AI Explanation Understanding from User's Perspective

in Healthcare Application

Retno Larasati
retno.larasati@open.ac.uk
Knowledge Media Institute

Supervisors name/s: Dr Anna De Liddo, Prof Enrico Motta
Starting date: 01/10/2018 (Full-Time)

I. Introduction

With Artificial Intelligence (AI) systems being implemented everywhere, including in healthcare, the interaction with AI embedded system is inevitable. However, modern AI algorithms are complex and difficult to understand. In critical decisions that involves individual's well-being, it is important to know the reasons behind such a critical decision. We assumed that in healthcare application, such as disease diagnosis or risk assessment, users want to know about the AI reasoning. Explainable AI (XAI) is an emerging research area that focus on providing a layer of explanation which helps end users to make sense of AI results [1]. There are different types of user that possibly interacting with AI healthcare application: Medical Professional (doctor), AI/Machine Learning(ML) Experts (system developer), Laypeople (patient).

When it comes to human interaction, trust is one of the important factors influencing the adoption of AI systems. AI systems in healthcare are expected to help diagnose diseases, to develop new medicine, to gain better insights into treatments and preventions that could benefit all of society. Developing trust is particularly crucial in healthcare because it involves an element of uncertainty and risk for the vulnerable patient [2]. It is still unclear what explainable AI approaches are available and applicable to healthcare and what are the factors that affect non-expert users to make trust judgments towards AI healthcare application. This is important because regulators already legislate for mandatory explanation to the data subjects, whom in the healthcare case, are the patients, while AI experts are still investigating how these explanations can be designed, and what impact they have on users' trust.

Additionally, while human explanation is already an important component in healthcare practices, explanation with AI in the loop is not common practice. When a medical professional uses AI to support their diagnosis, or in the cases of AI based self-managed health systems, the diagnosis explanation needs to be passed indirectly or directly to the patients, accounting for the AI system in the loop. The modalities and styles of explanation to improve trust judgments about AI medical support systems need to be explored, especially when the AI system targets non-experts. Finally, the lack of meaningful and usable user interfaces for XAI makes it hard to effectively assess the impact of explanation on end-users, which furthers hinder our capability to test hypothesis and advance our understanding of this complex research field.

The main goal of my PhD research is to study and understand the role of explanation in affecting non-expert user's trust judgment. Particular attention will be paid to the design of explanation interfaces and interactions that can effectively enable non-expert users to moderate their trust judgments. We will, design explanations that build on users' understanding and desiderata of trustworthy explanations. We will then test to what extent our newly proposed human-AI explanation interactions affect non-expert users' trust judgment towards AI medical support systems. In order to designing explanations that build on user's understanding, we must examine how users perceive AI given explanation. In this abstract, we are trying to investigate how users, especially non-expert users, understand an explanation given by an AI application, and recognise possible room of improvement.

II. Understandable Explanation

Abdul et al. [3] argues that previous research in Explainable AI is not strongly informed by Cognitive Psychology in terms of how humans can interpret the explanations, and did not evaluate the explanations with real users in interactive applications. There are several characteristics to be considered towards designing an understandable explanation for AI system, based on philosophy and psychology perspective, as shown in Table II.

Characteristics	Description	Ref
contrastive	the cause of something relative to some other thing in contrast	[4], [5], [6]
domain/role dependent	pragmatic and relative to the background context	[5], [7]
generalisable	simpler and broad explanation is preferable	[8] [9]
social/interactive	people explain to transfer knowledge, thus can be a social exchange	[5] [6]
truthful	how truthful each elements in an explanation is with respect to the underlying system	[10]
thorough	describes all of the underlying system	[10]

TABLE I. Characteristics of Understandable/Meaningful Explanation

III. Method

We did a group discussions with twelve participants, 2 of them are practising doctors in the UK, 2 are trained doctors but not currently practising, the rest have no professional medical training / laypeople. Examples of explanation given by AI

Systems were given to them, before we asked questions to the participants. The examples are from applications;

- Skin Vision, a skin cancer detection/risk assessment mobile application.
- Lunit, a breast cancer detection application using mammography image.
- IBreastExam, also a breast cancer early detection application. IBreastExam use its own portable device to examine the area.

Then, the participants were asked if they understand the given explanations, and if they have any comment about the explanations. The discussion was last for an hour.

The audio recording taken from the group discussion was transcribed, and any personally identifying information was redacted. Grounded theory analysis was undertaken to explore rather than to impose participants' comments on explanation [11].

IV. Results

The group discussion resulted in three major themes: the lack of explanation detail; the lack of system's information; and the lack of understanding. These themes occurred across three explanation examples, although some comments were particularly around certain explanation examples, which are noted below.

The Lack of Explanation Detail

An initial reaction from one of the participants was question about how the system got the results, after they read the explanation examples. When we asked about their understanding towards the explanation, some of the responses are asking about the detail.

"I understand that they are looking at the pattern, so I want to know what kind of pattern. Is it the tone colour or what?"[PA]

"It only stated irregular pattern but what pattern? Not really clear."[PB]

The responses are in the form of questions, which shows participants lack of understanding and willingness to know more. Those questions above could be addressed with Thorough Characteristic (See Table II). More comments about the lack of detail, that could be addressed by giving explanation with Truthful and Contrastive characteristic, respectively.

"They don't say the evidence on why they come up with this statement."[PF]

"Or maybe how the image looks like if it is normal, how the image looks if it's not."[PG]

The Lack of System's Information

Participants also raised comments about some of the system's information, for example, they were asking about the data used by the system. Here is one of the conversation between participants.

"AI has a big database, so they recognised the pattern" [PC] "but what kind of data?" [PA]

The same participant noted the needs of detail on system's information, including how big the data the system used.

"Like if they got high risk result, then maybe they want to know how big the data they used is, if it's a million images or hundreds of images. Or some extra information about the system. [PC]

The Lack of Understanding

When first asked about participants understanding ("Do you understand the explanation?"), only one participant responded. In the attempt to simulate the discussion, similar question was asked: "So do you think the texts are ideal? How about the image? Do you understand it?". Two participants responded that they don't understand it [PD][PE]. It followed by the discussion explained above.

V. Discussion

This study has several limitations that should be noted. We only conducted one short group discussion with participants recruited based on the availability. We did not screen the participants and during the discussion we found out, only three participants are familiar with AI. Because of that, some of the discussion time was spent to explain about AI and explainable AI. This unfamiliarity also resulted in the indifference towards AI and AI explanation, which resulted in disengagement and not a stimulating discussion.

By the end of group discussion, they conclude that users don't need an explanation from an AI system, they want an expert opinion directly. We will put this conclusion as a consideration in our future work. For the future works, we are currently in the middle of explanation design phase, which includes finalising explanation user's mental model, iterative design, and explanation evaluation.

References

[1] D. Gunning, "Explainable artificial intelligence (xai)," 2017.

[2] A. Alaszewski, "Risk, trust and health," 2003.

[3] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 582.

[4] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, 1990.

[5] D. J. Hilton, "Conversational processes and causal explanation." *Psychological Bulletin*, vol. 107, no. 1, p. 65, 1990.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2018.

[7] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press, 2006.

[8] S. J. Read and A. Marcus-Newhall, "Explanatory coherence in social explanations: A parallel distributed processing account." *Journal of Personality and Social Psychology*, vol. 65, no. 3, p. 429, 1993.

[9] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[10] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 2013, pp. 3–10.

[11] B. G. Glaser and A. L. Strauss, *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 1967.