Interaction Between Human And Explanation in Explainable AI System For Cancer Detection and Preliminary Diagnosis

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

Nowadays, Artificial Intelligence (AI) systems are everywhere and AI helps to make decisions for us is our daily occurrence. AI provides for us from recommendations product on Amazon and video recommendations on YouTube, to tailored advertisements on Google search result pages. Even though they appear powerful in terms of results and predictions, AI algorithms suffer from transparency problem. Modern AI algorithms are complex and difficult to get the reasoning and the insight into AI algorithms work mechanism. However, in critical decisions that involves individuals well-being such as disease diagnosis or prognosis, it is important to know the reasons behind such a critical decision. An emerging research area called Explainable AI (XAI) looks at how to solve this problem by providing a layer of explanation which helps end users to make sense of AI results. The overall assumption behind XAI research is that explicability can improve trust and social acceptability of AI assisted predictions. In our research, we specifically look at cancer detection and diagnosis and hypothesize that appropriately designed Explainable AI systems can improve trust in AI assisted medical predictions.

II. Trust, Interaction and Explainable AI

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 and to gain better insights into treatments and prevention 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 [1]. The UK government issued a policy paper that declared its vision for AI to ”transform the prevention, early diagnosis and treatment of chronic diseases by 2030” [2]. However, many doctors are still skeptical about the AI healthcare system. Study found that among the 30% of clinicians respondent lack trust in AI [3]. Not only doctors, 61% general public correspondents in the UK are unwilling to engage with AI for their healthcare needs [4]. The lack of explainability, transparency, and human understanding of how AI works, are several reasons why people have little trust in AI healthcare system. Transparency [5] and understandability [6] would help to enhance trust in AI systems. According to the Defense Advanced Research Projects Agency (DARPA), Explainable AI is essential to enable human users to understand and appropriately trust a machine learning system [7]. Some of the previous studies shows that explanations improves trust, however the characteristics of explanation have not been explored. This lead us to our research questions.

III. Research Questions

The following are the research questions:

RQ1: Does explanation improve trust in AI healthcare system/application?

With the subquestions:
RQ 1.1: what are the factors that affect trust in human-AI interaction?
RQ 1.2: what are the characteristics of explanation that is meaningful and acceptable?
RQ 1.3: what are the relation between trust and explanation?

RQ2: What modality and style of interactions of explanation need to be presented?

IV. Current Work

To address those questions, we plan to collect and analyse both qualitative and quantitative data via two main methods: an online survey and a focus group. The online survey and the focus group has been designed. The online survey is structured in 3 main sections: a set of baseline questions; a two-pages dramatizing vignette, and a post-vignette set of questions. The focus group will be carried out to gather additional qualitative data, still with the usage of dramatizing vignette.

The aim of the vignette is to elicit reflections on a fictitious scenario in which AI assisted health assessment is possible and accessible to everyone for preliminary cancer diagnosis. Dramatizing Vignette is a research method based on design fiction, and it is appropriate to elicit users feedback on the implications of possible futures yet to be realised. The dramatization of the scenario is by design, and it aims at stretching peoples thinking toward opposing views, contested actions or unexpected consequences. The aim is to trigger participants critical thinking on the situation to be analysed, before personal feedback and opinions are elicited. The questions for online survey and focus group are based on the following framework.

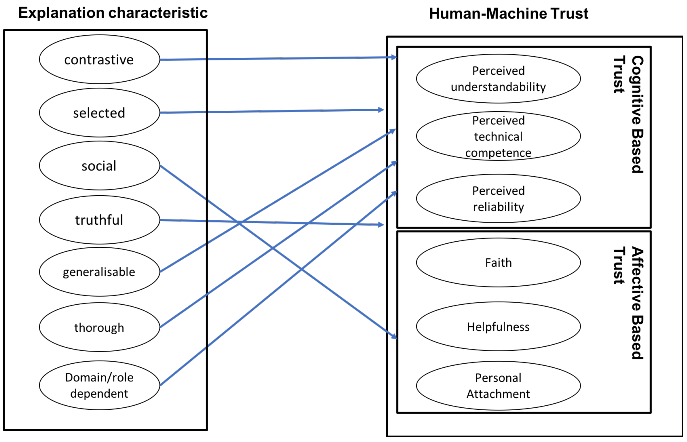

Fig. 1. Proposed Trust-Explanation framework

A. *Framework for interpreting explicability and trust*

At our current state, we have 7 characteristics of user-friendly explanations. First, explanations are contranstive. People usually ask for explanation of why a certain prediction was made instead of another prediction [8][9]. Second, explanations are selected. People usually select one or two causes from a variety of possible causes as the explanation [10]. User can choose based on their domain knowledge and cognitive ability. Third, explanations are social. The process of explaining something in order to transfer knowledge is a social exchange [10][11]. This is particularly relevant with healthcare, because some of the factors which encourage patient trust are information sharing, and their confidence in patient's ability to manage their illness [12]. Forth, explanations are truthful. This characteristic is greatly related with overall trust. User expect a robust and truthful explanation [13]. Fifth, explanations are general. People usually prefer simpler and more general explanations[14]. Sixth, explanations are thorough. Thorough explanation would cover necessity and sufficiency causes, which are strong criteria for preferred explanatory [8]. It also shows machine expertise. Seventh, explanations are domain or role dependent. Malle [15] considers a good explanation must have a pragmatic goal and role dependant [16], because explanations for medical professional is probably different than explanations for general public.

We conceptualised a general framework for trustworthy Explainable AI in healthcare. It consist of two components: explanation characteristic and human-machine trust (see: Fig. 1). Human Machine trust here is divided by two types of trust, cognitive based trust and affect based trust. The explanation characteristics are based on the items mentioned in the previous paragraph. The human-machine trust items are based on several literature about human-computer and human-machine trust [17] [18] [19]. From the total 23 trust items, we merged items that are overlapped and removed items that are not relevant. Contrastive, generalisable, thorough and domain/role dependant characteristics are related with understanding, we hypothesize the correlation with cognitive-based trust. We also speculate selected and trustful characteristics correlation with both cognitive and affect based trust. Lastly, social characteristic is correlated with affect-based trust. However, the relationships described above have yet to be investigated.

V. Future Work

We are planning to undertake a data collection and analysis mentioned above to investigate the relation between explanation and trust in healthcare, validate the items inside the framework, and gain insights about the challenges and the opportunities on developing a trustworthy explainable AI in healthcare.

References

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

[2] GOV.UK, "The future of healthcare: our vision for digital, data and technology in health and care - gov.uk," (Accessed on 02/10/2019).

[3] Intel, "U.s. healthcare leaders expect widespread adoption of artificial intelligence by 2023 — intel newsroom," 2018, (Accessed on 02/10/2019).

[4] "Survey results: Why ai and robotics will define new health: Publications: Healthcare: Industries: Pwc," 2016, (Accessed on 02/10/2019).

[5] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[6] Z. C. Lipton, "The doctor just won't accept that!" *arXiv preprint arXiv:1711.08037*, 2017.

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

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

[9] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proceedings of the 11th international conference on Ubiquitous computing*. ACM, 2009, pp. 195–204.

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

[11] D. Hilton, "Social attribution and explanation," in *The Oxford Handbook of Causal Reasoning*, 2017.

[12] M. R. Dibben* and M. Lean, "Achieving compliance in chronic illness management: illustrations of trust relationships between physicians and nutrition clinic patients," *Health, Risk & Society*, vol. 5, no. 3, pp. 241–258, 2003.

[13] T. Lombrozo, "Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions," *Cognitive Psychology*, vol. 61, no. 4, pp. 303–332, 2010.

[14] 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.

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

[16] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? a role-based model for analyzing interpretable machine learning systems," *arXiv preprint arXiv:1806.07552*, 2018.

[17] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th australasian conference on information systems*, vol. 53. Citeseer, 2000, pp. 6–8.

[18] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on Management Information Systems (TMIS)*, vol. 2, no. 2, p. 12, 2011.

[19] Z. Yan, R. Kantola, and P. Zhang, "A research model for human-computer trust interaction," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*. IEEE, 2011, pp. 274–281.