Explainability design patterns in clinical decision support systems

Mohammad Naiseh[1]

Faculty of Science and Technology, Bournemouth University, United Kingdom
mnaiseh@bournemouth.ac.uk

Abstract. This paper reports on the ongoing PhD project in the field of explaining the clinical decision support systems (CDSSs) recommendations to medical practitioners. Recently, the explainability research in the medical domain has witnessed a surge of advances with a focus on two main methods: The first focuses on developing models that are explainable and transparent in its nature (e.g. rule-based algorithms). The second investigates the interpretability of the black-box models without looking at the mechanism behind it (e.g. LIME) as a post-hoc explanation. However, overlooking the human-factors and the usability aspect of the explanation introduced new risks following the system recommendations, e.g. over-trust and under-trust. Due to such limitation, there is a growing demand for usable explanations for CDSSs to enable the integration of trust calibration and informed decision-making in these systems by identifying when the recommendation is correct to follow. This research aims to develop explainability design patterns with the aim of calibrating medical practitioners trust in the CDSSs. This paper concludes the PhD methodology and literature around the research problem is also discussed.

Keywords: Explainability · Decision support systems · User-Centred Design · Trust

1 Context and motivation

The development of Clinical Decision Support Systems (CDSSs) has led to a surge of interest in systems optimised not only for expected task performance and accuracy but also other critical criteria such as safety, transparency, avoiding technical debt or providing explanations. While efforts to make CDSSs transparent and explainable have been demonstrated [18, 21, 20], failing to calibrate user trust is one of the new errors introduced by using these tools. For example, Bussone et al. [3] studied the effect of the explanation on trust and reliance. They concluded that overlooking the human factors and user experience in the design of CDSSs explanation could lead to medical professionals over-trust the system recommendation, even when it is wrong, i.e. over-reliance. In the same way, the explanation that does not provide enough information could lead to users rejecting the suggestions, i.e. self-reliance or under-trust [13].

Explainability in the medical domain is defined as a set of measurable, quantifiable, and transferable attributes associated with the intelligent system that the aim to calibrate medical practitioners trust [20]. While the increasing interest in explainable clinical systems, it has become important to develop explainability design solutions that better suit the clinical decision-marking with the focus on their trust as a crucial factor. Here, the research distinguishes between the explainable model which generates the explanation and the explainable interface which makes the explanation usable and useful for the medical practitioners. This research is limited only to the explainable interface and its relevant design factors and facets to support the medical practitioners' decision-making and reduce the errors of failing to calibrate user-trust issue, i.e. under-trust and over-trust.

Among the possible ways for trust calibration (e.g. algorithmic assurances [1], automation reliability [11], personalisation [9], and explainablity [3]), this research focuses on the latter aspect of the trust calibration. This research argues that the effectiveness of the explanation design in relation to CDSSs can be itself basis or solutions to contribute to calibrate medical practitioners trust. Furthermore, this research is limited to the post-hoc explanation capabilities which refer to explanation models that are applied after model training. This PhD aims to develop HCI design patterns for post-hoc explanations in CDSSs with the aim of reducing trust calibration errors.

Qualitative research is the baseline for achieving that goal. This is due to the intense medical nature of the problem and solutions and the need for intensive input from medical practitioners. Studies including a systematic literature review, semi-structured interviews and think-aloud protocol are used to provide a conceptualisation for various aspects of explainability in clinical decision support systems. The design patterns and the explainable interface are then will be evaluated by means of two case studies (Prescribing breast cancer treatment and screening Palbociclibii Cancer treatment prescription) on IQemoi[1] prescribing system to investigate the efficiency of the produced solutions in calibrating user trust.

2 Background and related work

2.1 Human-Computer Interaction (HCI) and Explainability

The Human-Computer Interaction research community has identified several benefits of generating explanations by artificial intelligence agents [15]. For example, Samek et al. [17] present four social benefits aspects that are important for users interacting with intelligent systems. Interaction techniques and user feedback with explainable agents such as recommender systems and expert systems have widely studied in the literature of HCI. For example, Kuleza et al. [8] developed an explanatory debugging system that explains the decision and incorporates user feedback, which was shown to lead to better predictions, sounder

[1] iQemo, from iQ HealthTech, is a complete managed chemotherapy patient management and prescribing module. https://www.iqhealth.tech

mental models and higher user satisfaction. During the development of the explainable system, the research community identifies a collection of explanation properties and requirements that are important to generate useful and usable explanations. Many of these aspects are built based on the literature of social sciences, psychology and education to mimic the human to human explanations. Sokol et al. [19] present 11 usability requirements for the explanations which are: Soundness, Completeness, Contextfullness, Interactiveness, Actionability, Chronology, Coherence, Novelty, Complexity and Personalisation. Also, the initial findings from this PhD research provide in-depth investigation about the conceptualisation of the personalisation aspect in a previous work [14]. Since the implementation of these aspects has been limited to low stake applications, these principles and findings may not translate to high stake applications where trust calibration and safety are crucial requirements. A lack of clearly defined user experience aspects that the explainable interface should be considered in high stake applications with trust-calibration is the main focus is still missing.

Additionally, the HCI research literature identifies the risks of the explainable interfaces on users decision-making and their perception of the system. These risks are likely to arise when the designers overlook the user experience factors. For instance, users may feel that the system is trying to manipulate them when the explanation does not contain enough information or consistent with their prior beliefs [6]. The ongoing research findings identify six different possible risks that could arise in the absence of user-centred approaches, which are: Over-trust, Under-trust, Refusal, Perceived loss of control, Information overload and Suspicious motivation [13]. Finally, the HCI research community argued the ability of the explainable systems to be engineered to work in a long-term and evolve during the time based on what has already explained to the end-users before [12].

2.2 HCI design patterns

HCI design patterns are predefined and reusable design solutions that describe and solve users' problems. Alexander [2] argued that the pattern should capture context where the pattern can be applied, the problem and its environments, and the design guidance. Designers of new systems can take benefit from the design pattern and save the efforts and resources to build usable systems. When designing the explainable system, the development process needs to consider the explainee characteristics, needs, usability aspect and safety requirements [19]. Design pattern could help the baseline for such requirements by identifying the possible design problem and make the design solution available for future practice. For instance, TELL project [5] uses the design patterns to support the understanding of the learning process that occurs within the network supported collaborative learning. To date, very little work has investigated HCI design patterns for the explainable interfaces e.g. Chromik et al. [4] present and discuss several dark design patterns that designers of the explainable interfaces should avoid it.

2.3 Trust calibration

Existing work has investigated how users develop their trust with the intelligent systems with focusing on the factors that affect user trust in complex systems (e.g. transparency) [7, 9]. Trust is a dynamic and complex psychological and sociological concept, when the trustee over-trust or under-trust the automation system could lead to critical consequences, especially in safety-critical domains. Madhavan et al. [10] defined the problem of failing to calibrate users trust as it is a failure in the system design in balancing the actual safety and the users' perceived safety. Providing explanations is meant to be one of the factors that may contribute to the problem. Users may over-trust the system when the explainable interface is not built on the user experience aspect [3]. Also, the explanation may lead to users under-trust the systems, when the explanation is perceived to have a limited quality or fitness to the user intentions and context [18]. The challenge for HCI is to define the design properties and activities for achieving the right balance between actual and perceived safety.

3 Research aim

This research aims to develop explainability design patterns based qualitative approach to calibrate user trust in CDSSs by making the medical practitioners aware when to follow the system recommendations or not and potentially avoid under-trust and over-trust issues. The PhD contributes to the literature by helping the elicitation and customisation of the variability in the requirements and design of CDSSs interface that support medical practitioners safe and effective decision-making.

3.1 Research questions

The research focuses on the explainbility user experience aspects and trust calibration in CDSSs by asking the following questions:

RQ1: What are the user experience aspects of explainability?
RQ2: What makes the CDSSs explainable for medical practitioners?
RQ3: What are the explainability aspects and features that may contribute to failing in calibrating user-trust?
RQ4: What explainability design features could future CDSSs have to cater to medical practitioners calibrate their trust?

4 Research objectives and methods

Objective 1: Conduct a Systematic Literature Review to explore the explainability user experience aspects in the literature and develop an understanding of relevant user trust calibration problem.

As a first step in achieving the goal of this research, the need to understand explainability from artificial intelligence and Human-computer interaction perspectives is important to formulate the explanation design space. The empirical literature regarding how researchers and practitioners in the field provide explanations to end-users is reviewed to provide a foundation for satisfying the research aim and also to inform the exploratory studies and the prototyping stage. The research reviews the literature concerning the explainability aspect from both Explainable Artificial Intelligence and Human-Computer Interaction perspective. In addition, trust calibration theories are reviewed with its diversity of design guidelines. The literature on decision-making in medicine is also reviewed to provide foundations to the solution and also to inform the exploratory studies.

Objective 2: An empirical investigation into the post-hoc explanation capabilities that may affect user trust in the CDSSs through series of qualitative approaches. The empirical investigation is built on the result of the first objective. This objective informs the research regarding the explanation design requirements from healthcare professionals' perspectives. Ultimately, the data collection of the qualitative approach with the healthcare professionals and the initial analysis of the results are in process with collaboration with IQ Healthcare[2] and three hospitals in the UK. To achieve this objective, several steps and methods are followed. The first study gauges the opinions of medical practitioners in relation to the functional and non-functional requirements for explainability (e.g. the framing of the explanation, the content, the delivery methods and modalities). This study uses a think-aloud protocol for the purposes of data collection to allow participants to discuss their opinions about the existing literature around explainability. The second study utilises a semi-structured interview to investigate what makes the CDSSs explainable and trustworthy for healthcare practitioners. This will help the design of the explanation in CDSSs that influence user-trust. This stage also will develop a taxonomy of these findings and their relation to trust-calibration.

Objective 3: Iterative prototyping through design sessions to build design patterns toolkit for the explainable CDSS interface that calibrate user-trust.

In this objective, the research will attempt to develop explanation design patterns for CDSSs that calibrate user trust. That means that the healthcare professionals will be better informed about the CDSSs recommendations so that the user trust is calibrated. In this stage, the consideration of participants' roles and requirements will be taken into account to help the analysis for better understanding the qualitative data. Ultimately, the researcher will attempt to find aiding design patterns that help healthcare practitioner to use the CDSS in an effective and safe way with a reduction to under-trust and over-trust errors. This objective is achieved by means of design sessions. In the design sessions, various scenarios will be shown with different interface designs. The participants will

[2] iQ HealthTech is a software development company, with a common goal, to improve the way IT systems are used for healthcare.

be encouraged to find better designers, so they are better informed about the CDSSs recommendations.

Objective 4: Design and conduct qualitative and quantitative user studies on the prototype to evaluate the user trust, the effectiveness of the prototype and validate the approach.

The resulting explainability design patterns for CDSSs will be tested for its trust-calibration and the effectiveness of the design. Target users are medical practitioners. Users will be asked to perform certain tasks with the developed interface. These tasks will be based on the expected functionalities of the explainable interface. Two case studies will be used in this stage i) Expert system to prescribe breast cancer chemotherapy treatment for breast cancer ii) Rule-based system for screening prescription for Palbociclib cancer treatment. IQemo prescribing system will be used to build and validate the design patterns toolkit.

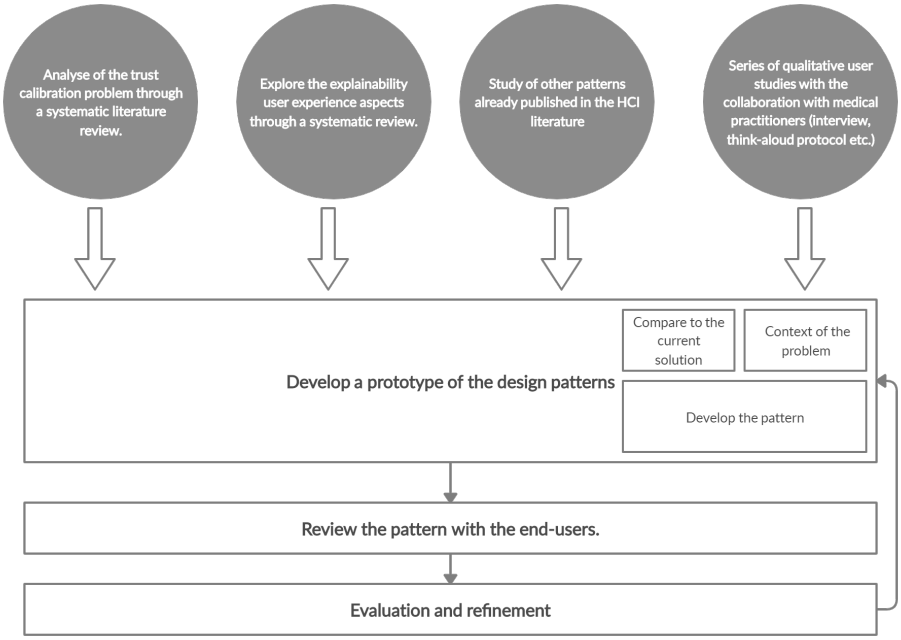

Fig. 1. The PhD methodology for forming the design patterns for CDSSs.

4.1 Research validity

To strengthen the external validity of the research, several aspects are addressed. First, the target participants in this research are selected by a strategy combining

convenience sampling and maximum variation sampling [16]. The use of this convenience sampling approach reflects the difficulty of gaining medical practitioners in this kind of research. Any possible bias traditionally related to convenience sampling tried to be mitigated by combining a maximum variation sampling so that the approached hospitals covered different characteristics regarding size, application domain, domain knowledge and practitioners experience. Second, the research has interviewed at least two practitioners per session to reduce the risk of bias and misinterpretation. Third, the participated hospitals were developing prescribing and diagnosis systems from the oncology department. It is possible that this factor may have an impact on our research findings. Therefore, it is important to highlight that the findings of this thesis might be considered more relevant to this type of expert systems (diagnosis and prescribing). However, future research is needed to validate the research findings in different application domain areas where the nature of the decision making strategy is different (e.g. dentist decision support system).

5 Conclusion and current progress

Driven by the increasing interest in decision support systems in the clinical settings, the understanding of potential user errors that might emerge is also essential. This paper present the ongoing PhD project that investigates explainability solutions with the aim to avoid failing to calibrate user-trust, i.e. over-trust and under-trust. Also, this paper elaborated on the status of the research problem and identified three distinct research strands in the literature that are relevant to the problem. Currently, the researcher is performing a continuous process of analysing the qualitative data that emerged from the qualitative user studies and identifying the properties of the explainability in the clinical settings. Once it has been done, the researcher will develop multiple prototypes of the design patterns and review the patterns with the potential end-users.

Acknowledgments

This work is partially funded by iQ HealthTech and Bournemouth university PGR development fund.

References

1. Aitken, M., Ahmed, N., Lawrence, D., Argrow, B., Frew, E.: Assurances and machine self-confidence for enhanced trust in autonomous systems. In: RSS 2016 Workshop on Social Trust in Autonomous Systems (2016)
2. Alexander, C.: A pattern language: towns, buildings, construction. Oxford university press (1977)
3. Bussone, A., Stumpf, S., O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: 2015 International Conference on Healthcare Informatics. pp. 160–169. IEEE (2015)

4. Chromik, M., Eiband, M., Völkel, S.T., Buschek, D.: Dark patterns of explainability, transparency, and user control for intelligent systems. In: IUI Workshops (2019)
5. Deliverable, W.: Introducing a framework for the evaluation of network supported collaborative learning
6. Eiband, M., Buschek, D., Kremer, A., Hussmann, H.: The impact of placebic explanations on trust in intelligent systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–6 (2019)
7. Glass, A., McGuinness, D.L., Wolverton, M.: Toward establishing trust in adaptive agents. In: Proceedings of the 13th international conference on Intelligent user interfaces. pp. 227–236 (2008)
8. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th international conference on intelligent user interfaces. pp. 126–137 (2015)
9. Liu, C.: Human-machine trust interaction: A technical overview. Trust Modeling and Management in Digital Environments: From Social Concept to System Development: From Social Concept to System Development p. 471 (2010)
10. Madhavan, P., Wiegmann, D.A.: Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science **8**(4), 277–301 (2007)
11. Merritt, S.M., Heimbaugh, H., LaChapell, J., Lee, D.: I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. Human factors **55**(3), 520–534 (2013)
12. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
13. Naiseh, M., Jiang, N., Ma, J., Ali, R.: Explainable recommendations in intelligent systems: Delivery methods, modalities and risks. In: The 14th International Conference on Research Challenges in Information Science. Springer (2020)
14. Naiseh, M., Jiang, N., Ma, J., Ali, R.: Personalising explainable recommendations: Literature and conceptualisation. In: WorldCist'20 - 8th World Conference on Information Systems and Technologies. Springer (2020)
15. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction **27**(3-5), 393–444 (2017)
16. Robinson, O.C.: Sampling in interview-based qualitative research: A theoretical and practical guide. Qualitative research in psychology **11**(1), 25–41 (2014)
17. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)
18. Schäfer, H., Hors-Fraile, S., Karumur, R.P., Calero Valdez, A., Said, A., Torkamaan, H., Ulmer, T., Trattner, C.: Towards health (aware) recommender systems. In: Proceedings of the 2017 international conference on digital health. pp. 157–161 (2017)
19. Sokol, K., Flach, P.: Explainability fact sheets: A framework for systematic assessment of explainable approaches. In: In Conference on Fairness, Accountability, and Transparency (FAT* '20) (2020)
20. Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: contextualizing explainable machine learning for clinical end use. arXiv preprint arXiv:1905.05134 (2019)
21. UCLA, E.: Outlining the design space of explainable intelligent systems for medical diagnosis (2019)