How may I persuade you to trust AI? Promote Customized Explainable AI through Information Vividness

Completed Research

Xiaocong Cui
Georgia State University
xcui4@gsu.edu

Jung min Lee
Georgia State University
jlee469@gsu.edu

J. Po-An Hsieh
Georgia State University
jjhsieh@gsu.edu

ABSTRACT

Artificial intelligence (AI) has been criticized for its black-box nature that confuses how outputs are derived. Some have proposed that explainable artificial intelligence (XAI) can address the issue and enhance users' trust in AI. Drawing on the lens of persuasion theory, we develop a research model that depicts how explanation with vividness and user characteristics independently and jointly shape trust in AI. To test the model and associated hypotheses, we conduct an online experiment. The results suggest that individual characteristics not only directly affect trust but also moderate the relationship between explanation vividness and trust.

Keywords

Human AI interaction, explainable AI

INTRODUCTION

The increasing power of advanced AI techniques (e.g., random forest, neural networks, support vector machines) has raised the public concern of whether humans should trust AI (Herman, 2017). In response, explainable AI (XAI) was proposed as a solution to improve individual trust in AI. One of the key agenda of XAI research investigates how to maximize the explainability of AI for users with different profiles to facilitate human trust in AI.

The literature on recommendation agent (RA) and persuasion theory suggest explanation form (e.g., graphics, text, and the length of explanation) as a critical factor affecting trust (Tintarev and Masthoff, 2010). Also, IS scholars have long maintained that user characteristics like domain knowledge may affect human trust in the RA (Wang and Benbasat, 2007). The above discussion leads to the following research questions:

RQ 1: How does information vividness affect user trust in AI?

RQ 2: How do individual characteristics affect user trust in AI?

A successful AI implementation requires the synthesis of a variety of knowledge, such as knowledge about machine learning, statistics, and the target domain (e.g., real estate) in which AI is applied. Individual characteristics, such as prior knowledge, may affect how users interpret XAI explanation and hence their trust. It is therefore important to examine how users with different characteristics would respond when provided with the same explanations from XAI, leading to our third research question:

RQ 3: How do individual characteristics moderate the impact of information vividness on trust?

TRUST IN INFORMATION SYSTEMS (IS)

Trust in IS can be traced back to trust in interpersonal relationships (Wang and Benbasat, 2007). From the cognitive perspective, McKnight et al. (2002) integrated trust research in e-commerce and proposed a trust model which suggests that trusting belief leads to trusting intention, which then leads to trust. However, studying trust from a rational choice perspective is insufficient to fully describe trusting behavior (Komiak and Benbasat, 2004; Komiak and Banbasat, 2006). Therefore, adopting both emotional and cognitive trust is more objective than just considering one of them. Later, studies of the customer's trust shifted to Recommendation Agent (RA) mediated e-commerce (Komiak and Banbasat, 2006), which is closely related to AI. Different from the traditional RA that only provides recommendations to users, the explanation-based RA provides explanations about why it offers such recommendations. Therefore, the explanation-based RA increases users' trust by helping users make the right decisions (Tintarev and Masthoff, 2010). Other factors that impact trust in RA include system features - including interface display, recommender algorithms, and user-system interaction. Factors that facilitate user and agent interaction also serve for trust formation. Effective design on the interface, such as website layout, typography, font size, and colors can increase credibility and trust (Fogg et al. 2003).

MODEL DEVELOPMENT

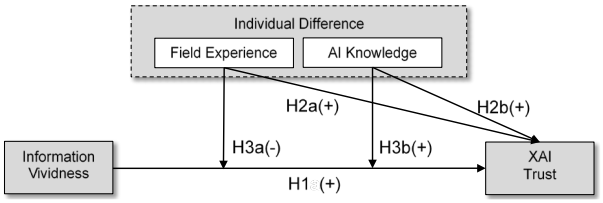


Figure 1. Research Model

XAI trust

This study is a preliminary examination of how the explanation of AI impacts a user's trust in AI. We focus on the user's trusting belief - a type of cognitive trust generated from the trustor's attributional processes when interacting with the trustee (Komiak and Benbasat 2004). The trustee in the context of our research is AI. By adopting McKnight et al. (2002)'s definition, trusting belief in XAI means that users believe the XAI has attributes that are beneficial to users.

Factors influencing XAI trust

Based on the literature review, factors influencing trust can be classified into three types: factors from the agent such as RA (e.g., the capacity of agent), factors from the user (e.g., trust disposition, knowledge learned about agent), and factors from interaction (e.g., interface design, provide explanation, explanation form). Because we focus on users' trusting belief on given XAI, this research focuses on factors from the users and factors from the interaction between the users and the XAI.

For users, the well-designed explanation can facilitate problem-solving, improve learning and performance, and lead to positive perceptions about a system (Gregor and Benbasat, 1999). Specifically, explanation content and presentation formats such as text-based explanations or explanations enhanced through images, graphics, or animation, have an impact on the user's perception of knowledge-based systems. Following the same logic, in XAI, how an explanation is presented could influence the credibility of AI. Therefore, in this study, we consider the impact of explanation presentation form on XAI trust by drawing on the concept of the vividness effect from the persuasion theory (Figure 1).

Information vividness

Vividness describes the stimulus quality of information by the extent that it attracts people's attention and stimulates people's imagination (Taylor and Thompson. 1982). It includes concrete and specific text, images, audio, and video (Taylor and Thompson. 1982). Vividness can be manipulated by using a different style of language or a different format (Kisielius and Sternthal, 1984).

Based on the vividness effect, the vivid presentation has special persuasive properties (Chang & Lee, 2010). A message can be decomposed into text arguments and vivid cues (Hafer et al., 1996). Persuasion can occur by a cognitive elaboration on arguments, or by paying attention to cues rather than arguments when a message is valid (Hafer et al., 1996). A valid message includes elaborative arguments and vivid cues that imply source expertise and "the attractiveness, likability or popularity of message source" (Hafer et al., 1996).

Based on the idea of the vividness effect, we proposed that in XAI, not only explanations of XAI plays the role of argument elaboration, but also of how we present explanations of XAI. In other words, information vividness can impact users' persuasiveness. We expect that by presenting explanations of XAI in a vivid format, users will perceive the explanation as more effective and persuasive, thereby stimulating a higher level of trust in XAI (Tintarev and Masthoff, 2010). As such, we present the following hypothesis:

H1: XAI with vivid information will positively affect trust in XAI.

Individual characteristics

According to Mueller et al. (2019), the value of explanation must consider the stance of explanation receivers. In addition to demographic characteristics, individual differences such as field-experience and knowledge about an agent are found to have an impact on trust in an agent (Wang and Benbasat, 2007).

1. The direct impact of field experience and AI knowledge on trust in XAI

Users' field experience in this paper refers to the users' experience in the implemented context of AI. According to Celsi and Olson (1988), users' field experience determines their level of involvement. Users with a high level of field experience are more involved in processing information than users with a low level of field experience. When AI information is presented to the users in a favorable way such as how AI solves the problems with high accuracy, the users are more likely to make a positive judgment about AI. Therefore, we arrive at the following hypothesis:

H2a: Field experience will positively affect trust in XAI.

Users' level of AI knowledge indicates their ability to process AI information (Celsi and Olson, 1988). When processing AI information, users with more AI knowledge, relative to those with little AI knowledge, can invoke more available AI information from memory. Therefore, users with a high level of AI knowledge will process and comprehend more AI information than users with a low level of AI knowledge (Gregor and Benbasat, 1999). Therefore, we propose the following hypothesis:

H2b: AI knowledge will positively affect trust in XAI.

Table 1. Descriptive, Internal Consistency, Convergent, and Discriminant Validity

Constructs	μ	SD	Cronbach's Alpha	CR	AVE	1	2	3	4
AI Belief[1]	3.04	0.59	N/A	N/A	N/A	N/A			
AI Knowledge	3.14	0.09	0.80	0.80	0.57	0.01 (b)	0.75 (a)		
Field experience[1]	0.30	0.46	N/A	N/A	N/A	0.28 (b)	-0.04 (b)	N/A	
Trust Competence	5.21	0.87	0.91	0.91	0.71	-0.09 (b)	0.16 (b)	0.18 (b)	0.84 (a)

a. Diagonals represent the square root of the AVE
b. Off-diagonal elements are the correlations among constructs
c. For discriminant validity, diagonal elements should be larger than off-diagonal elements
1. Single item measure
SD: Standard Deviation; CR: Composite Reliability; AVE: Average Variances Extracted

2. Moderating role of field experience and AI knowledge on trust in AI

More and more empirical research found that vivid information not only enhances persuasiveness but also undermines the persuasiveness. This conflict phenomenon can be explained by the resource matching hypothesis (Meyers-Levy and Peracchio, 1995). The hypothesis suggests that when consumers are not motivated, they will only devote a low level of available resources to process the advertisement information. In this situation, the vivid cues (such as the color in color ads compared with black-and-white ads) are more likely to draw the attention of consumers and facilitate the consumers to perceive the inherent goodness of the product. In contrast, when consumers are highly motivated, their available resources will be invoked to commensurate with resources required to process the content of ads. Therefore, consumers can make a judgment based on the validity of ad content. That is to say, users with a low level of field experience are more likely to be influenced by the vivid AI information; while users with high level of field experience are more likely to be influenced by the nonvivid AI information. Therefore, we posit that:

H3a: Field experience will negatively moderate the effect of information vividness on XAI trust.

Another factor that may influence how users interpret vivid and nonvivid AI information is their AI knowledge (Hafer et al., 1996). AI algorithms are usually presented in the forms of figures, formulas, and/or programming codes (e.g., Goodfellow et al., 2016). When processing vivid AI information (e.g., figures, formulas, programming codes), individuals who have a high level of AI knowledge will have a greater extent of cognitive elaboration. Therefore, we can arrive at the following hypotheses:

H3b: AI knowledge will positively moderate the effect of information vividness on trust in XAI.

EXPERIMENT DESIGN

Overview. To examine the validity of the research model, we designed a three-groups experiment. We used Random Forest (RF) to analyze the Boston Housing dataset and make housing price prediction. The dataset includes 506 records with 13 variables. The control group is a pure text-based description of RF; the first treatment group is a text-based description of RF with an image that describes the working process of RF, and the second treatment group is a text-based description of RF with a mathematical formula representing the cost function of RF.

Procedures. We created an online survey using Qualtrics. Data was collected via Amazon's Mechanical Turk (MTurk). We conducted a pilot study with 30 subjects to refine the experiment before the official study. The survey starts with measurements on users' knowledge about AI, followed by measurements on field-experience of the housing market. Subject were randomly assigned to each group. We then measure the subjects' trust towards the RF. We used attention checks as the rejection criteria.

Variables and measurements

According to McKnight et al. (2002), trusting belief includes competence, benevolence, and integrity. Because AI benevolence and integrity are less relevant to AI, this study focuses on competence rather than all these three aspects. The four-items measurement of competence is adapted from McKnight et al. (2002). AI knowledge is a three-items construct adapted from Rhodes et al. (2014), ranging from 1 (Not familiar at all) to 5 (Very familiar). We measured users' field experience of house buying with one item about their experience or knowledge of housing prices (Taylor and Todd, 1995). We controlled the effect of users' basic demographic characteristics (i.e., age, education level, and gender). Users' belief in AI is also controlled, which is measured as a four-items construct adapted from Rhodes et al. (2014).

RESULTS

There were 105 valid records after data cleaning, 60 males and 45 females. 90% of the subjects were between the age of 18 to 50. 42% of subjects have an educational degree lower than the undergraduate level, and 58% of subjects have an educational degree equivalent to or higher than an undergraduate degree.

We first examined the reliability and convergent and discriminant validity of the constructs (Table 1). The values of Cronbach's Alpha of both AI knowledge and Trust Competence are higher than 0.707, suggesting good internal consistency and reliability (MacKenzie et al. 2011). The values of AVE of AI Knowledge and Trust

Table 3. Results of ANCOVA (Dependent Variable: Trust Competence)

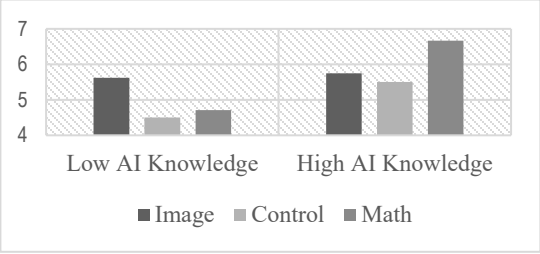
Source	df	SS	MS	F	P	Hypothesis	Supported
Information Vividness	2	0.22	0.11	0.22	p>0.1	H1	No
Field experience	1	3.26	3.26	6.58	**p<0.05**	H2a	**Yes**
AI knowledge	13	17.17	1.32	2.67	**p<0.01**	H2b	**Yes**
Information Vividness × Field experience	2	0.84	0.42	0.85	p>0.1	H3a	No
Information Vividness × AI knowledge	18	15.70	0.87	1.76	**p<0.05**	H3b	**Yes**
AI Belief	2	1.61	0.81	1.63	p>0.1		
Age	1	0.45	0.45	0.90	p>0.1		
Gender	1	1.33	1.33	2.68	p>0.1		
Degree	1	2.46	2.46	4.97	**p<0.05**		
Error	59	29.22	0.50				

Competence both exceed the required 0.5 (Fornell and Larcker, 1981). All the correlations are lower than the threshold of 0.707 (Mackenzie et al. 2011). As can be seen in Table 2, all items load higher than the recommended 0.707 on their corresponding constructs and are much higher than their cross-loadings on others.

There are 34 valid records in the control group, 39 in the image group, and 32 in the math group. The results of the ANCOVA analyses are shown in Table 3. To begin with, information vividness does not exercise a direct effect on the user's trusting belief. H1 is not supported. Next, AI knowledge has a significant positive effect on trust competence ($F(13, 59) = 2.67$, $p<0.001$), suggesting that users with more AI knowledge can better understand the competence of RF. H2a is supported. Field experience also has a significant positive impact on trust competence ($F(1, 59) = 6.58$, $p<0.01$). Users who have more housing knowledge can better understand the explanation of RF and have a stronger perception of the competence of RF. H2b is also supported.

With regard to the moderation effect, the results show that field experience does not moderate the relationship between information vividness and trust, rejecting H3a ($F(2, 59) = 6.58$, $p>0.1$). In contrast, AI knowledge has a positive and significant moderating effect on the relationship between information vividness and trust competence. H3b ($F(18, 59) = 6.58$, $p<0.05$) is, therefore supported.

Figure 2. The interaction effect between AI knowledge and information vividness



We plotted Figure 2 to further illustrate the interaction effect of the level of AI knowledge and information vividness. For users who have a low level of AI knowledge, (a) the vividness effect of the image is stronger than that of the mathematical formula, and (b) the vividness effects of both image and formula are stronger than the effect of the text. For users with a high level of AI knowledge, (a) the vividness effect of the mathematical formula is larger than that of the image, (b) the vividness effects of both image and formula are stronger than the effect of the text.

For users with either high or low AI knowledge, the vividness effect has a significant yet differential impact on their trust toward RF. In other words, they may interpret the image and mathematical formula differently. For knowledgeable users, the mathematical formula is a strong cue to invoke the vividness effect on trust because these users have the foundation to interpret the formula. In contrast, for the users with little AI knowledge, they are short of the base to understand the formula; in this case, image, relative to formula, can serve better to invoke the vividness effect on trust.

Table 2. Item loadings and cross-loadings

Factors	Items	Trust Competence	AI Knowledge
Trust Competence	Comp_3	**0.89**	-0.01
	Comp_2	**0.86**	0.05
	Comp_4	**0.82**	0.04
	Comp_1	**0.81**	0.01
AI Knowledge	AI_K2	0.04	**0.79**
	AI_K1	0.05	**0.75**
	AI_K3	-0.07	**0.74**

CONCLUSION AND DISCUSSION

Our results show that (a) users' AI knowledge and field experience both contribute to users' understanding of the explanation from XAI and enhance their trust in AI and that (b) information vividness and AI knowledge interactively affect trust in AI; that is information vividness is influential for users with a higher level of AI knowledge but not for users with little AI knowledge. Although RA and XAI are both algorithm-based agents, the insights derived from RA literature not be directly applicable to XAI research; for instance, while explanation form is a key factor in RA, we find, this factor exerts an effect on XAI trust only for users with high AI knowledge.

Like most empirical research, there are some limitations to this study. First, only the random forest algorithm is considered in this study; more algorithms should be tested in the future to determine the robustness of these findings. Also, in this study, we only consider two types of information vividness: images and mathematical formulas.

For robustness consideration, we recommend interested scholars to include other types of information vividness such as terminology, visuals, and audio cues for a more compressive assessment. Third, in this research, we only considered users' field experience and AI knowledge. More factors about individual characteristics (e.g., users' belief in AI) should be considered in the future.

This study opens the avenue to some possible future directions. First, since XAI explanation can come in a variety of formats, we urge more examination on the effect of explanation style on trust in XAI. For example, interested scholars can compare explanations with simple language versus example-based explanations. Another factor that deserves attention is the compatibility between explanation form and the underlying algorithm of XAI. For instance, the rule-based extraction method may as well be presented as a decision tree form, whereas the variable importance in the random forest may be presented as table form or histogram, to attain the desirable outcomes.

REFERENCES

1. Agarwal, R. and Karahanna, E. (2000) Time Flies when You're having Fun: Cognitive Absorption and Beliefs about Information Technology Usage, *MIS Quarterly*, 24, 4, 665-694.
2. Celsi, R L. and Olson J C. (1988), The Role of Involvement in Attention and Comprehension Processes, Journal of Consumer Research, 15 (September), pp. 210-224.
3. Chang, C.-T., and Lee, Y.-K. (2010) Effects of message framing, vividness congruency and statistical framing on responses to charity advertising, International Journal of Advertising (29:2), pp. 195–220
4. Fogg, B., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., and Brown, B. (2001) Web credibility research, CHI 01 extended abstracts on Human factors in computing systems - CHI 01
5. Fornell, C., and Larcker, D. F. (1981) Evaluating Structural Equation Models with Unobservable Variables and Measurement Error, Journal of Marketing Research (18:1), pp. 39-50.
6. Goodfellow, Bengio and Courville. 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org
7. Gregor, S., and Benbasat, I. (1999) Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice," MIS Quarterly (23:4), p. 497
8. Hafer, C. L., Reynolds, K. L., and Obertynski, M. A. (1996) Message Comprehensibility and Persuasion: Effects of Complex Language in Counterattitudinal Appeals to Laypeople, Social Cognition (14:4), pp. 317–337
9. Herman, B. (2017) The promise and peril of human evaluation for model interpretability". arXiv preprint.
10. Hoff, K. A., and Bashir, M. (2014) Trust in Automation, Human Factors: The Journal of the Human Factors and Ergonomics Society (57:3), pp. 407–434.
11. Kisielius, J., and Sternthal, B. (1984) Detecting and Explaining Vividness Effects in Attitudinal Judgments, Journal of Marketing Research (21:1), p. 54
12. Komiak, and Benbasat. (2006) The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents, MIS Quarterly (30:4), p. 941
13. Komiak, S. X., and Benbasat, I. (2004) Understanding Customer Trust in Agent-Mediated Electronic Commerce, Web-Mediated Electronic Commerce, and Traditional Commerce, Information Technology and Management (5:1/2), pp. 181–207.
14. Lee, J. D., and See, K. A. (2004) Trust in Automation: Designing for Appropriate Reliance, Human Factors: The Journal of the Human Factors and Ergonomics Society (46:1), pp. 50–80.
15. MacKenzie, S. B., Podsakoff, P. M., and Podsakoff, N. P. (2011) Construct Measurement and Validation Procedures in MIS and Behavioral Research: Integrating New and Existing Techniques, MIS Quarterly (35:2), pp 293-334
16. Mcknight, D. H., Choudhury, V., and Kacmar, C. (2002) Developing and Validating Trust Measures for e-Commerce: An Integrative Typology, Information Systems Research (13:3), pp. 334–359
17. Mueller S.T., Hoffman R.R., Clancey W, Emrey A, and Klein G. (2019) Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI." DARPA XAI Program
18. Rhodes, R. E., Rodriguez, F., and Shah, P. (2014) Explaining the alluring influence of neuroscience information on scientific reasoning., Journal of Experimental Psychology: Learning, Memory, and Cognition (40:5), pp. 1432–1440
19. Taylor, S. E., and Thompson, S. C. (1982) Stalking the elusive 'vividness' effect., Psychological Review (89:2), pp. 155–181
20. Tintarev, N., and Masthoff, J. (2010) Designing and Evaluating Explanations for Recommender Systems, Recommender Systems Handbook, pp. 479–510.
21. Wang, W., and Benbasat, I. (2008) Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for E-Commerce, Journal of Management Information Systems (24:4), pp. 249–273