
[a]*Department of Educational Information Technology, East China Normal University, Shanghai, China*
[b]*Shanghai Institute of AI for Education, East China Normal University, China*
* bjiang@deit.ecnu.edu.cn

Abstract: Trust has been thoroughly investigated in human-human collaboration. As large language models advance, human-AI collaboration is becoming the future trend. Generative AI can act as a collaborator. However, limited research dedicated to exploring trust in human-AI collaboration. The degree of trust is intricately connected to both the user's reliance on the system and the system's perceived usefulness. Based on a human-AI collaborative writing dataset, this work employed cluster analysis to explore collaborative patterns in the process of human-AI collaboration. The results show that trust is dynamic, two-sided, and vague element. Meanwhile, based on the changes in trust, human-AI collaboration can be categorized into three types: increasing, curvilinear, and decreasing.

Keywords: Human-AI collaboration, Human-AI interaction

1. Introduction

Trust is a subjective factor contributing to effective collaboration(Paul et al., 2016). It is related to cooperation (Tseng et al., 2009), knowledge sharing, team performance (Baruch & Lin, 2012) and engagement (Zhang et al., 2019). Factors like frequency and quality of communication (Greenlee & Karanxha, 2010) and affective commitment (Tiplic et al., 2020) influence trust in collaboration. Currently, most works focus on understanding trust in human-human collaboration, seldom consider human-AI settings.

With the development of large language models (LLM), Human-AI collaboration has emerged. An anthropomorphic AI can actively engage during collaboration. It guides human to solve the problem and improves human's innovative thinking (Siemon et al., 2020). Some objective indicators, such as accuracy, stability, mutuality and innovation, are used to evaluate the human-AI collaboration process, as well as the result indicators like artifacts and knowledge gains (Lee, Liang, et al., 2022). Procedural and subjective indicators like trust are also considered to be important factors affecting collaboration (Lee, Srivastava, et al., 2022). The performance of intelligent collaborative systems are steadily improving (Lee, Srivastava, et al., 2022). However, the impact of trust to the effectiveness of human-AI collaboration is still unclear.

This works aims to answer following questions: RQ1: What is the definition, characteristics and changing path of trust in human-AI collaboration? RQ2: What are the patterns of human-AI collaboration when taking trust into account? We design the trust indicator in human-AI collaboration. With an open-source dataset CoAuthor (Lee, Liang, et al., 2022), we explore the patterns of human-AI collaboration from the perspective of trust.

2. Related Work

Human-AI collaboration has become a very popular paradigm in the LLM era. AI collaborators have been endowed with the ability to answer questions, generate articles (Coenen et al., 2021; Lee, Liang, et al., 2022), provide suggestions(Siemon et al., 2020) and even increase creativity

(Hitsuwari et al., 2023). Some researchers focused on humans' behaviors and feelings in collaboration. Zhang investigated human's expectations of AI teammates and found that AI collaborators' competence is most valued and shared understanding with human teammates should be included (Zhang et al., 2021). However, existing research has neglected the subjective feelings of humans. Lee et al. pointed that the evaluation of human-AI collaboration should include procedural, subjective and preference indicators (Lee, Srivastava, et al., 2022).

Trust is an index that include procedural, subjective and preference attributes. The concept of human-AI trust is developed from interpersonal trust. The difference between AI and human is that AI lacks will and moral subject (Mcknight et al., 2011). Therefore, majority of studies take human-AI trust as a representation of people's willingness to adopt AI technology (Ghazizadeh et al., 2012). In Siau's study, trust to AI is defined as people's attitude that the agent will help them in uncertain situation or in risk, which can influence people's reliance on AI (Siau & Wang, 2018). In the field of automatic driving(French et al., 2018; Wojton et al., 2020), the stage of trust has been mentioned, which divided trust into tendentious trust, factual trust and post-task trust. Some researchers found that system performance like interpretability, social presence, transparency(Liu, 2021) and features of interaction like complexity, comfort and enjoyment(Bao et al., 2021) enhanced trust. Bao also believes that complexity, comfort and pleasure of interaction are influencing factors of trust. However, over-trusting is a problem (Okamura & Yamada, 2020).

To sum up, limited research existed on human-AI trust in the field of collaboration. Further exploration of the role of trust as a feature is warranted within the context of human-AI interaction. This paper aims to validate the relationship of trust and other features in collaboration as well as the evolution of trust in human-AI collaboration.

3. EXPERIMENT: Design for patterns human-AI collaboration from the perspective of trust

3.1 Data

The CoAuthor dataset(Lee, Liang, et al., 2022) is used to evaluate the proposed framework. CoAuthor is a GPT3-based collaborative writing tool. As shown in Figure 1, 58 authors from Amazon Mechanical Turk attended the experiment, writing creative articles with AI collaborators. The platform uses a text editor. They can interact with AI collaborator in two ways: a) AI writes the beginning of the article. b) Authors can press the tab key to acquire 5 suggestions from AI collaborator. The dataset includes writing sessions and survey responses. This paper extracts four features: equality, mutuality, trust and work. CoAuthor dataset is the first publicly available human-LLM collaboration dataset that contains interaction process data and users' feedback.

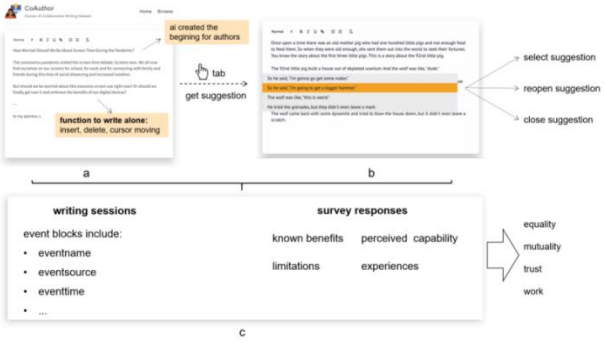

Figure 1. introduction of CoAuthor dataset

3.2 Method

3.2.1 Feature extraction

Trust is the main character we need to use in our experiment. This study defines the connotation of trust in human-AI collaboration as the collaborator's readiness to accept potential harm resulting from receiving incorrect responses or decisions from the AI partner during the collaborative process. This willingness stems from believing that AI partners can improve the quality of problem solving. Meanwhile, another two characteristics were extracted to explore the relationship between trust and collaboration: collaborative behavior and quality of works, as shown in Table 1.

Table1: The three features extracted from CoAuthor creative writing dataset

type	features	definition	source
Collaboration behavior	equality	the evenness of human's and AI's contribution to the final work.	(Storch, 2002)
	mutuality	the frequency that people and AI interact with each other and participate in each other's contributions	
Quality of works	creativity	the ability to generate new ideas	(Lee, Liang, et al., 2022)
	fluency	the ability to use words and their meanings and express smoothly	(Taylor, 1947)
	accuracy	negatively correlated with the number of errors in words and sentences	(Storch, 2005)

- Collaboration behavior feature

We rearrange the process events to make a result table taking a sentence as a unit with the attribute of source and a process table including events, source and object. Then we combined process data and result contributions to compute equality and mutuality scores. In result table, let:

$$H = \{source = human\}; A = \{source = api\}$$

represent human's and AI's contributions in the final work respectively. In the process table, let:

$$E = \{insert, delete, get, open, select, cursor - move\};\ \ I = \{E | source \neq object\}$$

represent the set of behaviors written by the human collaborator and the set of behaviors that human and AI interact with each other respectively. Given a set of events$\{e_i\}$, we define:

$$equality = 1 - \frac{\sum_i [e_i \in H] - \sum_i [e_i \in A]}{\sum_i [e_i \in H] + \sum_i [e_i \in A]}$$

where [P] = 1 if P is true and 0 if not. We also define:

$$mutuality = \frac{\sum_i [e_i \in I]}{\sum_i [e_i \in E]} \times 100\%$$

- Trust to AI

We calculate total trust in terms of both positive and negative indicators:

Trust is a subjective attitude, which can't be systematically translated into behaviors. We use the number of some specific event blocks to represent writer's trust in the process. Besides, we choose some items in the questionnaire to score negative and positive feelings on trust like "I am confident in my ability to write a story with the help of the system". Let the times a collaborator gets suggestions from AI (G) be the sets presenting writers' willingness to get help from AI collaborator. We defined:

$$trust = zscore(positive) - zscore(negative) + \frac{\sum_i [e_i \in G]}{\sum_i [e_i \in E]}$$

in which positive and negative scores are from the results of questionnaire.

- Quality of works

We used indicators to measure the quality of articles: Creativity refers to the new ideas' formed in the collaborative writing process such as roles, locations, event and etc.. Fluency means how capable authors are to express what they want with the assistance from AI collaborator, related to the transitions in passage. Accuracy is the basis of the integrity of the composition, negatively related to the amount of lexical and grammatical errors. GPT-3 is used to mark the composition.

3.2.2 Data Analysis Methods.

To answer RQ1, we extracted the characteristics of trust in human-AI collaboration. To answer RQ2, the hierarchical clustering algorithm was used to analyze the collaborative writing process with equality, mutuality and trust in human-AI collaboration, and get patterns of human-AI collaboration.

3.3 Results

3.3.1 Distribution of features

The frequency histogram (Figure 2) shows the distribution of all the features: collaborative behavior, trust, and works quality. In general, most authors divide the writing tasks evenly with the AI (avg=0.68). But authors have a relatively low score in mutuality (avg=0.05) as well as in human-AI trust (avg=-0.04) in the process of collaborative writing. The scores of the articles in creativity, fluency and accuracy are high (avg=8) and concentrated (SD < 0.8).

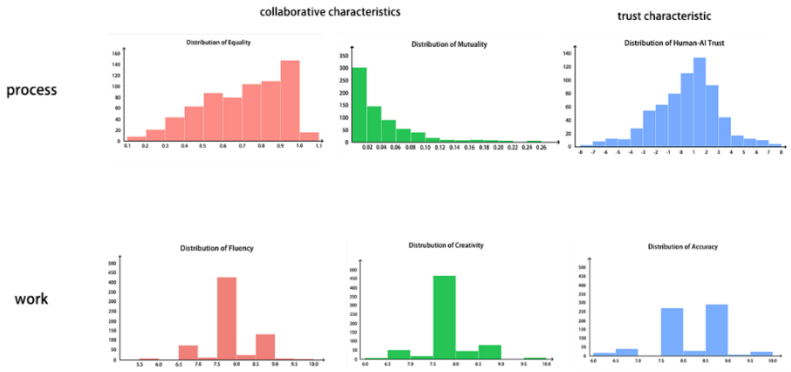

Figure 2. Frequency histogram of equality, mutuality, trust, fluency, creativity and accuracy.

3.3.2 Patterns of human-machine collaboration in the context of trust.

Trend analysis is used to perform cluster analysis of trust change trends. To take changes in trust during collaboration into account, we converted the trust data obtained per minute into a percentage of the total trust value, based on the total time it took the authors to complete the collaborative writing. It was found that trust is indeed dynamic during collaboration, and there are three main patterns in human-AI collaboration as Figure 3: growth pattern in which the authors increasingly trust and get suggestions from the AI collaborator; curve pattern in which authors' trust increases at first and then decreases; decreasing pattern in which authors keep a high level of trust at the beginning but gradually become distrustful of AI collaborators. To focus on the collaborative characteristics of the following three patterns, it is not difficult to find that the authors of the first two modes have a higher frequency of interaction with AI and better task allocation. However, in terms of the results of collaborative writing, there is also no significant difference in the creativity, fluency and accuracy of the compositions completed under the three collaborative writing patterns.

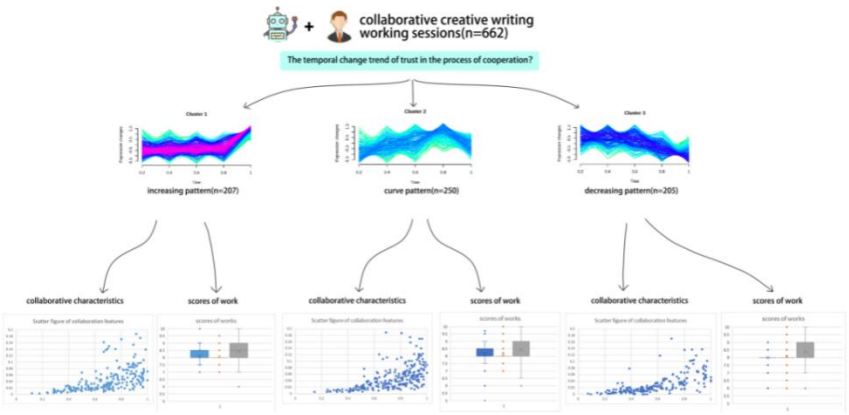

Figure 3. The time series clustering analysis.

4. Discussion—a conceptual framework of trust in human-AI collaboration

Currently, the existing research on human-AI trust is still concentrated in the stage of theory and measurement. Some studies have found the effect of trust on collaboration, or enhanced human-AI trust through design, but few study explored the characteristics and effects of trust as a process factor. Based on the data set of collaborative writing, this study conducted experiments on trust and proposed a theoretical framework of trust in human-AI collaboration.

We extracted human-AI trust indicators from process data and questionnaire results. The average value and distribution of trust are consistent with definition, which proves that the indicator can be used for measurement of human-AI trust. Further, some unique characteristics of trust in this context have been found: 1) **Dynamic.** Through the results of cluster analysis, we found that trust of almost all participants during the collaboration changes. 2) **Two-sided.** In cluster analysis, there is no difference in final work between different types of trust collaboration, which suggests that trust can be two-sided. 3) **Vague.** From the results of the distribution of trust, we found that in human-AI collaboration, one's trust in AI does include the cognition of AI's ability and emotion as multiple roles like a tool or a team member, indicating that trust is ambiguous.

When the trend of trust is considered, the process of human-AI collaboration can be divided into three patterns: trust rising, the author's trust gradually increases with the interaction process; In the arc trust type, the author's trust in AI rises at the beginning, and then decreases after reaching the peak; Trust decline, the author initially maintained a high level of trust in AI, and then trust gradually decreased. In the first case, this may be because AI is similar to author's understanding of the article, or gives feedback that pleases the author, and in the third case the situation is opposite. At the same time, we found no significant difference in the creativity, fluency, and accuracy of human-machine collaborative works regardless of the type of clustering, which may be because when parallel human-AI collaborative writing is performed, the author's writing ideas tend to responses, which has been demonstrated in the design field.

5. Conclusion

In this work, we identified the importance of researching procedural and subjective characteristics, especially trust in human-AI collaboration. We argued that trust can influence learner's attitude to AI and high mutuality will improve learner's knowledge of AI collaborators. We also found that over-trusting may be existed in human-AI collaboration. In the future work, we will use human-AI questionnaires to measure trust score and design the experiment to demonstrate the result in other human-AI collaboration activities and explore more factors that can influence trust and how to design a system that can control human-AI trust within bounds.

References

Bao, Y., Cheng, X., De Vreede, T., & De Vreede, G.-J. (2021). Investigating the relationship between AI and trust in human-AI collaboration.

Baruch, Y., & Lin, C.-P. (2012). All for one, one for all: Coopetition and virtual team performance. *Technological Forecasting Social Change*, *79*(6), 1155-1168.

Coenen, A., Davis, L., Ippolito, D., Reif, E., & Yuan, A. (2021). Wordcraft: a human-ai collaborative editor for story writing. *arXiv preprint arXiv:.07430*.

French, B., Duenser, A., & Heathcote, A. (2018). Trust in automation–a literature review. *Commonwealth Scientific Industrial Research Organisation*.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. *Cognition, Technology Work*, *14*, 39-49.

Greenlee, B. J., & Karanxha, Z. (2010). A study of group dynamics in educational leadership cohort and non-cohort groups. *Journal of Research on Leadership Education*

5(11), 357-382.

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior*, *139*, 107502.

Lee, M., Liang, P., & Yang, Q. (2022). Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. Proceedings of the 2022 CHI conference on human factors in computing systems,

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., . . . Rong, F. (2022). Evaluating human-language model interaction. *arXiv preprint arXiv:.09746*.

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384-402.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems*, *2*(2), 1-25.

Paul, R., Drake, J. R., & Liang, H. (2016). Global virtual team performance: The effect of coordination effectiveness, trust, and team cohesion. *IEEE Transactions on Professional Communication*, *59*(3), 186-202.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, *31*(2), 47-53.

Siemon, D., Li, R., & Robra-Bissantz, S. (2020). Towards a model of team roles in human-machine collaboration. *Forty-First International Conference on Information Systems*.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language learning*, *52*(1), 119-158.

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of second language writing*, *14*(3), 153-173.

Taylor, C. W. (1947). A factorial study of fluency in writing. *Psychometrika*, *12*(4), 239-262.

Tiplic, D., Elstad, E., Brandmo, C., Steingrímsdóttir, M., & Engilbertsson, G. (2020). Perceived organizational antecedents of emerging collaborative learning activities among icelandic beginning teachers. *Scandinavian Journal of Educational Research*, *64*(6), 801-815.

Tseng, H., Wang, C., Ku, H.-Y., & Sun, L. (2009). Key factors in online collaboration and their relationship to teamwork satisfaction. *Quarterly Review of Distance Education*, *10*(2), 195-206.

Wojton, H. M., Porter, D., T. Lane, S., Bieber, C., & Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *The Journal of social psychology*, *160*(6), 735-750.

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1-25.

Zhang, X., Meng, Y., de Pablos, P. O., & Sun, Y. (2019). Learning analytics in collaborative learning supported by Slack: From the perspective of engagement. *Computers in Human Behavior*, *92*, 625-633.