Determinants of User Trust in an AI-enabled System in the Development Stage

Pi-Yang Weng
Department of Management Information Systems
National Chengchi University
Taipei, Taiwan
email: piyangong@gmail.com

***Abstract*—Explainable Artificial Intelligence (XAI) has provided a noticeable foundation for user trust building in recent years, especially in the high-risk decision scenarios, such as medical and healthcare domains. Building trust in an AI-enabled system is one of the important issues for users, which would start from the development stage. User trust could be enhanced by understanding the so-called black-box model. However, trust could be built by an emotional factor like user satisfaction in addition to scientific factors, such as XAI. In this paper, we present a framework named Three-Pillar User Trust to identify the underlying determinants of user trust in an AI-enabled system. We propose that the introduction of XAI can enhance user trust in the stages of model evaluation and validation by improving their comprehensibility with the AI system outputs and algorithms. Moreover, we propose that user satisfaction, as an emotional factor, would be an important component to influence user trust. To validate our framework, we will recruit some students from one university to participate in our experiment. This research will aim to build a three-pillar user trust framework with model interpretability, user satisfaction, and instance explainability.**

Keywords-XAI; interpretability; explainability; satisfaction; trust.

I. INTRODUCTION

In this research, the AI-enabled system users are the domain experts in healthcare domain, such as nurses or long-term care personnel, in the nursing homes. Recent studies have indicated that AI with explanations allows users to have more confidence in an AI-enabled system and have faith and trust in the algorithm results [1]. In order to obtain a better AI system output performance, domain experts are required to engage in the Machine Learning (ML) pipeline to assist in building an AI-enabled system [2]. It is also important to have domain experts kept in the loop to optimize the ML model [3]. However, ML is a complicated process, especially for deep learning. It is inevitable for domain experts to consider it as a black box even though its inputs and outputs are useful mappings. Therefore, it is essential that an AI-enabled system output is able to be explainable and comprehensible for domain experts to understand, which is instrumental to validate the quality of an AI system output [4]. During the interaction between AI engineers and domain experts in the ML pipeline, domain experts' satisfaction with the AI algorithm interpretation and its output explanation would also influence domain experts' trust in the AI system.

In Section 2, we review related concepts on XAI, Trustworthy AI, and User Satisfaction. In Section 3, we propose a conceptual model named Three-Pillar User Trust. In addition, we propose a research methodology with Hypotheses, AI Artifact, and Experiment Design to validate our framework. In Section 4, we make a preliminary conclusion for this research and propose our future work.

II. LITERATURE REVIEW

The literature review of this research will consist of three parts: Explainable AI, Trustworthy AI, and User Satisfaction.

A. Explainable AI (XAI)

Clinicians might feel uncomfortable with black-box AI, leading to recommendations that AI should be explainable in a way that clinical users can understand [5]. In the machine learning pipeline, users or domain experts are required to participate in model evaluation and system output validation to obtain high-quality training datasets [6]. XAI is a useful tool to unveil the black box and provides an explanation for each AI system output [7], which aims to explain the information behind the black-box model of deep learning that reveals how decisions are made [8]. It is necessary to explain the decision of the AI system to increase the user trust in the system. Therefore, a general model interpretability might not be sufficient for users to build their trust in an AI system. A collection of features to contribute to the output of one specific AI system would be a helpful add-on explanation to enhance user trust [9], which could be defined as instance explainability. Local Interpretable Model-Agnostic Explanations (LIME) [10], one of the XAI tools, will be used in this research.

B. Trustworthy AI

The Defense Advanced Research Project Agency (DARPA) launched a program known as Explainable Artificial Intelligence, whose motivation was to make AI systems explainable and trustworthy [11]. User trust needs to be addressed directly in all the contexts in which AI-enabled systems are being used or discussed [12]. Explainability serves as a fundamental factor that determines the user trust in AI technology [13].

Explainability could be defined as a collection of features of the interpretable domain that have contributed, for a given example, to the production of a decision [14]. To build a trustworthy AI system, a specific instance explainability would be essential for users, especially in the case that the user decision based on the AI system outputs would have a huge impact on its outcomes. (e.g., in the medical and financial domains).

C. *User Satisfaction*

User satisfaction with the explanation of AI algorithms, which is performed by AI engineers or data scientists could be defined as the degree to which users feel that they sufficiently understand the AI system or the process explained to them [14]. In addition to understanding algorithms in terms of rationality, user satisfaction, as an aspect of emotion, could be an important factor to enhance user trust in the AI system. Recent studies indicated that user interaction with AI-enabled systems would influence user satisfaction with the user-AI system interaction [15]. Therefore, the user would perceive satisfaction with the AI system during the model evaluation and validation while collaborating with AI engineers.

III. Research Methodology

AI system users would enhance their comprehensibility with the AI model by incorporating XAI into the model evaluation and validation process. Furthermore, the user comprehensibility would be composed of two components, which are model interpretability and instance explainability, serving as two pillars to support the user trust building. In addition, user satisfaction would be a significant factor in influencing user trust in the AI system. Therefore, we propose a conceptual model as our framework named Three-Pillar XAI for user trust building, as shown in Figure 1.

Based on this framework, we develop our hypotheses and experiment design as follows:

A. *Hypotheses*

It is essential that the AI system provides users with a reasonable explanation for one instance, especially in a high-risk scenario, such as healthcare domain. Therefore, we develop hypothesis H1 as follows:

H1: Users with understanding about instance explainability would lead to a higher level of trust than users with understanding about model interpretability.

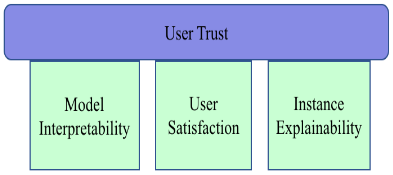

Figure 1. Three-Pillar User Trust.

It is required that domain experts need to be involved in the model evaluation and validation for high-quality training datasets and have a fundamental understanding about the AI algorithm. Then, further build their trust in the AI system. Therefore, we develop hypothesis H2 as follows:

H2: Users with understanding about model interpretability would lead to a higher level of trust than users without any understanding about both instance explainability and model interpretability.

Since user satisfaction with the explanation about the AI system or algorithm would influence his trust in the AI system, we develop hypothesis H3 and H4 as follows:

H3: Higher user satisfaction with the model interpretability would lead to a higher level of trust.

H4: Higher user satisfaction with the instance explainability would lead to a higher level of trust.

We expect that the user trust level with the understanding about instance explainability would be higher than that with the understanding about the general AI model interpretability, especially in the high-risk decision settings. The reason is that users would need to know the reason for one specific system output to ensure that their decision-making is based on logic. Also, we expect that the user trust level with understanding about AI model interpretability is higher than that without any understanding about AI model interpretability and instance explainability. The reason is that users would need to have fundamental understanding about the operational mechanism of the black-box model to build their trust in the AI system. Likewise, we expect that the user satisfaction with model interpretability or instance explainability would be higher than that without any understanding about XAI.

B. *AI Artifact*

We select the AI-enabled fall detection system as an AI artifact, which is shown in Figure 2. In this research, a mmWave radar is used to detect the moving human body in consideration of privacy, which is a camera-free device. Then, we will use a local explanation tool named LIME to show us the feature importance, such as the speed of movement at different portions of the human body, which indicates the major reason for the fall event and the possible type of fall.

We will show participants the point cloud change in shape, generated from the mmWave radar, while the human body is moving around. Also, we will simulate a fall event to have the LIME generate an output with feature importance.

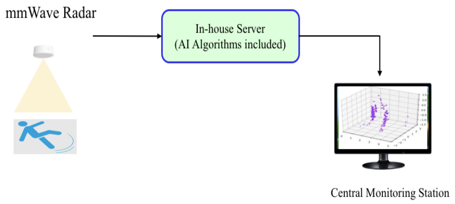

Figure 2. AI-enabled fall detection system with point clouds.

C. Experiment Design

More than 90 students from one university will participate in this experiment and will play the role of long-term care personnel. All students will be randomly divided into three groups, which are group A, group B, and group C. We will design three different courses for different groups, which are described as follows:

Group A: Participate in the model evaluation/validation with instance explainability.

The course outline includes:

- Introduction to fall detection system architecture and functions
- Introduction to model learning process (i.e., ML pipeline)
- Introduction to instance explainability (i.e., system output explanation)

Group B: Participate in the model evaluation/validation with model interpretability.

The course outline includes:

- Introduction to fall detection system architecture and functions
- Introduction to model learning process
- Introduction to model interpretability (i.e., AI algorithm)

Group C: As a control group, without any XAI. Just receive a brief introduction to this AI system, including the system architecture and functions.

IBM SPSS tool will be used for the significance analysis on trust level. In addition, we will check whether the collinearity between these three pillars is not strong, which is required to construct three-dimensional pillars to support this framework.

We design four parts of questions in the questionnaire with 5-point Likert scale, which are partially described as below:

- Model Interpretability
  I understand that the fall detection system uses an AI model, such as the KNN or SVM algorithm.
  I can understand that the change in point cloud shape indicates a certain kind of movement.
- Instance Explainability
  I realize that the AI system will output a reason to show the feature importance for each instance, such as the different moving speed at different portion of the human body.
  I can tell the difference in the human body movement by reading the different feature importance.
- User Satisfaction
  I am satisfied with the model interpretability or instance explainability.
  I think the explanation of the system output is reasonable. (For group A)
  I think the model interpretation is comprehensive. (For group B)
- User Trust
  I realize that this AI system can capture the detecting logic and produce a reasonable output.
  I can rely on the detection result of the fall detection system.
  I can trust this AI system and would like to use it as an auxiliary tool to perform my care work.

Model interpretability could be considered as the first step for domain experts to build their trust in the AI system, providing a general understanding about the AI algorithm. Instance explainability would provide the domain experts with the AI system output reasons. We would anticipate its potential application to expand to a loan application. For example, a bank financial specialist, as a domain expert, may need to know the reasons why an individual loan application will be approved or disapproved, which are generated from the AI system with the capability of instance explainability. Moreover, satisfaction with the model interpretability and instance explainability could be a sense that domain experts perceive the usefulness of the AI system, which is also an important factor for the enhancement of user trust.

IV. Conclusion and Future Work

In this work-in-progress research, we proposed a Three-Pillar User Trust framework based on reviews in the literature, which shows three pillars to support the trust level: Model Interpretability, Instance Explainability, and User Satisfaction. User trust could be built through the user satisfaction with the AI model interpretability or the instance explainability and the user comprehensibility with the AI system output reasons in addition to the user understanding with the AI model interpretability.

User satisfaction is a sense of feeling sufficient and understandable in the AI algorithm and / or system output reason, which is carried out by AI engineers. Therefore, AI engineers would face a challenge in their ability to explain an AI algorithm and the reason behind the output of the system in a way that domain experts can understand.

The introduction of XAI into the ML pipeline would trigger the interaction between domain experts and AI engineers in the collaboration of training dataset generation, model evaluation, and model validation. Moreover, it is a mutual learning process for both domain experts and AI engineers in terms of domain knowledge and ML workflows. Since the result of the model training and the output of the AI system are informed through AI engineers, we might consider it is also an interaction between domain experts and the AI system, which is a human-AI collaboration.

It is possible that this framework could be applied to another high-risk application context, such as the decision on loan application approval. Financial specialists would be highly concerned with recommendations based on the outputs of the AI system because of the huge impact on the consequences of decision making.

Our future work would include more discussions on user satisfaction influenced by the interaction of users and the AI system. Furthermore, we are also interested in constructing an evaluation model for the measurement of user satisfaction.

References

[1] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI". International Journal of Human-Computer Studies, Vol. 146, 102551, pp. 1-40, 2020.

[2] M. Maddi, H. Khorshidi, and U. Ackelin, "A review on human-AI interaction in machine learning and insights for medical applications". International Journal of Environmental Research and Public Health, Vol. 18, pp. 1-27, 2021.

[3] G. Futia and A. Vetro, "On the integration of knowledge graph into deep learning models for a more comprehensible AI: Three challenges for future research". Information, Vol. 11, No. 122, pp. 1-10, 2020.

[4] D. Pedreschi et al., "Meaningful explanations of black box AI decision systems". The Thirty-Third AAAI conference on Artificial Intelligence, pp. 9780-9784, 2019.

[5] M. Ghassemi, L. Oakden-Rayner, and A. Beam, "The false hope of current approaches to explainable artificial intelligence in health care". Lancet Digital Health, Vol. 3, No. 11, e745-e750, 2021.

[6] Google Cloud, "MLOps level 0: Manual process", Google Cloud Architecture Center, 2020. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines [retrieved: December, 2024].

[7] A. Shaban-Nejad, M. Michalowski, and D. Buckeridge, "Explainability and interpretability: Keys to deep medicine. Explainable AI in Healthcare and Medicine", Vol. 914, pp. 1-10, 2021.

[8] A. Chaddad, J. Peng, J. Xu, and A. "Bouridane, Survey of explainable AI techniques in healthcare". Sensors, Vol. 634, No. 23, pp. 1-19, 2023.

[9] M. Rebeiro, S. Signh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, 2016.

[10] D. Kaur, S. Uslu, K. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review". ACM Computing Surveys, Vol. 55, No. 2, Article 39, pp. 1-38, 2022.

[11] T. Bach, Khan, A., Hallock, H., Beltrao, G., and Sousa, S., "A systematic literature review of user trust in AI-enabled systems: An HCI perspective". International Journal of Human-Computer Interaction, 40:5, pp. 1251-1266, 2024.

[12] B. Li et al., "Trustworthy AI: From Principles to Practices". ACM Computing Surveys, Vol. 55, No. 9, Article 77, pp. 1-46, 2023.

[13] G. Montavon, W. Samek, and K. Muller, "Methods for interpreting and understanding deep neural networks". Digital Signal Processing, Vol. 73, pp. 1-15, 2017.

[14] R. Hoffman, S. Mueller, G. Klein, and J. Litman, "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance", Frontiers in Computer Science, pp. 1-15, 2023.

[15] C. Rzepka and B. Berger, "User interaction with AI-enabled system,: A systematic review of IS research". Thirty Ninth International Conference on Information Systems, San Francisco, pp. 1-18, 2018.