Building a Trustworthy Explainable AI in Healthcare

Retno Larasati and Anna DeLiddo

Knowledge Media Institute, The Open University, UK
retno.larasati@open.ac.uk

Abstract. The lack of clarity on how the most advanced AI algorithms do what they do creates serious concerns as to the accountability, trust and social acceptability of AI technologies. These concerns become even bigger when people's well being is at stake, sucah as healthcare. This calls for systems enabling to make decisions transparent, understandable and explainable for users. This paper briefly discusses the trust in AI healthcare system, propose a framework relation between trust and characteristics of explanation, and possible future studies to build trustworthy Explainable AI.

Keywords: Trust · Explainable AI · AI Healthcare

1 Introduction

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]. How do we get to trust an AI system in such sensitive contexts in which people's health is at stake? What are the factors that affect people's trust in AI healthcare systems? And what does a good explanation looks like? In this paper we discuss the importance of trust in AI healthcare systems, describe some key factors that influencing user friendly explanations, and propose a framework to explore the relationships between trust and explicability. We conclude by indicating trajectories for future studies.

2 Background and Motivation

2.1 Trust in AI Healthcare System

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" [1]. However, many doctors are still skeptical about the AI healthcare system. Study found that among the 30% of clinicians respondent lack trust in AI [2]. Not only doctors, 61% general public correspondents in the UK are unwilling to engage with AI for their healthcare needs [3]. 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 [7] and understandability [10] would help to enhance trust in AI systems.

2.2 Trust and Interaction in Healthcare

Trust is the foundation of relationships and is important to build a better relationship between medical professional and patient. Some of the factors in trusting a medical professional are their care and concern for the patient as an individual, and the confidence in a patient's ability to manage their disease [4][16]. Being viewed as competent by a medical professional also increased patient trust [15]. Some other factors which encourage patient trust are the clinician's technical competence, information sharing, and their confidence in patient's ability to manage their illness [2].

2.3 Explainable AI and Trust

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 [3]. 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; what kind of explanation is needed for users to trust the healthcare intelligent system?

3 Framework for interpreting explicability and trust in healthcare

At our current state, we have 6 characteristics of meaningful explanations. First, explanations are **contranstive**. People usually ask for explanation as the cause of something relative to some other thing in contrast [9] [6]. Second, explanations are **domain or role dependent**. People usually select one or two causes from a variety of possible causes as the explanations [6]. People select the causes based on their domain knowledge and cognitive ability [12]. The process of explaining something in order to transfer knowledge is a social exchange [6][5], therefore

[1] https://www.gov.uk/government/publications/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care

[2] https://newsroom.intel.com/news-releases/u-s-healthcare-leaders-expect-widespread-adoption-artificial-intelligence-2023/

[3] https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/survey-results.html

explanations are **social/interactive**. People expect explanations to be **truthful** and **thorough** explanation [8]. People usually prefer simpler and more **general** explanations[14].

This paper 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 human-machine trust items are based on several research studies about human-computer and human-machine trust [11] [13] [17]. However, the relation between the two is still a speculation and has yet been investigated.

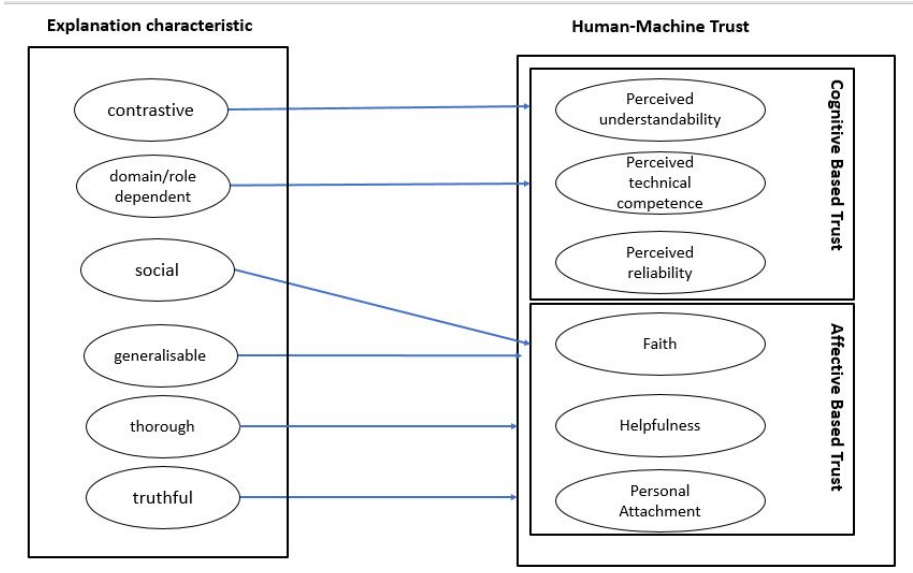

Fig. 1. trustworthy explainable AI in healthcare framework

4 Discussion and implication for future research

This paper proposed a framework of trustworthy explainable AI in healthcare. We derived characteristics of user-friendly explanations, and component of trust from previous studies. We are planning to undertake a qualitative and quantitative study 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. Alaszewski, A.: Risk, trust and health (2003)
2. Dibben*, M.R., Lean, M.: Achieving compliance in chronic illness management: illustrations of trust relationships between physicians and nutrition clinic patients. Health, Risk & Society **5**(3), 241–258 (2003)
3. Gunning, D.: Explainable artificial intelligence (xai) (2017)
4. Henman, M., Butow, P., Brown, R., Boyle, F., Tattersall, M.: Lay constructions of decision-making in cancer. Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer **11**(4), 295–306 (2002)
5. Hilton, D.: Social attribution and explanation. In: The Oxford Handbook of Causal Reasoning (2017)
6. Hilton, D.J.: Conversational processes and causal explanation. Psychological Bulletin **107**(1), 65 (1990)
7. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
8. Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., Wong, W.K.: 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. pp. 3–10. IEEE (2013)
9. Lipton, P.: Contrastive explanation. Royal Institute of Philosophy Supplements **27**, 247–266 (1990)
10. Lipton, Z.C.: The doctor just won't accept that! arXiv preprint arXiv:1711.08037 (2017)
11. Madsen, M., Gregor, S.: Measuring human-computer trust. In: 11th australasian conference on information systems. vol. 53, pp. 6–8. Citeseer (2000)
12. Malle, B.F.: How the mind explains behavior: Folk explanations, meaning, and social interaction. Mit Press (2006)
13. Mcknight, D.H., Carter, M., Thatcher, J.B., Clay, P.F.: Trust in a specific technology: An investigation of its components and measures. ACM Transactions on Management Information Systems (TMIS) **2**(2), 12 (2011)
14. Read, S.J., Marcus-Newhall, A.: Explanatory coherence in social explanations: A parallel distributed processing account. Journal of Personality and Social Psychology **65**(3), 429 (1993)
15. Rowe, R., Calnan, M.: Trust relations in health carethe new agenda. The European Journal of Public Health **16**(1), 4–6 (2006)
16. Thorne, S.E., Robinson, C.A.: Health care relationships: The chronic illness perspective. Research in Nursing & Health **11**(5), 293–300 (1988)
17. Yan, Z., Kantola, R., Zhang, P.: A research model for human-computer trust interaction. In: Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on. pp. 274–281. IEEE (2011)