Artificial Intelligence, Social Computing and Wearable Technologies, Vol. 113, 2023, 304–315
https://doi.org/10.54941/ahfe1004203

The Impact of AI Transparency and Reliability on Human-AI Collaborative Decision-Making

Xujinfeng Wang, Yicheng Yang, Da Tao, and Tingru Zhang

Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China

ABSTRACT

Human-AI collaborative decision-making has become a prevalent interaction paradigm, but the lack of transparency in AI algorithms presents challenges for humans to understand the decision-making process. Such lack of comprehension can lead to issues of over-reliance or under-reliance on AI recommendations. In this study, we focused on a human-AI collaborative income predicting task and investigated the influence of AI transparency and reliability on task performance. The results revealed that when AI reliability was high (75% and 90%), transparency had no significant effects on human decision-making. However, at a lower level of reliability (60%), higher transparency levels led to increased compliance with AI suggestions, thereby demonstrating a persuasive effect. Further analysis indicated that compliance rates only improved when AI made correct decisions, rather than when AI made incorrect ones. However, transparency did not significantly impact humans' ability to correctly reject erroneous recommendations from AI, suggesting that increasing transparency alone did not enhance humans' error detecting ability. In conclusion, when the reliability of AI is low, heightening transparency can promote appropriate dependence on AI without elevating the risk of over-reliance. Nevertheless, further research is necessary to explore effective strategies that can assist humans in identifying AI errors effectively.

Keywords: Human-AI collaboration, Transparency, Reliability, Compliance

INTRODUCTION

With the rapid development of artificial intelligence (AI) technology, the traditional paradigm of human-machine interaction is transitioning towards human-AI collaboration (Rieth & Hagemann, 2022). AI possesses not only formidable computing capabilities but also cognitive abilities such as logical reasoning and learning, which hold immense potential for enhancing performance efficiency. However, while this has significantly improved the accuracy of AI algorithms, it has also caused challenges in terms of explaining and comprehending the calculation process and causality behind AI decisions, often referred to as the "black box" challenge (Zhang et al., 2020). When human are unable to understand how AI arrives at decisions, it can lead to a reduced sense of security and trust in it, and may even result in reluctance

to use it (Glikson & Woolley, 2020). In addition, the opaque nature of black box makes it more difficult for humans to identify AI errors. As a result, when AI makes mistakes or malfunctions, humans may still blindly follow its decisions, leading to issues of so called over-trust or over-reliance (Bussone et al., 2015). Several fatal accidents of self-driving cars in recent years have highlighted the severity of this problem. Therefore, it is crucial to develop strategies to enhance human comprehension of AI, foster calibrated trust and usage behavior, and ultimately improve human-AI collaboration performance.

Improving AI transparency has been proposed as an effective way to make AI decision-making understandable to human. Chen proposed the Situation Awareness-based Agent Transparency (SAT) model, which is a transparency model based on situational awareness theory (Chen & Barnes, 2014). Situation Awareness (SA) refers to the internal representation of individuals coping with the changing external environment, including the perception, understanding and prediction of various elements in the environment. The SAT model proposes that AI transparency involves conveying information such as the state, intention, reasoning process, and future plans of the machine to the user through a well-designed interactive interface to assist the user in comprehending the output of the machine (Bhaskara et al., 2021). The model posits that three types of information, corresponding to the three stages of SA, can be provided to humans. The first type of information is related to the purpose or intention of AI, supporting perception SA. The second type involves providing the rationale for AI decisions, such as why AI recommends to take one action instead of another, to support understanding SA. Moreover, AI can supply information regarding uncertainties related to future outcomes, supporting prediction SA. It should be noted that these three types of transparent information should be presented in a sequential and progressive manner to gradually increase the transparency level of AI.

Within the theoretical framework outlined above, several studies have been carried out to investigate the effects of AI transparency on trust and human-AI collaboration performance. Early studies primarily focused on the cooperative scheduling tasks of unmanned aerial vehicles (UAVs) and unmanned vehicles (UVs) (Bhaskara et al., 2021; Lyons et al., 2017; Mercado et al., 2016; Stowers et al., 2020). For example, Mercado et al. (2016) explored how the transparency level of intelligent agents affects the performance, trust, and workload of operators in UV dispatching tasks. Their results revealed that increased transparency improved task performance and trust, but it also induced greater workload and longer response time. Lyons et al. (2017) examined the impact of transparency of the automated assistance equipment on trust in aircraft emergency landing tasks and discovered that the higher transparency levels corresponded to higher levels of trust. More recent studies have expanded their focus to encompass a wider array of collaborative tasks such as disease diagnosis (Fischer et al., 2018), human-machine co-driving (Kunze et al., 2019; Oliveira et al., 2020), and battlefield enemy situation analysis (Selkowitz et al., 2017; Wright et al., 2020). For instance, Fischer et al. (2018) investigated the issue of trust in human-AI interactions in the context of healthcare, specifically in a scenario

where a human and a robot collaborated to measure blood pressure. The results demonstrated that increasing transparency, specifically in terms of AI explaining its behaviors and capabilities, consistently influenced users' trust and perceived comfort.

Despite the valuable insights obtained from the studies above, consistent conclusions have not yet been reached, and many of these investigations did not account for potential confounding effects of other significant factors. Therefore, this study aimed to develop different levels of transparency for a human-AI collaborative decision-making task based on the SAT theory and evaluate the effects of AI transparency on trust and task performance. Besides, we developed AI with different reliability and explored if reliability would moderate the effects of AI transparency. Results of this study will contribute to the design of AI interface and therefore can promote the development and practical application of AI technology.

METHODS

To simulate a human-AI collaborative decision-making task, we conducted an income predicting task in which participants were asked to predict whether an individual's annual income would exceed $50K. Detailed task and experiment design are presented below.

Participants

We recruited 54 participants (27 males, all Chinese) to participate in this experiment. Their average age was 22.8 years old (SD = 2.2). The average duration of the experiment was 60 min, and the subjects were compensated for their participation and experiment performance.

Income Predicting Task

The dataset utilized in this experiment came from a publicly available dataset in the UIC Machine Learning Repository (https://www.kaggle.com/competitions/-cs5228/data). The dataset comprises 24,421 instances of individuals, each described by 12 attributes such as annual income, age, and educational level. Participants were required to predict whether an individual's annual income would be above or below $50K based on seven features, including age, job type, educational level, marital status, occupation, capital gain, and working hours per day.

We have developed an Income Predicting AI (IPAI) by dividing the dataset into training set and test set with a 7:3 random split. The model was trained by random forest algorithm. By randomly selecting data and controlling the proportion of data that AI suggestions are true or false, the reliability of IPAI could be manipulated to different levels.

The experiment system includes three graphical interfaces (Figure 1): decision-making interface, feedback interface, and trust reporting interface. The decision-making interface consists of three sections: basic information section, AI prediction section, and decision-making section. The basic information section displays the information of the individual to be predicted. The AI prediction section presents transparency related information. At the low

transparency level, only the AI prediction result (area ① in Figure 1) is shown. At the medium transparency level, in addition to the prediction result, explanations regarding why AI reaches the decision are provided by displaying information about the two most important features used in AI prediction. For instance, area ③ in Figure 1 shows the proportion of individuals with an income of more than $50K with different educational levels and different job types, with the bar related to the specific individual highlighted in red. At the high transparency level, besides the information mentioned above, AI confidence score, which is output by the algorithm reflecting the probability that IPAI's prediction is correct, is provided (area ② in Figure 1).

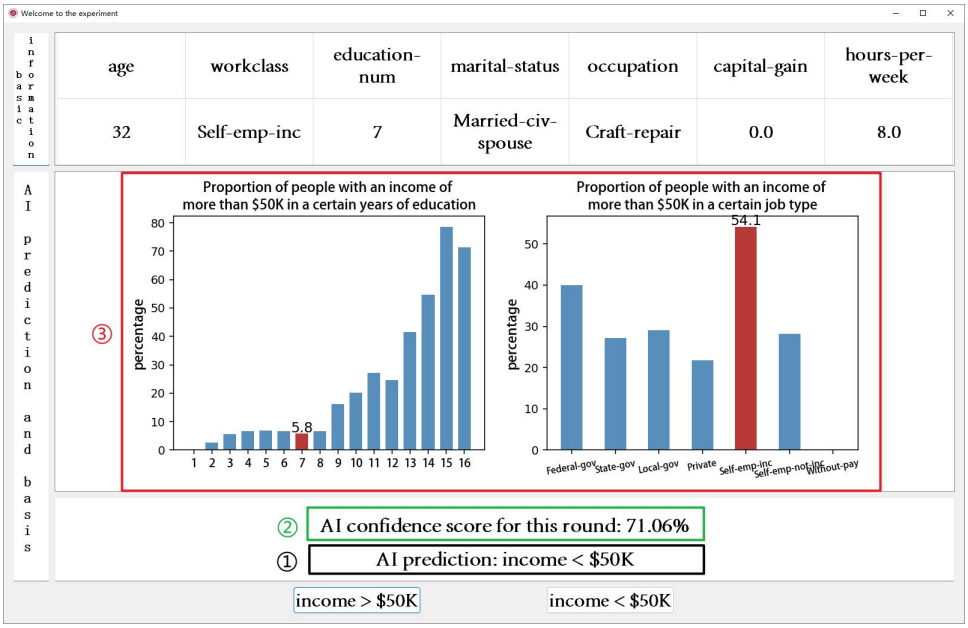

Figure 1: The decision-making interface.

After making their decisions, participants were presented with one of the four feedback interfaces (Figure 2), depending on the correctness of both human and AI decisions. Subsequently, the trust reporting interface would appear and require participants to report their trust towards IPAI in this round on a 7-point Likert scale (1 represents "very distrustful" and 7 represents "very trustworthy").

Experiment Design

This study employed a 3 (AI transparency: low, medium, and high) × 3 (AI reliability: 60%, 75%, and 90%) × 2 (gender: male and female) factorial design. AI transparency and gender were considered as between-subject variables, while AI reliability was a within-subject variable. AI reliability refers to the proportion of correct suggestions given by IPAI. For instance, a reliability of 60% means that 12 of the 20 AI predictions in each experiment trial were correct.

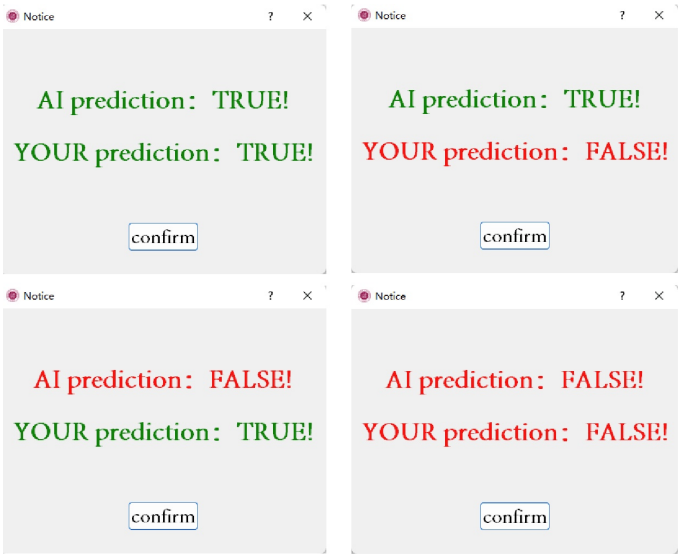

Figure 2: The feedback interface.

The dependent variables include AI compliance rate, human decision type (correct acceptance, incorrect acceptance, correct rejection, and incorrect rejection) and decision-making time. The dependent variables and their meanings are shown in Table 1.

Table 1. Dependent variables and meanings.

Dependent variables		Meaning
AI compliance rate		The percentage of trials when the participant's decision is consistent with the IPAI's suggestion to the total number of decisions.
Decision type	Correct acceptance	Number of trials that IPAI is correct and the participant agrees with IPAI.
	Incorrect acceptance	Number of trials that IPAI is incorrect and the participant agrees with IPAI.
	Correct rejection	Number of trials that IPAI is incorrect and the participant disagrees with IPAI.
	Incorrect rejection	Number of trials that IPAI is correct and the participant disagrees with IPAI.
Decision-making time		The time elapsed from the appearance of the decision-making interface to the participant pressing the decision-making button.

Procedure

The schematic representation of the experiment procedure is shown in Figure 3. All participants first provided informed consent and filled out

the pre-experiment questionnaires prior to the experiment. Then they practiced to get familiar with the prediction task and the experiment interface. In the formal experiment, each participant was required to complete 60 rounds of decision-making tasks, 20 under each of the three AI reliabilities, with the assistant of the IPAI showing either the low-, medium-, or high-transparency. The order of the three reliability levels followed the Latin square design to eliminated potential confounding effects of order. There was five minutes break after completing all decisions under one AI reliability. After the experiment, participants were asked to fill out a questionnaire about their subjective perception of IPAI, which included perceived explainability, perceived transparency, and perceived usefulness, etc.

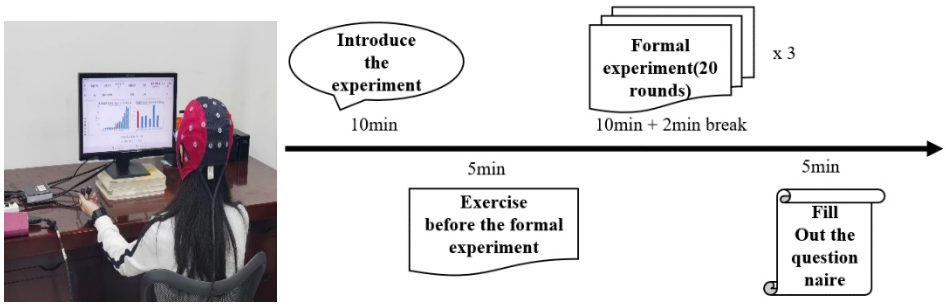

Figure 3: The flow chart of the formal experiment.

Statistical Analysis

A 3×3×2 mixed-effects ANOVA was applied to investigate the effects of AI transparency, AI reliability, and gender on dependent variables. Only the main effects and the two-way interaction effects were analyzed.

RESULTS

The ANOVA results of the effects of AI transparency, Reliability, and Gender on decision-making are summarized in Table 2.

Table 2. Summary of the ANOVA results on decision-making.

Dependent variable	Transparency	Reliability	Gender	Transparency * Reliability	Transparency * Gender	Reliability * Gender
AI compliance rate	F = 0.192, p=0.826	**F = 20.45, p<0.001**	F = 0.234, p=0.624	**F = 5.536, p<0.001**	F = 0.669, p=0.502	F = 0.486, p=0.617
Correct acceptance	F = 0.461, p=0.633	**F = 207.4, p<0.001**	F = 1.293, p=0.261	**F = 3.876, p=0.005**	F = 0.478, p=0.623	F = 0.539, p=0.585
Incorrect acceptance	F = 0.117, p=0.890	**F = 320.6, p<0.001**	F = 0.916, p=0.343	F = 1.418, p=0.234	F = 1.272, p=0.290	F = 0.177, p=0.838
Correct rejection	F = 0.068, p=0.934	**F = 89.21, p<0.001**	F = 0.780, p=0.381	F = 1.308, p=0.272	F = 1.189, p=0.313	F = 0.134, p=0.874
Incorrect rejection	F = 0.419, p=0.660	F = 0.372, p=0.691	F = 1.189, p=0.280	**F = 4.086, p=0.004**	F = 0.450, p=0.640	F = 0.487, p=0.616

AI Compliance Rate

The results of AI compliance rate revealed that the interaction effect of transparency and reliability ($F_{(4,153)} = 5.536$, $p < 0.001$) was significant, while the interaction effect of transparency and gender, gender and reliability were not significant. Figure 4 shows the results of the simple main effect analysis on this significant interaction effect. As depicted in Figure 4a, when AI reliability was 60%, medium ($p = 0.006$) and high ($p = 0.026$) transparency levels resulted in a significantly higher compliance rate than the low transparency condition, while the difference between medium and high was not significant ($p = 0.863$). When the reliability is 75% and 90%, no significant effects of transparency on compliance rate were identified. Moreover, Figure 4b suggested regardless of the transparency level, the compliance rate tended to be significantly higher when AI reliability was 90% compared to the other two reliabilities.

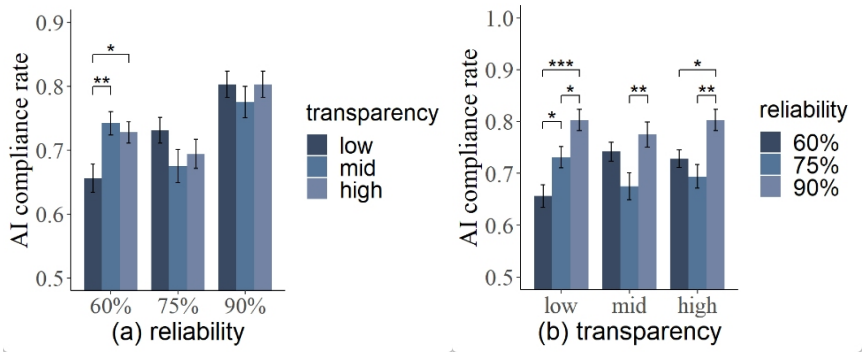

Figure 4: The results of the interaction effect analysis of transparency and reliability on AI compliance rate.

Types of Decision-Making

AI compliance consists of two types of decision, namely correct acceptance and incorrect acceptance. We separately analyzed the effects of transparency, reliability, and gender on these two types of decision and presented the results in Table 2. With regard to correct acceptance, transparency and reliability had a significant interaction effect ($F_{(4,153)} = 3.876$, $p = 0.005$). Simple main effect analysis (Figure 5a) showed that when the reliability was 60%, the number of correct acceptances under medium ($p = 0.007$) and high ($p = 0.020$) transparency were significantly higher than that under low transparency, while that between medium and high transparency had no significant difference ($p = 0.919$). The simple main effect of transparency was insignificant at both 75% and 90% reliability (Figure 5a). Moreover, as expected, correct acceptance increased with reliability (Figure 5b). Interestingly, for incorrect acceptance, no transparency-related effects were found. These results suggested that when AI reliability was low (i.e., 60%), higher transparency increased AI compliance rate of the operator, by promoting more correct acceptances, not more incorrect acceptances.

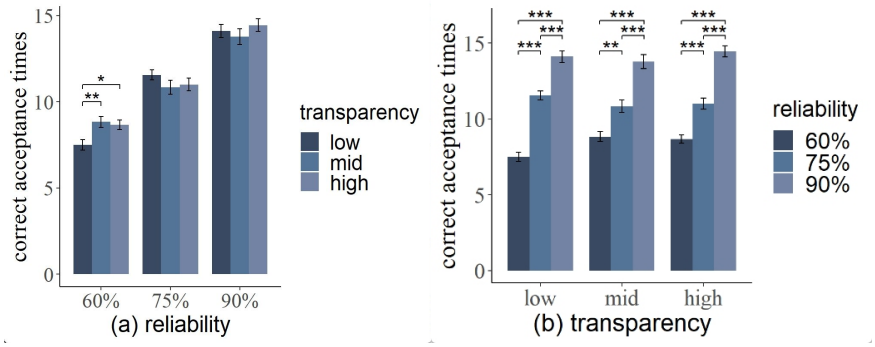

Figure 5: The results of the interaction effect analysis of transparency and reliability on correct acceptance times.

Similar analysis was performed on correct rejection and incorrect rejection. The results demonstrated that transparency and reliability had a significant interaction effect on incorrect rejection ($F_{(4,153)} = 4.086$, $p = 0.004$). Specifically, when the reliability was 60%, there were significantly fewer incorrect rejections under the medium ($p = 0.007$) and high ($p = 0.020$) transparency levels than those under the low transparency level (Figure 6a). Interestingly, the differences in incorrect rejection across different reliability levels were not significant (Figure 6b). For correct rejection, no transparency-related effect was found.

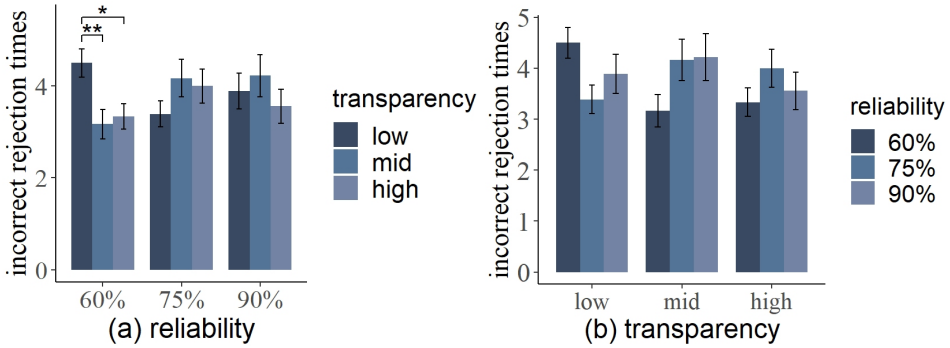

Figure 6: The results of the interaction effect analysis of transparency and reliability on incorrect rejection times.

Decision-Making Time

The effects of AI transparency, Reliability and Gender on decision-making time are summarized in Table 3.

The results revealed no significant interaction effects and the main effects of reliability and gender on decision time were not significant. Only the main effect of transparency ($F_{(2,51)} = 3.303$, $p = 0.045$) reached the significance level. Post-hoc pairwise comparison (Figure 7a) showed that the decision-making time in the high transparency level was significantly longer

than that in the low level of transparency ($p = 0.035$), while the difference between the medium and low transparency level ($p = 0.456$), and the medium and high transparency level were not significant ($p = 0.367$).

Table 3. Summary of the ANOVA results on decision-making time.

Dependent variable	Transparency	Reliability	Gender	Transparency * Reliability	Transparency * Gender	Reliability * Gender
Overall	**F = 3.303, p=0.045**	F = 1.619, p=0.203	F = 0.754, p=0.389	F = 0.646, p=0.631	F = 0.947, p=0.395	F = 1.194, p=0.307
correct acceptance	F = 2.677, p=0.078	F = 1.734, p=0.182	F = 0.920, p=0.342	F = 0.737, p=0.569	F = 1.385, p=0.260	F = 1.000, p=0.372
incorrect acceptance	F = 2.897, p=0.064	F = 1.741, p=0.180	F = 0.001, p=0.971	F = 0.498, p=0.737	F = 0.536, p=0.589	**F = 3.330, p=0.039**
correct rejection	**F = 4.842, p=0.011**	**F = 29.27, p<0.001**	F = 0.415, p=0.522	F = 1.511, p=0.205	F = 1.128, p=0.332	F = 0.059, p=0.943
incorrect rejection	F = 3.083, p=0.055	F = 1.352, p=0.263	F = 0.665, p=0.418	F = 1.684, p=0.160	F = 0.217, p=0.805	F = 0.650, p=0.524

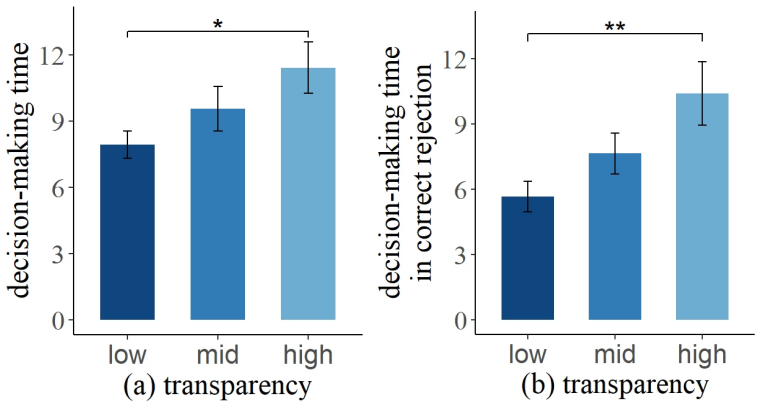

Figure 7: a) A post-hoc test of transparency on decision time. b) The results of the main effect analysis of transparency on decision time in correct rejection.

To thoroughly understand the effect of transparency, we separately analyzed the effects of transparency on decision time under different decision types. The results (Table 3) showed that only decision time under correct rejection was significantly affected by AI transparency. Particularly, transparency showed a significant main effect ($F_{(2,51)} = 4.84$, $p = 0.011$) and no significant interaction effects. Post-hoc pairwise comparison (Figure 7b) suggested that the decision-making time under the high transparency level was significantly longer than that under the low transparency level ($p = 0.009$). With regard to other decision types, no significant transparency-related effects were identified. Taking together, these results indicated that increasing transparency level only had a negative effect on decision-making time when the decision type is correct rejection.

DISCUSSION

This study aimed to investigate the role of AI transparency on Human-AI collaborative decision-making between humans and AI. Possible mediating effects of AI reliability and gender have also been considered.

The results showed that a higher level of transparency promoted a greater AI compliance rate, but only when AI reliability was relatively low (i.e., 60%). This is consistent with Mercado et al. (2016), who also observed a persuasive effect when AI disclosed more information about its decision-making process. A detailed analysis showed that higher transparency level only promoted more correct acceptances without inducing more incorrect acceptances. This finding suggested that the potential automation bias associated with an overload of information was not found. This is consistent with Vasconcelos et al. (2023) who also reported that automation bias was not a problem when AI explained its decision-making mechanism to users. However, in contrast, Vered et al. (2023) found that explanations did not reduce automation bias and, in some cases, even increased it. Possible explanations for such controversy might be attributed to differences in experimental tasks. This study and Vasconcelos et al. (2023) used similar predicting tasks while Vered et al. (2023) have adopted a detection task

Furthermore, the results indicated that increasing transparency level contributed to reduced incorrect rejection when AI reliability was low. This suggested that increasing transparency level can help mitigate unwarranted distrust towards AI. Unwarranted distrust might be a challenge in collaboration when AI reliability was low as human may maintain a generally low trust. Our results suggest that explanations about AI decision-making can be an effective way to identify possible correct predictions provided by AI and therefore enhance the usefulness of AI. The effect of transparency on correct rejection was not significant, suggesting that increasing transparency did not contribute to an improved ability in identifying AI errors. This was aligned with the results of Poursabzi-Sangdeh et al. (2021), who also found that increasing the level of transparency did not significantly improve human performance in identifying AI errors and, in some cases, may even have a detrimental effect. However, while the effect of transparency on correct rejection was not significant, we did observe a longer decision-making time when participants made a correct rejection under high transparency condition. This implied that information about important features and confidences might have inspired more analytical thinking when participants were about to make a correct rejection. Unfortunately, the prolonged thinking might have not transferred to improved error identification.

To sum up, we believe that when the reliability of AI is low, increasing the level of transparency can help increase individuals' correct dependence on AI without increasing the risk of over-reliance, and can bring about a greater improvement in work performance. However, the impact of a higher level of transparency on identifying AI errors still needs to be verified by further research.

ACKNOWLEDGMENT

This research was supported by Guangdong Provincial Natural Science Foundation General Project (2021A1515011610), Shenzhen Basic Research General Project (JCYJ20210324100014040) and National Natural Science Foundation of China (72071170).

REFERENCES

Bhaskara, A., Duong, L., Brooks, J., Li, R. Y., McInerney, R., Skinner, M., Pongracic, H., & Loft, S. (2021). Effect of automation transparency in the management of multiple unmanned vehicles. Applied Ergonomics, 90, Article 103243.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems 2015 International Conference on Healthcare Informatics.

Chen, J. Y. C., & Barnes, M. J. (2014). Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. IEEE Transactions on Human-Machine Systems, 44(1), 13–29.

Fischer, K., Weigelin, H. M., & Bodenhagen, L. (2018). Increasing trust in human–robot medical interactions: effects of transparency and adaptability. Paladyn, Journal of Behavioral Robotics, 9(1), 95–109.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. Ergonomics, 62(3), 345–360.

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., Smith, D., Johnson, W., & Shively, R. (2017). Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. In (pp. 127–136). Springer International Publishing.

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. Human Factors, 58(3), 401–415.

Oliveira, L., Burns, C., Luton, J., Iyer, S., & Birrell, S. (2020). The influence of system transparency on trust: Evaluating interfaces in a highly automated vehicle. Transportation Research Part F: Traffic Psychology and Behaviour, 72, 280–296.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021, 2021). Manipulating and Measuring Model Interpretability.

Rieth, M., & Hagemann, V. (2022). Automation as an equal team player for humans? - A view into the field and implications for research and practice. Appl Ergon, 98, 103552.

Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. C. (2017). Using agent transparency to support situation awareness of the Autonomous Squad Member. Cognitive Systems Research, 46, 13–25.

Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. (2020). The IMPACT of Agent Transparency on Human Performance. IEEE Transactions on Human-Machine Systems, 50(3), 245–253.

Vasconcelos, H., Jörke, M., Grunde-Mclaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–38.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. Artificial Intelligence, 322, 103952.

Wright, J. L., Chen, J. Y. C., & Lakhmani, S. G. (2020). Agent Transparency and Reliability in Human-Robot Interaction: The Influence on User Confidence and Perceived Reliability. Ieee Transactions on Human-Machine Systems, 50(3), 254–263.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.