Artificial Intelligence, Social Computing and Wearable Technologies, Vol. 113, 2023, 126–135
https://doi.org/10.54941/ahfe1004186

User Trust Towards an AI-Assisted Healthcare Decision Support System Under Varied Explanation Formats and Expert Opinions

Da Tao[1], Zehua Liu[1], Tingru Zhang[1], Chengxiang Liu[1], and Tieyan Wang[2]

[1]College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, 518060, China
[2]Xiamen Meiya Pico Information Co., Ltd, Xiamen, 361000, China

ABSTRACT

While Artificial Intelligence (AI) has been increasingly applied in healthcare contexts, how AI recommendations should be explained to achieve higher user trust is yet to be determined. This study was aimed to investigate users' trust towards an AI-assisted healthcare decision support system under varied explanation formats and expert opinions. Twenty participants participated in a lab-based experiment where they were asked to complete a series of dosage adjustment tasks in chronic disease care scenarios with the help of a simulated AI-assisted decision support system. Four explanation formats and three types of expert opinions were examined. Data on subjective trust, task performance and physiological measures were collected. The results showed that explanation formats had significant effects on subjective trust, task performance and skin conductance. Expert opinion had significant effects on subjective trust and task performance. There existed an interaction effect on compliance rate between explanation format and expert opinion. It appears that AI recommendations that are explained by counterfactual reasoning way and supported by medical experts are likely to achieve higher user trust. The findings can provide references for better design of explainable AI in AI-assisted healthcare contexts.

Keywords: Explainable AI, Trust, Healthcare, Explanation format, Expert opinion

INTRODUCTION

In recent years, artificial intelligence (AI) systems have been widely applied in healthcare (Contreras and Vehi, 2018; Kavakiotis et al., 2017; Schachner et al., 2020). A number of governments have launched national-level initiatives to promote the use of AI-assisted healthcare decision support systems (AIHDSSs) to allow for enhanced healthcare quality, and improved health services for users (Petersson et al., 2022). For example, the national strategy report for next generation of AI in China has emphasized the development of intelligent medical care that aims to establish intelligent healthcare service system, develop collaborative medical robots and intelligent diagnosis assistants, and strengthen intelligent healthcare and health management. However, AI's inherent black box properties have prevented AI from penetrating more deeply into the healthcare field. This is because that users who lack of professional knowledge on AI may suffer from difficulty in accepting AI recommendations due to the complexity of AI algorithms (Rüping, 2006). In addition, AI with black-box attributes may generate unpredictable risks and accidents during their applications, which further triggers a crisis of trust in AI among users (e.g., distrust and under-trust) (Longoni et al., 2019; Shaffer et al., 2013).

One of effective ways to address the AI trust problem is to enhance the explainability of AI, which can be considered as the degree to which AI has the ability to explain how and why the algorithm comes to particular outcomes (Arrieta et al., 2020; Shin, 2021). Explainability is widely recognized as one of basic elements of explainable AI (XAI), and as a prerequisite for AI fairness and credibility (Sullivan et al., 2020). It is suggested that when AI systems can be well explained to be understood by users, they are more likely to be trusted (Shin, 2021). Conversely, an AI system that is hard to understand may reduce user trust and even lead to reduced task performance. In light of this, XAI has gained increasing research attention in recent years (Gunning and Aha, 2019). A number of studies have examined the ways to explain AI to users (Holzinger et al., 2019; Kim et al., 2020). For example, Silva et al. examined a wide range of AI explanation types, and found varied effectiveness of the explanation types (Silva et al., 2023). In particular, they highlighted the advantages of counterfactual explanations and emphasized the shortcomings of confidence scores in explaining AI. While previous studies mostly examined explainability in general AI scenarios, few studies have examined explainability in healthcare domain (Loh et al., 2022; Van der Waa et al., 2021). In addition, current explainability methods may be unlikely to engender trust for patient-level decision support (Ghassemi et al., 2021). How AI recommendations should be explained in healthcare deserves further examination.

Expert opinion is another important factor that could affect user trust. People often tend to consult experts' opinions when they make decisions, especially in high-stakes scenarios like healthcare. For example, on Twitter, a professional doctor is regarded with higher credibility than a layperson, even though both have many followers (Lee and Sundar, 2013). There are also cases that patients prefer doctors' independent diagnosis to the use of computer-assisted decision-making systems (Shaffer et al., 2013). Users are more inclined to medical services provided by humans instead of AIHDSSs (Longoni et al., 2019). In contrast, Wang et al. (2020) showed that AI systems are as influential as human for online recommendations if both act as experts. To what extent the endorsement on AIHDSS by medical experts would affect user trust in the AI systems appears unclear.

In summary, although AIHDSSs have shown growing popularity in healthcare, the black box attributes of AI have present difficulty for users to trust in their recommendations for healthcare activities. In light of this, the present study aimed to examine the effects of explanation format and expert opinion on user trust towards a simulated AIHDSS in chronic disease care scenarios.

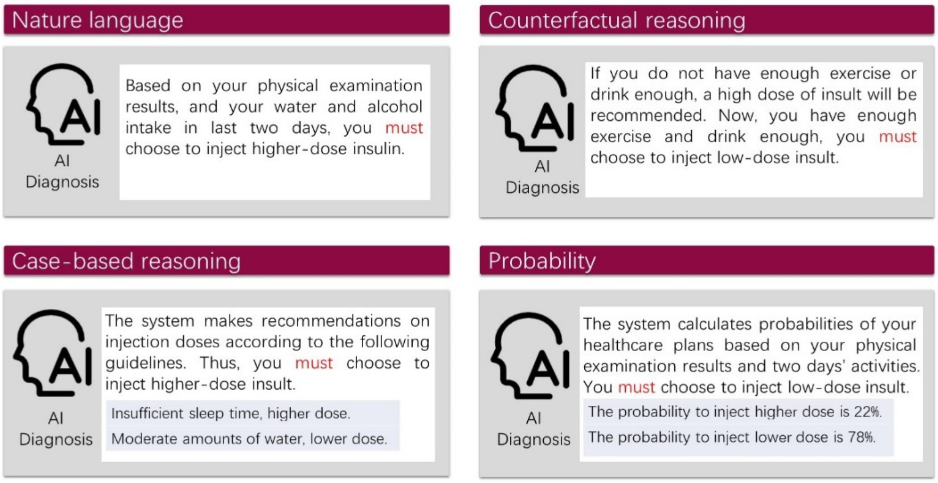
METHODS

Participants

Twenty university students (10 males and 10 females, mean age = 22.7 years (SD = 1.2 years)) participated in the experiment. On average, their self-reported health status during the past three months was 5.2 (SD = 0.8) in a 7-point Likert scale. Their average electronic health literacy was 4.9 (SD = 0.1), as measured by a 5-item 7-point Likert scale. All participants reported to have normal or corrected-to-normal vision. The participants gave informed consent before the experiment.

Experimental Design

This study employed a two-factor within-subjects design, with explanation format (i.e., natural language explanation, counterfactual reasoning explanation, probability explanation and case-based reasoning explanation) and expert opinion (i.e., expert in support of and against AI decision-making, and absence of expert opinion) serving as independent variables. The four AI explanation formats (Figure 1) were selected based on previous studies, as they are more likely to affect user perceptions on AI (Silva et al., 2023). Multidimensional indicators such as subjective trust (i.e., trust, and reliance perceptions), task performance (i.e., compliant rate and decision-making time) and physiological measures (i.e., skin conductance level and heart rate variability, indicated by RMSSD) were collected from the participants during the experiment.



Based on your physical examination results, and your water and alcohol intake in last two days, you must choose to inject higher-dose insulin.

If you do not have enough exercise or drink enough, a high dose of insult will be recommended. Now, you have enough exercise and drink enough, you must choose to inject low-dose insult.

The system makes recommendations on injection doses according to the following guidelines. Thus, you must choose to inject higher-dose insult.

Insufficient sleep time, higher dose.

Moderate amounts of water, lower dose.

The system calculates probabilities of your healthcare plans based on your physical examination results and two days' activities. You must choose to inject low-dose insult.

The probability to inject higher dose is 22%.

The probability to inject lower dose is 78%.

Figure 1: Four typical AI explanation formats (in Chinese in the experiment).

Materials and Tasks

A 23-inch dell desktop computer was used to run the experimental application, which was created to present the experimental tasks and record participants’ task performance. The ErgoLAB physiological system was used to collect data on heart rate variability and skin conductance level. Experimental tasks were adapted from typical self-care activities in chronic disease management (Van der Waa et al., 2021) and included a series of dosage adjustment tasks for typical chronic diseases. In the tasks, participants would be first presented with information on their latest physical examination results and their basic activities in the past several days, and were required to carefully read the information. After that, they were presented with AI recommendations on dosage injection from a simulated AIHDSS with varied explanation formats and expert opinions and were asked to make decisions on whether they comply with the AI recommendations or not.

Procedures

Participants were briefed on the procedures of the experiment upon their arrival, and were instructed to complete practice tasks. Then, they started the main experiment, where they were required to complete three dosage judgment tasks under each of 12 experimental conditions of explanation form and expert opinion combinations. The order of the experimental conditions was counterbalanced across the participants. After completing tasks in each experimental condition, they were asked to answer their trust and reliance perceptions towards the AI system with scales adopted from previous studies (Hoffman et al., 2019).

Statistical Analysis

The repeated measures analyses of variance (RMANOVAs) were used to assess the effects of explanation format and expert opinion on users' trust perceptions, task performance and physiological measures. Greenhouse-Geisser adjusted degree of freedom and p-values were applied for data that violated the spherical assumption. Pair-wise comparisons was performed with Bonferroni post-hoc tests. Data analysis was performed with IBM SPSS 22 with a significance level of 0.05.

RESULTS

Trust Perceptions

Both AI explanation format (F(3, 57) = 4.248, $p = 0.009$) and expert opinion (F(2, 38) = 20.549, $p < 0.001$) had significant effects on trust (Table 1). Expert opinion had a significant main effect on reliance (F(2, 38) = 18.385, $p < 0.001$). No significant interaction effects were found (Figure 2).

Task Performance

Expert opinion had a significant effect on compliance rate (F(1.430,27.164) = 30.582, $p < 0.001$) (Table 2). There was also a significant interaction between explanation format and expert opinion (F(6,114) = 2.225, $p = 0.046$) (Figure 3). In the case of expert endorsement or expert objection, there is no significant difference in the compliance rate of various explanation formats, but in the case of no expert advice, the difference of compliance rate for various explanation formats are significant. Both explanation form (F(2.124, 40.356) = 3.909, p = 0.026) and expert opinion (F(2, 38) = 13.487, p < 0.001) had significant main effects on decision-making time (Figure 3). No other significant effect was detected.

Table 1. Effects of explanation format and expert opinion on compliant rate and decision-making time.

	Compliance rate (%)				Decision-making time (s)			
	Mean	SD	F	P	Mean	SD	F	P
AI explanation format			2.615	0.060			3.909	**0.026**
Natural language	73	15			6.8^{d}	3.0		
Counterfactual reasoning	84	11			9.5^{d}	4.5		
Case-based reasoning	79	13			7.6	3.1		
Probability	78	11			7.1	3.1		
Expert opinion			30.582	**< 0.001**			13.487	**<0.001**
Support	96^{ab}	5			7.1^{c}	2.9		
Against	64^{a}	14			9.3^{ac}	3.3		
Absence	76^{b}	15			6.8^{a}	2.3		
AI explanation format× expert opinion			2.225	0.046				0.157

Notes: Values labelled with superscript letters 'a' and 'b' differed at a level of 0.001, 'c' differed at a level of 0.01, 'd' differed at a level of 0.05.

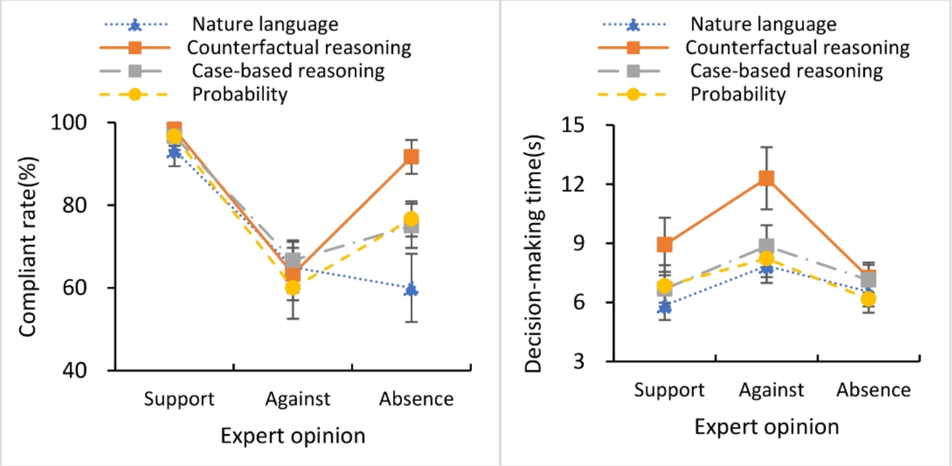


Figure 2: Task performance by explanation format and expert opinion.

Physiological Measures

Table 3 presents the effects of AI explanation format and expert opinion on physiological measures. Explanation form had a significant effect on skin conductance (F(3, 57) = 4.314, p = 0.008) and heart rate variability (F(3, 57) = 3.106, p = 0.037) (Figure 4). No other significant effect was detected.

DISCUSSION

In spite of the development and popularization of AI-assisted decision support systems in healthcare, the systems are often difficult for users to maintain a high level of trust for their recommendations due to AI's black-box attributes (Ghassemi et al., 2021). This study aims to examine the effects of explanation format and expert opinion on users' trust in recommendations from an AIHDSS. The results showed that AI explanation format and expert opinion had varied influence on users' subjective trust, task performance, and physiological measures.

Table 2. Effects of explanation formats and expert opinion on user's trust perceptions.

	Trust				Reliance			
	Mean	SD	F	P	Mean	SD	F	P
Explanation format			4.248	0.009			2.678	0.056
Natural language	5.03[a]	0.90			4.93	0.71		
Counterfactual reasoning	5.60[a]	0.57			5.48	0.80		
Case-based reasoning	5.52	0.58			5.30	0.73		
Probability	5.20	0.78			5.03	0.89		
Expert opinion			20.549	< 0.001			18.385	< 0.001
Support	5.79[bc]	0.47			5.59[bc]	0.57		
Against	5.04[b]	0.64			4.94[b]	0.69		
Absence	5.19[c]	0.71			5.03[c]	0.56		
Explanation format × expert opinion			0.464	0.834			1.807	0.104

Note: Values labelled with superscript letters 'a' differed at a level of 0.05, and values labelled with superscript letters 'b', 'c' differed at a level of 0.001.

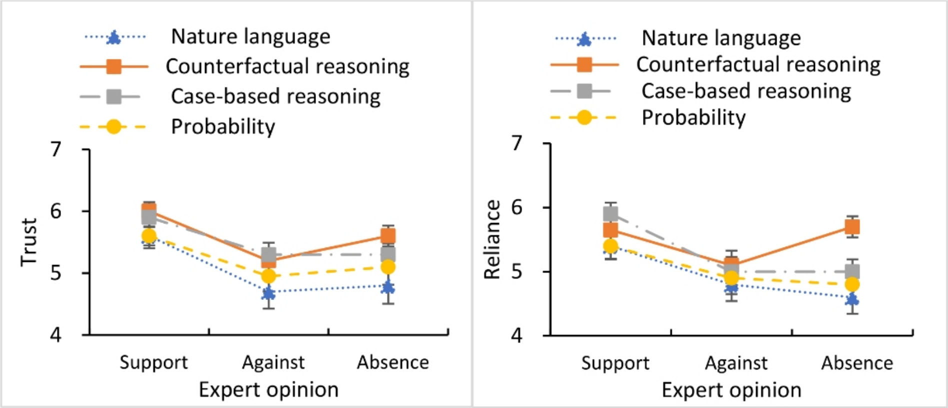


Figure 3: User trust and reliance by AI explanation format and expert opinion.

The results showed that explanation format had a significant impact on user perceived trust towards recommendations from the AIHDSS. Users had the highest level of trust on recommendations by counterfactual reasoning explanation compared with other explanation types. This result is consistent with evidence from a recent study (Silva et al., 2023), which also showed the benefits of counterfactual reasoning explanation in obtaining higher trust level. Counterfactual reasoning presents explanation information in alternatives that are contrary to what are usually expected by users and would bring more logic thinking for users (Van Hoeck et al., 2015). Particularly, in our study counterfactual reasoning presented the consequences in a way that are usually not favoured by users (e.g., assuming that the users do not adhere to self-care instructions, and thus they had to take more medication). Therefore, users might be alerted more by this explanation format, and be more likely to agree with the logic in this format. This would lead to their increased trust in AI recommendations.

Table 3. Effects of explanation format and expert opinion on physiological measures.

	Skin conductance (μs)				RMSSD (ms)			
	Mean	SD	F	P	Mean	SD	F	P
Explanation format			4.314	0.008			3.016	0.037
Natural language	4.4	3.8			43	53		
Counterfactual reasoning	4.4^{a}	3.7			43	55		
Case-based reasoning	4.5	3.5			48	47		
Probability	5.3^{a}	4.3			44	53		
Expert opinion			0.913	0.410			0.327	0.610
Support	4.6	3.8			41	57		
Against	4.6	3.7			42	49		
Absence	4.8	3.8			47	59		
AI explanation format× expert opinion			1.687	0.190			0.612	0.721

Notes: Values labelled with superscript letters ‘a’ differed significantly.

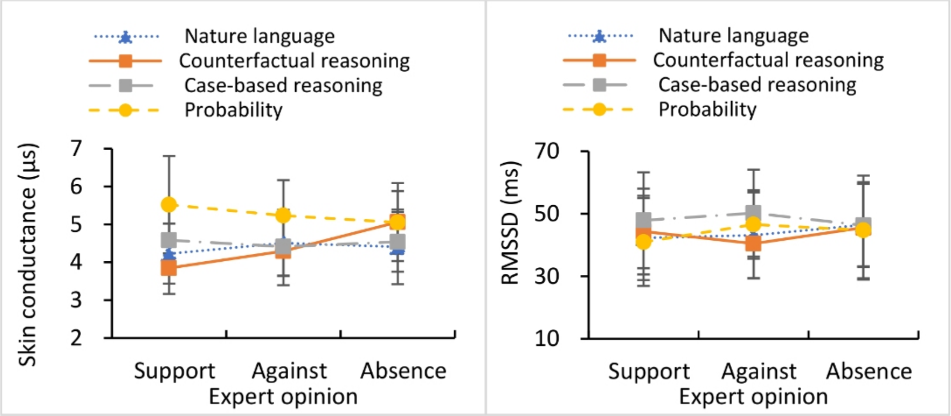


Figure 4: Skin conductance and RMSSD by explanation format and expert opinion.

While explanation format had no effect on user’s compliance rate, it had a significant effect on users’ decision-making time. Counterfactual reasoning required more time for decision-making compared with natural language explanation. The result appears to contradict findings in previous studies (Kim et al., 2020), which showed that sufficient trust could reduce users’ decision-making time. It may be that, although counterfactual reasoning obtained higher trust level, its counterfactual nature in explanation may have aroused more in-depth thinking that caused more time to respond to AI recommendations.

Explanation format also had a significant effect on users' skin conductance, which was significantly lower for counterfactual reasoning than that of probability score. This result is consistent with the study by Kahawaji et al. which showed that skin conductance is negatively correlated with interpersonal trust in smart phone- and computer-based text-chat environment (Khawaji et al., 2015). The finding could be intuitive. As users show more trust towards the AIHDSS, they would be more relax in physiological response, thereby resulting in less sweat secretion and smaller skin conductance.

Expert opinion showed significant effects on both trust perceptions and task performance. AI recommendations that were supported by medical experts yielded more higher trust and reliance levels, compared with AI recommendations against by medical experts, or no expert opinion. While previous studies suggested that human experts have similar influence with AI systems (Kim et al., 2020), our results indicated that expert endorsement could indeed reinforce the influence of AI recommendations. In addition, there is a significant difference between expert opposition and expert endorsement for users' complaint rate and decision-making time. Consistent with their trust perceptions, users would be more likely to accept AI recommendations, if the recommendations were supported by medical experts. In contrast, users took longer time for decision-making when the medical expert disagreed with the AI recommendations, reflecting that users might think carefully about whether they should follow the AI recommendation or not. Moreover, AI recommendations supported by medical experts yielded lower skin conductance compared with that of no expert opinion, though the difference was not significant.

Expert opinion also interacted with explanation formats for compliance rate. It showed that compliance rate for various explanation formats was comparable in the case of expert endorsement or expert objection, but it showed significant differences for various explanation formats in the case of no expert opinion. This may mean that the influence of explanation format could be largely mediated by expert opinion, no matter whether the opinion is in support of or against the AI recommendations (Wang et al., 2020).

The results from this study provide implications for practice and future studies. First, our study, with empirical evidence, demonstrated that AI recommendations in healthcare that are explained with different formats could elicit varied levels of user trust, as indicated by trust perceptions, task performance and physiological measures. Counterfactual reasoning could be a promising way to convey how and why AI makes healthcare recommendations. Second, expert opinion could still be powerful in influencing users' trust and decision-making, even in scenarios where AI is recognized as accurate as human experts in basic diagnosis tasks. Expert opinion could mediate the influence of explanation format. Thus, future AIHDSS developers and managers could be careful when using varied explanation formats and expert opinions in the design and implementation of AI systems to enhance the system' explainability, as users' trust towards the AI systems should be kept in appropriate levels, neither over-trust nor under-trust. Finally, our study provided promising evidence that physiological measures (such as skin conductance) could be sensitive indicators of human-AI trust, and thus could serve as objective measures of human-AI trust. Future studies could make further analysis of the physiological measures with machine learning algorithms, in an attempt to obtain more accurate assessment of human-AI trust.

CONCLUSION

This study examined the impact of explanation format and expert opinion on users' trust towards an AIHDSS, as measured by trust perceptions, task performance and physiological indicators. The results showed that explanation format and expert opinion had varied influence on users' subjective trust, task performance, and physiological measures. Users showed higher trust in AI recommendations explained by counterfactual reasoning and supported by medical experts. In addition, it appears that the influence of explanation format could be mediated by expert opinion. The findings can provide references for better design of explainable AI and for more accurate assessment of human-AI trust in AI-assisted healthcare contexts.

ACKNOWLEDGMENT

This work was partly supported by the National Natural Science Foundation of China (Grant No. 72101161 and 32271130) and the Foundation of Shenzhen Science and Technology Innovation Committee (Grant No. JCYJ20210324100014040). Tieyan Wang is the corresponding author and can be contacted at: whoiam945010295@hotmail.com.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., and Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Contreras, I., and Vehi, J. (2018). Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. *Journal of Medical Internet Research, 20*(5), e10775. doi:10.2196/10775

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745-e750.

Gunning, D., and Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program. *AI Magazine, 40*(2), 44–58. doi:10.1609/aimag.v40i2.2850

Hoffman, R., T. Mueller, S., Klein, G., and Litman, J. (2019). Metrics for Explainable AI: Challenges and Prospects. *Arxiv (Cornell University)*. doi: arXiv:1812.04608

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Mueller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery, 9*(4), e1213. doi:10.1002/widm.1312

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal, 15*, 104–116. doi:10.1016/j.csbj.2016.12.005

Khawaji, A., Zhou, J., Chen, F., and Marcus, N. (2015). *Using galvanic skin response (GSR) to measure trust and cognitive load in the text-chat environment.* Paper presented at the Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, Republic of Korea. 1989–1994. https://doi.org/10.1145/2702613.2732766

Kim, B. H., Koh, S., Huh, S., Jo, S., and Choi, S. (2020). Improved Explanatory Efficacy on Human Affect and Workload Through Interactive Process in Artificial Intelligence. *IEEE Access, 8*, 189013-189024. doi:10.1109/access.2020.3032056

Lee, J. Y., and Sundar, S. S. (2013). To tweet or to retweet? That is the question for health professionals on Twitter. *Health Communication, 28*(5), 509–524.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., and Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine, 226*, 107161.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research, 46*(4), 629–650. doi:10.1093/jcr/ucz013

Petersson, L., Larsson, I., Nygren, J. M., Nilsen, P., Neher, M., Reed, J. E., Tyskbo, D., and Svedberg, P. (2022). Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. *BMC Health Services Research, 22*(1), 850. doi:10.1186/s12913-022-08215-8

Rüping, S. (2006). *Learning Interpretable Models.* (Ph. D), University of Dortmund.

Schachner, T., Keller, R., and Wangenheim, F. V. (2020). Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review. *Journal of Medical Internet Research, 22*(9), e20701. doi:10.2196/20701

Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., and Medow, M. A. (2013). Why Do Patients Derogate Physicians Who Use a Computer-Based Diagnostic Support System? *Medical Decision Making, 33*(1), 108–118. doi:10.1177/0272989x12453501

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, 102551.

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., and Gombolay, M. (2023). Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human-Computer Interaction, 39*(7), 1390-1404. doi:10.1080/10447318.2022.2101698

Sullivan, Y., de Bourmont, M., and Dunaway, M. (2020). Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. *Annals of Operations Research, 308*, 525–548. doi:10.1007/s10479-020-03702-9

Van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, 103404. doi:10.1016/j.artint.2020.103404

Van Hoeck, N., Watson, P. D., and Barbey, A. K. (2015). Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience, 9*. doi:10.3389/fnhum.2015.00420

Wang, J., Molina, M. D., and Sundar, S. S. (2020). When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence. *Computers in Human Behavior, 107*, 106278. doi:10.1016/j.chb.2020.106278