Human-AI Collaboration Cycle in the Development Stage of An AI-enabled System

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

Rua-Huan Tsaih
Dept. of Management Info Systems
National Chengchi University
Taipei, Taiwan
email: tsaih@mis.nccu.edu.tw

Hsin-Lu Chang
Dept. of Management of Info Systems
National Chengchi University
Taipei, Taiwan
email: hlchang@mis.nccu.edu.tw

***Abstract*—Explainable Artificial Intelligence (XAI) has garnered attention in the AI system development in recent years, especially in the high-stakes decision scenarios, such as medical and healthcare domains. In this paper, we present a framework named Human-AI Collaboration Cycle. The framework emphasizes the collaboration between domain experts and AI system in the development stage of an AI-enabled system through an introduction of XAI. We propose that the introduction of XAI can enhance domain experts' engagement in the stages of model evaluation and validation, then further review and engage in the data preprocessing, which in turn, improves their comprehensibility and trust toward the system. To validate our framework, we will conduct a field experiment in a hospital, in which nurses, as domain experts, and AI engineers will work together to develop an AI-enabled fall detection system with model explainability. We will evaluate the role of Local Interpretable Model-agnostic Explanations (LIME), one of the noted XAI tools, in the proposed Human-AI Collaboration Cycle.**

Keywords-engagement; domain expert; XAI; comprehensibility; trust.

I. Introduction

Organization for Economic Cooperation and Development (OECD) has promoted the concept of the responsible AI, suggesting that AI-related actors need to advocate human-center value, transparency, explainability, and accountability [1]. In order to obtain a better output performance, past research suggests that 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, a system developed by AI technology is not usually based on a clear statistics and probability theory. 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 necessary that machine learning and AI systems need to be explainable and comprehensible in human terms, which is instrumental for validating the quality of an AI system outputs [4]. The output of the black box needs a reasonable explanation for domain experts to trust in the AI-enabled system. In response to this issue, XAI has been more widely recognized in recent years.

It is essential that domain experts increase their trust in the AI-enabled system and further optimize the AI model and adopt it during Human-AI Collaboration (HAC). During HAC, a better output performance is expected through the domain experts' engagement in the higher quality training data generation [2]. Therefore, domain experts need to engage in the data preprocessing, such as data cleaning, data labeling, and feature selections. In recent years, AI Model Explainability has been receiving greater attention as well. However, user trust is not easy to build due to lack of transparency, especially in high-risk decision contexts, such as medical and healthcare domains [5]. We will present an XAI tool to unveil the black box to build user trust in an AI-enabled system.

This research findings will provide AI-enabled system designers with a Human-AI Collaboration Cycle framework as a guideline for developing a responsible AI system. Also, this research will highlight the importance of domain experts' engagement in the ML pipeline in the development stage of an AI-enabled system and highlight the functionality of XAI incorporated in the model evaluation/validation process, which could enhance user trust in an AI-enabled system.

In Section 2, we reviewed current concepts on Human-AI collaboration, ML pipeline, and XAI. In Section 3, we proposed a conceptual model named Human-AI Collaboration Cycle. In Section 4, we proposed a research methodology with IT Artifact, Hypotheses, and Experiment Design to validate our framework. In Section 5, we made a preliminary conclusion for this research and proposed our future work.

II. Literature Review

The literature review of this research will be composed of three parts: Human-AI Collaboration (HAC), Machine Learning Pipeline and Explainable AI (XAI).

A. Human-AI Collaboration (HAC)

Human experts and AI have different yet complementary capabilities by which they can work together to have an effective decision-making [6]. AI is not just a tool; it may become a teammate to enhance team performance [7]. Humans and AI can have mutual learning through which AI or algorithms can learn from humans and humans can acquire insights from AI or algorithms [8]. ML needs methods that engage domain experts directly into the ML process and have them in the loop until the desired results are received [2]. After building an AI

model, data scientists often need to find a domain expert to help interpret the test results and validate whether they make sense or not [9].

Humans and AI can work together as a symbiotic system through which humans can gain intelligence augmentation and AI can learn from humans' feedback through interactions [10]. AI system designers could consider a human-AI team building based on the core competencies brought in by humans and the core capabilities of the AI teammates [7]. Therefore, it is required that domain experts need to collaborate with AI systems through AI engineers in the development stage of an AI system in order to obtain a better system performance.

B. Machine Learning Pipeline

The ML pipeline starts with data extraction and analysis and then obtains a trained model after model evaluation and validation [11]. The pipeline is shown in Figure 1.

The key tasks for each ML step are described as follows:

- Data extraction and analysis
  Select and integrate the relevant data from various data sources for the ML task. Also, identify the data preparation and feature engineering that are needed for the model.
- Data preparation
  It involves data cleaning and data splitting into training data and test data.
- Model training
  The data scientist implements different algorithms with the prepared data to train various ML models.The output of this step is a trained model.
- Model evaluation and validation
  The model is validated on a holdout set to evaluate the model quality. The output of this step is a set of metrics to assess the quality of the model.

The domain experts need to join the training data labeling task, in the case of supervised learning, for obtaining high-quality training datasets and avoiding garbage in, garbage out results [12]. Also, they are required to engage in the model evaluation and validation [2]. Before a trained model is accepted by the domain experts, usually the system users, they are required to stay in the ML pipeline, especially in the stages of data extraction and analysis and model evaluation back and forth.

C. Explainable AI (XAI)

XAI is a useful tool to unveil the ML black box and provides an explanation for each AI system output [13]. XAI is especially instrumental in medicine and healthcare to ensure that the AI system outputs produced by the AI model are correct and justifiable [14]. It is necessary to explain the AI system's decision to increase the users' trust in the system. If AI system users can clearly understand the particular reasons for each system output, they will tend to trust in the AI system [15].

As domain experts have more understanding on the AI algorithm and the explanation for each system output, they tend to provide more feedback on the data preprocessing, such as data extraction and feature selections, further engaging in the ML pipeline and, hence, contributing their domain knowledge into the AI model refinement. Therefore, with XAI incorporated, domain experts will stay in the loop of ML pipeline until an acceptable AI model is achieved.

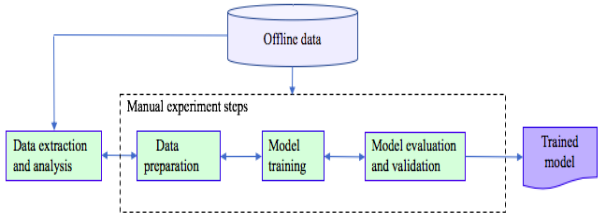

Figure 1. Machine learning pipeline.

III. Conceptual Model

As domain experts are the key AI-enabled system users, the HAC requires domain experts' engagement in the development stage for building a high-quality training dataset and achieving a better model for deployment. Moreover, AI system users will enhance their comprehensibility with the AI model by incorporating XAI into the model evaluation and validation process [16]. With this comprehensibility, AI system users will have greater trust in the AI system and, therefore, adopt the system.

It is possible that the newly trained model would fail in the next few tests before deployment. One of the possible reasons is that the model does not cover some real-world cases, which may be attributed to the introduction of XAI. With XAI, it could facilitate domain experts' engagement in the model evaluation and validation with new test data. Hence, the domain experts will need to re-engage in the ML pipeline for the training dataset review. Therefore, it constructs a cycle in the HAC. We coin it as Human-AI Collaboration Cycle, which is shown in Figure 2.

There are four components in the HAC cycle:

- Data Engagement
  In this research, Data Engagement refers to domain experts' engagement in the data preprocessing including data extraction, data cleaning, data labeling, and feature selections. With domain experts' engagement, the training data would have higher quality to train a better model. Therefore, data engagement in ML pipeline implies that domain experts would have partial responsibility for a better trained model.
- User Comprehensibility
  XAI is a technology tool to unveil the black box, which could help domain experts comprehend the model and algorithm logic. LIME, as one of XAI tools, will present some key features for each instance, i.e., each input [13]. Its output format is shown in Figure 3.

  In order to measure the speed of human movement, we use a human skeleton marked with four key coordinates [17], as shown in Figure 4. Point 1 to Point 4 represents

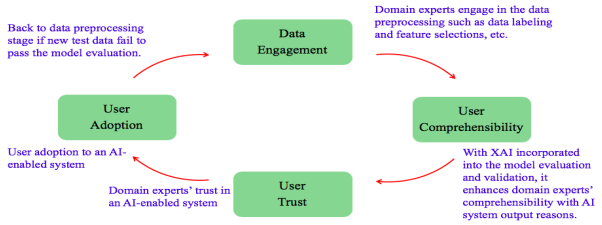

Figure 2. Human-AI collaboration cycle.

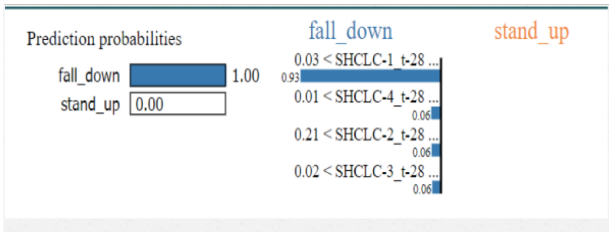

Figure 3. LIME output format for an AI-enabled fall detection system.

the central coordinate for shoulders, hips, knees, and ankles respectively. With the measurement of the Speed of Human Center Line Coordinate (SHCLC), we could identify the different kind of falls.

Figure 5 shows one kind of fall with higher speed on the movement of point 1 (i.e., SHCLC-1 on Figure 3). Therefore, the LIME outputs may provide us with the key features and reasons why the AI system judges the fall event. Also, it will indicate the specific kind of fall, such as fall over, fall down, and fall off, etc.

With domain experts' comprehensibility with the model, their trust in the AI-enabled system could be enhanced.

- User Trust

  In this research, domain experts are the key users. It is a mutual learning process during the interaction between domain experts and the ML pipeline; domain experts input their domain knowledge into the data preprocessing to confirm the training data quality for building a better model; the AI-enabled system provides insights by its data-driven analytical capabilities.

  In addition to XAI tool incorporated into the ML pipeline, domain experts provide more valuable feedback into the model refinement and training data revision through the interaction with the ML pipeline, which also help increase the user trust. An important path leading to better adoption rates identified is trust-building [18].

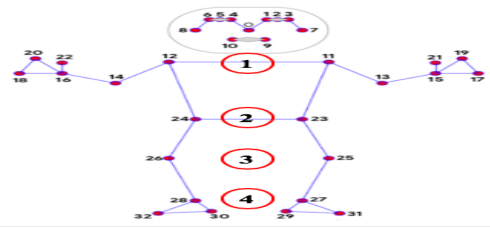

Figure 4. Human skeleton with four key coordinates on center line.

Figure 5. Different movement speed on different portion of a human while falling.

- User Adoption

  In addition to trust, Technology Acceptance Model (TAM) [19] provides us with a guideline to follow in achieving user adoption on the AI-enabled system. With TAM, the usefulness and ease of use perceived are important principles for the AI-enabled system design.

The HAC cycle starts with data engagement and then guides AI system designers and domain experts to go through the cycle to achieve a better model for deployment.

IV. Research Methodology

In this research, we design a field experiment to validate the framework of the HAC cycle. The IT artifact, hypotheses, and experiment design are described as follows:

A. IT Artifact

We select the AI-enabled fall detection system as an IT artifact, which is shown in Figure 6. There are various fall detection methods, including wearable devices with threshold setting, non-wearable device like mmWave radar and vision-based video camera. Each fall detection sensor has its advantages and disadvantages. With non-wearable sensors, people do not need to attach them on their bodies. However, they can not be used outdoors and are limited to a small area inside the detection range. However, in this research, the vision-based fall detection system is applicable for the indoor use.

B. Hypotheses

The hypotheses on domain experts' trust level are shown in Figure 7. We proposed three hypotheses(H1, H2, and H3) as follows for this research:

H1: AI-enabled system users participating in data preprocessing and model evaluation/validation with XAI incorporated would have higher trust level than users participating in data preprocessing and model evaluation/validation but without XAI incorporated.

H2: AI-enabled system users participating in data preprocessing and model evaluation/validation but without XAI incorporated would have higher trust level than users without participating in data preprocessing and model evaluation/validation, also without XAI incorporated.

H3: AI-enabled system users participating in data preprocessing and model evaluation/validation with XAI incorporated would have higher trust level than users without participating in data preprocessing and model evaluation/validation, also without XAI incorporated.

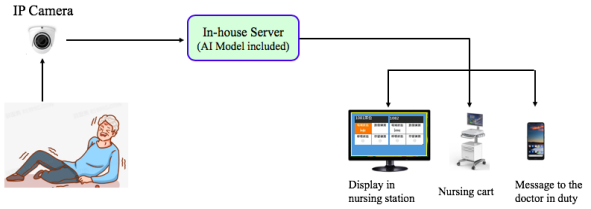

Figure 6. AI-enabled fall detection system.

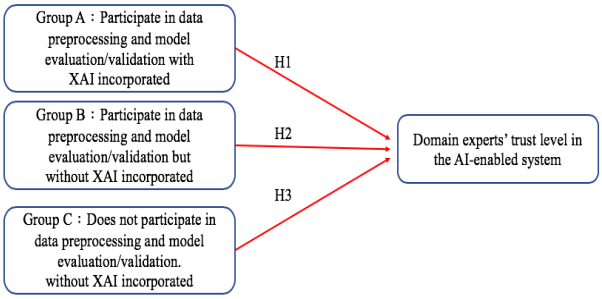

Figure 7. Hypotheses on trust level.

C. Experiment Design

More than 90 nurses, as domain experts, from one local hospital will participate in this experiment. Since user trust is one of the key components in the HAC cycle, in this research, firstly we conduct the significance level test on it.

All nurses will be divided into three groups, which are group A, B, and C. Each group will watch the same video demonstrating the brief introduction to the AI-enabled fall detection system. We designed different interaction modes with the AI-enabled system for each group, which are described as follows:

Group A: Participate in data preprocessing and model evaluation/validation with XAI incorporated.

Group B: Participate in data preprocessing and model evaluation/validation but without XAI incorporated.

Group C: As a control group, without HAC and XAI, to be informed of the system performance only.

The Likert scale will be used for trust level evaluation. The items are rated on a bipolar scale going from "I agree strongly" to "I disagree strongly", which are modified from [20]. Questions are as follows:

- I have confidence in the AI system performance.
- The AI system performance could be improved gradually.
- The output of the AI system is very predictable.
- The AI system is very reliable.
- The AI system is easy to use.
- The AI system is very efficient.
- The AI system can act as part of my team.
- I like to use the AI system.

ANOVA tool will be used for the significance analysis on trust level between groups. The timings of pre-test and post-test for each group are illustrated in Figure 8.

The nurses in Group A will be expected to have a better understanding on the key reasons of fall event identified by

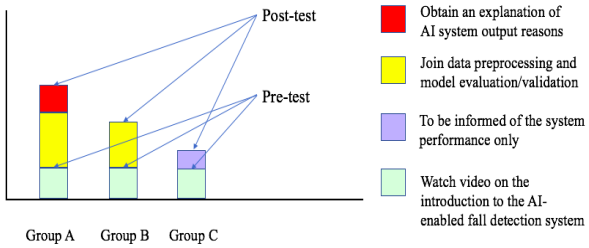

Figure 8. The timings of pre-test and post-test for each group.

the AI system. Therefore, they would have higher confidence in gradual improvement of the AI system in the development stage.

Pilot test with 9 nurses, 3 in each group, and manipulation check will be conducted to ensure the effectiveness of XAI treatment. Our basic assumption is that most nurses are rational with respect to the interpretation, provided by AI engineers, on the LIME outputs, i.e., key features.

In addition to the quantitative analysis, we will observe the differences in their interaction modes with the AI-enabled system in each group and make a complete record for qualitative analysis. For example, we have interest in the nurses' feedback or response to the XAI output explanation for one specific instance, which may encourage their engagement with the training data and test data review to assist in a better model building.

In the event that the significance level shows that nurses in Group A have the highest trust level by the introduction of XAI, the HAC cycle could be constructed with the user comprehensibility with the AI model and user adoption to the AI system. Also, a few more new test data, attributed to more data engagement, provided by the nurses in Group A would guide them to go into the second cycle for building a better model, especially in the development stage. The implementation of the HAC cycle might be considered as an approach to differentiate the user trust levels among the three groups.

V. Conclusion and Future Work

In this research, we proposed HAC Cycle based on the literature reviews, which includes four components: Data Engagement, User Comprehensibility, User Trust, and User Adoption. Also, the ML black box could be unveiled by LIME, an XAI tool, which provides the AI system users with an explanation for each instance. Hence, user trust could be built through user comprehensibility with the AI system output reasons and user adoption could be achieved under TAM.

The HAC Cycle might be considered as an approach to differentiate the user trust levels. Also, the AI model could be optimized by the implementation of this cycle with a few runs in the development stage of an AI-enabled system.

However, user comprehensibility is not limited to the user's understanding with the reasons for one specific AI system

output, which could be considered a scientific factor. User comprehensibility could also be enhanced by the model or algorithm explanation done by AI engineers. In this case, the emotional factor would be involved in the user comprehensibility. Therefore, we would propose that user comprehensibility might need to be split into two sub-components, which are AI model interpretability done by AI engineers and AI system output explainability done by XAI. The former is related to an emotional factor and the latter is related to a scientific factor. Hence, we may need both Group A1 and Group A2 to explore the differences in user trust level affected by different factors mentioned above.

ACKNOWLEDGMENT

The authors thank Alex Hsu, an AI engineer, for his assistance in the operation of XAI tool.

REFERENCES

[1] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL 0449 G, 2022. https://www.oecd.org/digital/artificial-intelligence [retrieved: February, 2024].

[2] M. Maadi, H. A. 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] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, pp. 1-11, 2020.

[6] M. H. Jarrahi, Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Kelley School of Business, Indiana University, BUSHOR-1478, pp.1-10, 2018.

[7] I. Seeber et al., Machines as teammates: A research agenda on AI in team collaboration. Information and Management, Vol. 57, pp. 1-22, 2020.

[8] M. J. Saenz, E. R. Revilla, and C. Simon, Designing AI systems with human-machine teams. MIT Sloan Management Review, Reprint 61430, 2020.

[9] D. Wang et al., Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. Proceedings of the ACM on Human-Computer Interaction, Vol. 3, Issue CSCW, Article 211, pp. 1-24, 2019.

[10] L. Zhou, S. Paul, H. Demirkan, L. Yuan, and J. Spohrer, Intelligence augmentation: Towards building human-machine symbiotic relationship. AIS Transactions on Human-Computer Interaction, Vol. 13, Issue 2, pp. 243-264, 2021.

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

[12] R. S. Geiger et al., "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data ? Quantitative Science Studies, Vol. 2, No. 2, pp. 1-42, 2021.

[13] M. T. 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.

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

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

[16] 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.

[17] S. Jeong, S. Kang, and I. Chun, Human-skeleton based fall-detection method using LSTM for manufacturing industries. 34th International Technical Conference on Circuit/Systems, Computers and Communications, JeJu, Korea(South), pp. 1-4, 2019.

[18] P. Bedue and A. Fritzsche, Can we trust AI ? An empirical investigation of trust requirements and guide to successful AI adoption. Journal of Enterprise Information Management, Vol. 35, No. 2, pp. 530-549, 2022.

[19] F. D. Davis, A technology acceptance model for empirically testing new end-user information systems: Theory and results (Doctoral dissertation, Massachusetts Institute of Technology), pp. 1-291, 1985.

[20] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, Measuring trust in the XAI context. Technical Report, DARPA Explainable AI Program, pp. 1-26, 2018.