AI Healthcare System Interface: Explanation Design for Non-Expert User Trust

Retno Larasati, Anna De Liddo and Enrico Motta

Knowledge Media Institute, The Open University. Walton Hall, Milton Keynes, United Kingdom

Abstract

Research indicates that non-expert users tend to either over-trust or distrust AI systems. This raises concerns when AI is applied to healthcare, where a patient trusting the advice of an unreliable system, or completely distrusting a reliable one, can lead to fatal incidents or missed healthcare opportunities. Previous research indicated that explanations can help users to make appropriate judgements on AI Systems' trust, but how to design AI explanation interfaces for non-expert users in a medical support scenarios is still an open research challenge. This paper explores a stage-based participatory design process to develop a trustworthy explanation interface for non-experts in an AI medical support scenario. A trustworthy explanation is an explanation that helps users to make considered judgments on trusting (or not) and AI system for their healthcare. The objective of this paper was to identify the explanation components that can effectively inform the design of a trustworthy explanation interface. To achieve that, we undertook three data collections, examining experts' and non-experts' perceptions of AI medical support system's explanations. We then developed a User Mental Model, an Expert Mental Model, and a Target Mental Model of explanation, describing how non-expert and experts understand explanations, how their understandings differ, and how it can be combined. Based on the Target Mental Model, we then propose a set of 14 explanation design guidelines for trustworthy AI Healthcare System explanation, that take into account non-expert users needs, medical experts practice, and AI experts understanding.

Keywords

Explanation, Trust, Explainable Artificial Intelligence, AI Healthcare, Design Guidelines, Participatory Design

1. Introduction

Trustworthiness, the capability to independently establish the right level of trust in an AI system, is progressively becoming an ethical and societal need. Trust is humans' primary reason for acceptance [1], without which the fair and accountable adoption of AI in healthcare may never actualise. 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], and this might not be achieved if there is an impediment to AI adoption and AI usage from the general public (non-expert healthcare customers).

Developing trust is particularly crucial in healthcare because it involves uncertainty and risks for vulnerable patients [3]. However, the lack of explainability, transparency, and human understanding of how AI works are key reasons why people have little trust in AI healthcare applications; and research indicates that transparency [4] and understandability [5] can be effectively used as means to enhance trust in AI systems. Explainable AI is argued to be essential "to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" [6]. Nevertheless, the lack of trust is not the only problem. Previous research indicates that non expert users tend to over-trust and continue to rely on a system even when it malfunctions in some circumstances [7]. To help non-expert healthcare customers to appropriately trust AI systems, not over-trust or distrust, the system should be able to give an appropriate understandable explanation for that specific target audience. This paper aims at identifying the explanation components of AI healthcare system interfaces, for non-expert users to appropriately inform their trust in the AI system. We carried out a user study to determine these explanation components and then used them to inform a set of design guidelines for trustworthy AI Healthcare Systems explanation interfaces.

We chose a stage-based participatory method, adapted from Eiband et al. [8], that has been previously successfully applied to design explanation of recommender systems in fitness applications [8]. This method particularly fits our case since it enables an individual investigation of expert and non-expert views on the problem and then provides a framework to combine expert and non-expert knowledge to inform design requirements. The stage-based participatory process consists of two phases. The first phase focuses on "what" to explain through the construction of an Expert Mental Model (what "can be explained") and a User Mental Model (what "needs" to be explained). The second phase focuses on synthesising the two models in a Target Mental Model, which

Joint Proceedings of the ACM IUI 2021 Workshops, April 13-17, 2021, College Station, USA

retno.larasati@open.ac.uk (R. Larasati); anna.deliddo@open.ac.uk (A. De Liddo); enrico.motta@open.ac.uk (E. Motta)

0000-0002-6412-2598 (R. Larasati); 0000-0003-0301-1154 (A. De Liddo); 0000-0003-0015-1952 (E. Motta)

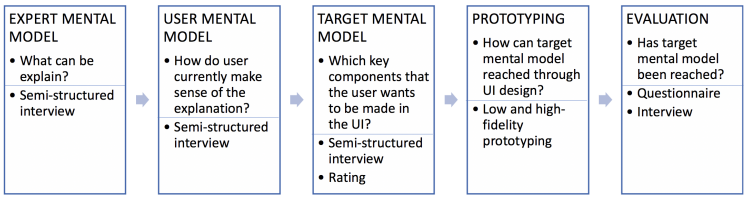

Figure 1: The stage-based participatory process for our case. Inside the box: guideline question and data collection method

describes "how to convey the explanation" by design and developing a prototype technology.

To build the Expert Mental Model, depicting the key components of explanation that need to be communicated to patients, we carried out a series of interviews with medical professionals. Second, we conducted semi-structured interviews with non-experts to identify the User Mental Model, which captures users' needs and expectations in terms of AI explanation. Finally, we conducted a third set of semi-structured interviews with both AI experts and non-experts to determine how explanation content could be communicated to the non-expert users to respond to the identified users needs (Target Mental Model). From the Target Mental Model, we then derived a list of design guidelines, which we then used to develop a prototype explanation interface for an AI breast cancer risk assessment support systems. In particular, we focused on a self-managed breast cancer risk scenario, in which results of mammography scans are automatically analysed by an AI system and need to be communicated to the prospective patients. We choose a self-managed health scenario, because it represents the extreme case, in which non-expert users are presented with AI results, without any support from medial or AI experts, and therefore the explanation is the only mediating interface between patients and the AI system.

2. Background

In recent years, several studies explored different approaches to design explanation of the outputs from intelligent systems [9][10][11]. Some of the research focused on explanation designs for AI healthcare systems [12][13]. Despite the fact that many approaches have been proposed, the explanation design for AI healthcare system mostly targets expert user [14][15]. Explanation design specifically targeted to non-expert users has received scarce attention, despite the recognised importance of improving non-expert user's understanding of the AI system to positively affect users' trust in the system[16], and trust in the system recommendations [17].

To improve users' understanding of the AI system with explanation, we first need to determine how they make sense of an explanation (what does the users' mental model of an explanation looks like). Unlike previous studies [8][18], we did not have an available working system to understand the users' mental models. This difference affected how we elicited users' feedback. We conceptualised and used a hypothetical AI diagnosis system (inspired by similar commercial systems) to interrogate both expert and non-expert, and elicited their mental models from reflections on the system and previous experience with healthcare explanation. Our hypothetical system was a Breast Cancer Self-Assessment system, a medical system to assess breast cancer risk tailored for non-experts.

As mentioned above, following Eiband et al. [8], we carried out a stage-based participatory process consisting of two phases and five stages. The first phase focused on "what" to explain and consisted of two stages: the Expert Mental Model definition and the User Mental Model definition. The second phase focused on "how to convey the explanation" and consisted of three stages: the Target Mental Model construction, the Prototype development stage (to implement the Target Mental Model in a realistic application case), and the Evaluation stage, to further test the prototype technology. In this paper, we conducted four out of five stages in Figure 1, leaving further testing and evaluation of the prototype for future research. Each stage is described in details in the next section.

3. STUDY DESIGN AND METHODOLOGY: STAGE-BASED PARTICIPATORY PROCESS FOR EXPLANATION DESIGN

3.1. Experts Mental Model

The expert mental model definition stage aimed at capturing experts' understanding and vision of what an appropriate explanation of AI medical support system results to non-experts should look like. The experts involved in its development were both machine learning technologists and medical professionals. This research stage

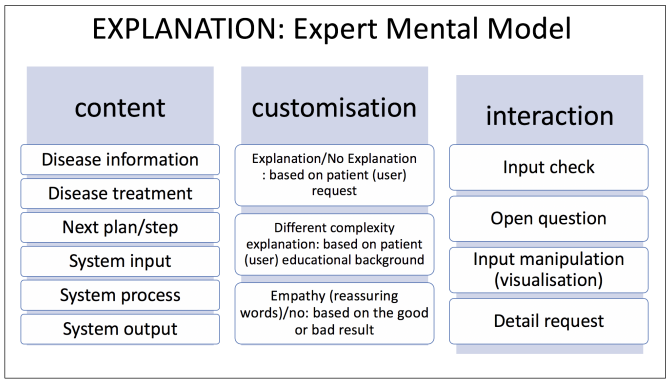

Figure 2: Expert Mental Model analysis result: explanation components

aimed at defining what can or should be explained to the wider public from an expert perspective, by distilling a series of explanation components, which represent the Expert Mental Model.

Six participants were recruited by email, from the authors' personal research and social networks, three were AI/machine learning developers/researchers, and the other three were doctor/physicians (general practitioners and oncology specialists). The main guiding questions that drove this stage were: what can be explained?; and what does an expert explanation for non-experts looks like? We asked the questions based on participants respective expertise (medical professionals and AI experts). We also showed participants two examples of breast cancer-related systems currently in commerce, to understand how experts make sense of AI systems' outputs and how they would explain the results to non-experts.

Result and Analysis

We analysed interviews' data using Grounded Theory [19]. Three sets of explanation components emerged. The first set of explanation components entailed the *Content* of the explanation, and described what information needs to be included in the explanation. The second set of explanation components entailed the required *Customisation* of the explanation, what needs to be considered when explaining, and changed accordingly on a case by case basis. The third set of explanation components entailed the explanation *Interaction*, the interactivity opportunities that need to be available to users during an explanation.

In terms of content of the explanation itself, AI experts answers were quite straight forward; users need to know about input, system process, and output. *"We have the inputs and intermediate results. The inputs are different variables, as a driver for the predictor and explanatory variables,[...]. They can interact with the app and see the simulation."*- A2

The system process answers varied considerably, and spanned from providing information such as features' importance, to providing the name of the algorithm or who made it. *"It's like, for example, if they're trying to recognise cancer in a certain image, so this is the feature that helped me the most having this conclusion"*- A1. *"You can try to show the formulation of the calculation. But some algorithms do provide explanation on how it works."*- A3 This means that even though this explanation component was deemed important, AI experts were not clear on what and how to present it to non expert users.

From the medical experts perspective on the other hand, explanations they usually gave to patients consist of disease information, possible treatments to choose, and the next step for the patient to take. They mentioned that explaining diagnosis works differently if the diagnosis result is bad. *"When we deliver the diagnosis to a patient, we consider the situation as well. [...] For breaking bad news, we usually deliver the news layer by layer. so not directly go to the diagnosis, we have some introduction first."*- M1. If the result is bad, reassuring words are needed to help patients feel less stressed and worried. If the diagnosis result is good or if there is no sign of distress from the patient, there is less need for reassuring words. *"I think one of the important things if it's about serious conditions, we need to put more empathy."*- M3 This is in line with previous research on medical explanations *How to Break Bad News: A Guide for Health Care Professionals* [20] and similar explanation protocols have been proposed and tested in the literature [21][22].

The medical experts mentioned that explanation was not given by default but based on patients' request and customised to patients' needs. *"It depends on how curious they are. If the patient just wants to know the diagnosis, then I may just tell them about it."*- M3. AI experts also

mentioned explanation should probably only be provided on request. According to the AI experts, they rarely explain how the system works to non-expert user in a real-life situations unless the user asks for it. *"if the app is working properly you don't need to explain. But if there is a problem, you need to explain what is going wrong."*- A3. One AI expert even argued that non-expert were not interested in knowing the logic behind/system process. *"I have never met a common user that is interested in artificial intelligence or the machine learning of it. Even the expert from the Ministry (people they work for), they were not really curious."* - A2.

The medical experts also reported that they assess what the patient knows and the patient's perception. One medical expert mentioned that people who live in a rural area might have different knowledge than people who live in a big city, meaning the explanation is customised to the patients' knowledge. *"People in the rural area, don't get the privilege to get a proper education, so it's challenging for them to absorbs the explanation."*- M2

The explanation components related to explanation interaction, reflect on the modalities in which experts communicate the explanation. Medical experts mentioned how they usually ask for confirmation about the patient symptoms and worries before making a diagnosis (input check). The second component related to the capability for non-experts to raise open questions. After giving patients their results, medical experts would always ask if there were any more questions. This interaction usually involves a back and forth exchange, until the patients has no further questions. *"...Then we will explain what's the next step. And we will ask if they have any questions or not. Including the diagnosis and the plan."*- M3. *"whenever patients ask, we then answer the questions directly."*- M1.

One AI expert mentioned that showing how the output changes could help non experts to understand the system better (input manipulation and visualisation). AI experts also mentioned how it could be overwhelming for the user to read all the explanation, and suggest it would be better to give users the option to request details if they need them (details request). *"We need the user to see the general output, but they can expand on some detail. Making it simple, just a few statements, and the general result, and if the user is curious, they can dig into it."*- A3. The Expert Mental Model outcome from the analysis can be seen in Fig 2.

3.2. User Mental Model

In the User Mental Model research stage, we captured users' understanding and their perspective on how explanation should be presented in an AI medical support system. The purpose of this stage was to acquire knowledge about how do users currently make sense of explanations. This acquired knowledge was then structured in several key components of explanation, which constitute the User Mental Model.

Szalma & Taylor (2011) showed that trust propensity is one of the human-related factors that could affect the response to an intelligence system [23]. To account for trust propensity, we sampled the participants based on their dispositional trust towards an AI medical support application and made sure there was a nearly equal number of people in each trust groups (the AI sceptic, the open-minded, the AI enthusiast). We recruited four participants for the three groups representing three levels of dispositional trust, with 12 participants in total. To identify the level of trust, we asked the perspective participants to answer the following question: "if there was a cancer risk assessment/self-detection application available on the market, how likely would you be to use it? Please rate the likelihood from 1-7". This question was sent in advance of the interview invitation. The participants were then grouped into three groups, the sceptic (1-3 likelihood responses), the open-minded (4-5 likelihood responses), and the enthusiast (6-7 likelihood responses). We sought to balance out the age range (twenties to forties) because research suggests that age could affect users' trust towards a system, where older adults are more likely to trust the system than younger adults in a medical management system (decision aid) [24]. We also balanced out the male-female participants by recruited one male in each group because we recognised despite male breast cancer is only accounting for less than 1% of all breast cancer diagnoses worldwide [25], sometimes men are included in the decision making towards the usage of a particular system for affected women close/related to them.

We followed the same interview structure as in the experts' interviews. The main guiding questions we asked the non-expert users were: how do users currently understand AI explanations?; what does a user explanation looks like? We then showed the participants two examples of breast cancer-related systems to probe non-expert users reflections and feedback on the AI system's result and explanation.

Result and Analysis

We carried out a Thematic Analysis [26] to analyse the interviews' data. The same three sets of explanation components could be identified *Content, Customisation, and Interaction.* In receiving a diagnosis, participants explained that they would like to know about the *disease information.* They mentioned: disease name, disease symptoms and the severity of the disease as key information they would like to receive. Participants also reported they would like to know about the *next step/action* or action that they could or should take, for example, information about the *disease treatment* that they should undergo after diagnosis, or if they have to make an appointment

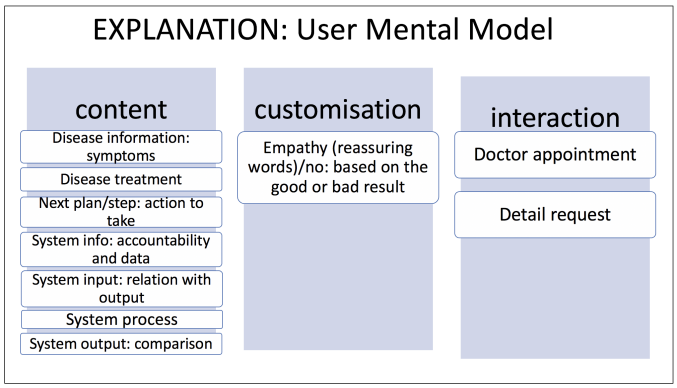

Figure 3: User Mental Model analysis result: explanation components

with their doctor or physician. *"you got cancer, and your options would be these, these, and these, and this is how I want to proceed. These are your options."* - E2. *"do I need to contact my physician directly or is there a next step that is also provided by the application itself?"*- OM1

These diagnosis-related explanations, both disease information and next step/action, could be considered more local/disease specific explanation. However, participants also wanted more general explanations about the AI system, *system information*, which was not related to either the inputs or the results. One of the participants asked for information about the *system process/algorithm. "I would want to know, what are they doing actually in the background to do this?"*- E1. However, not all participants expressed their interest in knowing the system process; some were not keen to know the information. They argued that in a stressful situation, such as a positively assessed cancer, their focus would not be on the system information and more on their well-being. *"Says that I have cancer, then I am not going to be interested in the system process"*- S1.

This arguments match with the AI experts opinion we mentioned above, recognising that non-experts may usually not be interested in knowing the technical side of a system. Non-expert users reluctance to know the technical information was a matter of timing and their emotional state after receiving a diagnosis. However, it also reflects their reluctance caused by the possibility of not understanding the technical terms used to explain the process. *"I mean, the very hard, fine grain details? it will be incomprehensible for me because I am not familiar with the technology and everything."*- E2. this confirms previous research, arguing that what people consider acceptable and understandable explanation depends on people's domain knowledge or role [27][28].

Another emerging explanation component was *system data.* non experts users mentioned they would like to have information about the volume of a database used to train the algorithm or the data features used for the prediction. *"So at least I have to know how big is their database."*- OM2. *"Explain to you the quality of features and characteristics; it is because this thing has this colour shape"*- E2. However, participants also talked about the data they provide, their personal input data, they expressed concerns of data privacy, and demanded specific information about that. *"And how am I sure that my breast picture will not be leaked to be utilised for other intentions and such."*- OM1. *"Where are these data going?"*- OM4.

Participants also talked about the system *accuracy, credibility. "However, for my health, I think it will be quite beneficial if I know how accurate it can be"*- OM1. The credibility they mentioned was related to the institute/company that developed the system. Credibility could also mean if the system has been tested and approved by the appropriate health institution.

Besides the information that should be included in the explanation, participants also talked about how the explanation should be delivered. Participants demanded for the AI results to be presented with care and *empathy*, especially if their result was not in their favour. Empathy is the ability to understand and share the feelings of others, and an empathetic statement should include phrases that help to establish a connection with the user. Participants mentioned empathy or reassuring words only in the case of "bad news" or presented if the result is not good; therefore, we put it under customisation in the User Mental Model.

"if I want to use text explanation, I think you should be, in terms of style of shaping the statement that you present to the user, I think you should always follow, sort of defensive language. So again, it might be quite direct and aggressive to say to the user, you have cancer, exclamation mark. [...] Be a little bit more reserved, rather than explicit, into your statements because it's quite a sensitive matter."- E2.

Participants who expected care and empathy were more concerned with the choice of words and how "delicately" the AI system delivers the diagnosis results.

Other than text and words, participants mentioned the use of graphics and images to communicate the explanation, for example, by showing comparison images of normal condition vs abnormal condition. The graphic/image to show comparison, we put under explanation content in User Mental Model, because regardless of the result (good/bad), the user wants to see the opposite case and decide themselves if the result makes sense to them or not. *"Perhaps have some examples of how affected breast looks like, how unaffected breast looks like. So you can compare yourself with what is being put in your input."*- E2. *"and then the image comparing, you know, both, my results and the healthy ones."*- S1. Participants requested to show the opposite case was in line with the literature in cognitive psychology, which states that human explanations are sought in response to particular counterfactual cases [29][30]. Our finding confirm that counterfactual case/contrastive explanation is argued to be an explanation that is understandable for user [31][32][28].

For interaction with the AI system, participants expressed their needs for a course of action, which is an additional feature of *doctor appointment* included in the explanation interface. In how they will interact with explanation, participants wanted to be able to request detailed information rather than presented with the full long explanation in one go. The mentioned that the explanation detail could be presented as a link to an outside source or as a piece of expandable information. The User Mental Model outcome from the analysis can be seen in Fig 3.

3.3. Target Mental Model

In the Target Mental Model research stage, we identified what key components of an explanation (from the expert perspective - Expert Mental Model) the users might want to be included in a AI explanation User Interface (UI). The Expert Mental Model's explanation components were combined with the explanation components from the User Mental Model to form the Target Mental Model. We conducted semi-structured interviews with the same group of non-expert participants involved in the User Mental Model definition.

During the interviews, the main guiding question was: which explanation components users want to be realised in a UI to explain AI results? We asked participants to reflect on the explanation components from the Expert Mental Model and discuss which one they considered most important and valuable. Participants were asked to explicitly reflect on each explanation component by giving a rating of importance (form 0-10) and expressing their opinion on each of them. Based on the critical analysis of User and Expert Mental Model, combined with the analysis of users' feedback of expert mental model views, we distilled the Target Mental Model.

content	disease info	treatment	next plan	input	process	output
	10	10	10	10	7	10
custom	explanation req	user education	empathy			
	8	4	10			
interact	input check	open question	input manip	detail		
	10	7	10	10		

Table 1
Median values of Expert Mental Model explanation components rating to inform Target Mental Model

Result and Analysis
The median values of explanation components' rating given by the users are reported in Table 1. Under the *Content* explanation components set in the Expert Mental Model, the *system process* was not seen as crucial, since not all participants were interested in knowing the technicality of how the AI system made a decision/prediction.

Under the *Customisation* set of explanation components, *empathy/reassuring word* was rated high by the participants. *User request* was also rated relatively high because some of the participants argued that explanation should always be available whether a user requested it or not. The lowest-rated component under the *Customisation* group was *user education* and was deemed unnecessary since explanation should be understandable for lay users regardless of their educational background. Under the *Interaction* group, all components were rated as important except for *open question*. Some participants were sceptical about openly asking questions to the AI system and preferred to wait to ask questions to a doctor.

The final Target Mental Model is shown in Figure 4. The explanation components included were obtained from the combination of explanation components from the Expert Mental Model and User Mental Model, then revised according on users' perceptions and preference on experts views. The explanation components with lower rating score are indicated with lighter text in figure. As an additional step, we went back to the experts and asked a follow-up question to the medical experts for the explanation components that appeared in the User Mental Model but not in the Expert Mental Model, such as *system accountability and data* and *doctor appointment*. According to them, system's certification and accountability were not essential to be included in the explanation. If the application is recommended by the healthcare authority (e.g., for the UK, NHS), it would be considered enough for them. The *doctor appointment* component did not come up in the interview before because they expected it as a given feature.

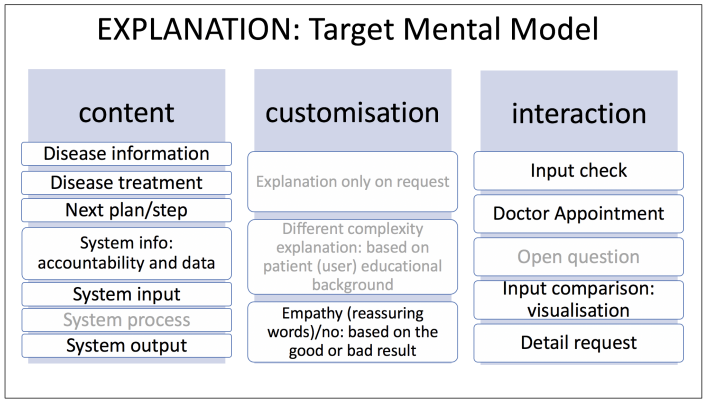

Figure 4: Target Mental Model analysis result: explanation components

Explanation Design Guidelines (EDG)		Descriptions	Requisite
Information Included (EDG1-EDG7)	Disease Information	general disease information e.g.: name, symptoms, caused	Yes
	Disease Treatment	treatment options and information	Yes
	Next Plan/Step	next step user could take following the result	Yes
	System Information	general system information e.g.: data used, system certification	Yes
	System Input	data the user inputted	Yes
	System Process	system algorithm or the technical process to gets its results	Optional
	System Output	system result e.g.:pre-diagnosis, recommendation	Yes
Information Delivery (EDG8-EDG9)	Empathy (Reassuring Words)	delicately deliver the results with carefully selected words	Yes
	Simple and General	uncomplicated wording that is acceptable for lay users from various education background and level	Optional
Interaction Included (EDG10-EDG14)	Input Check	for the user to check the input (is it correct or not)	Yes
	Doctor Appointment	for the user to make a doctor appointment	Yes
	Open Question	for the user to ask open questions	Optional
	Input Comparison (Visualisation)	for the user to compare the result with other data	Yes
	Detail Request	for the user to request detailed information	Yes

Table 2
Our 14 explanation design guidelines, categorised by information included, information delivery, and interaction included.

3.4. Design Guidelines and Prototype

By reflecting on the findings of the Target Mental Model, we propose 14 explanation components/design guidelines for trustworthy AI medical support system interfaces (See Table 2). Those guidelines were grouped into three categories that mirrored the Target Mental Model's explanation components sets: Explanation Content/Information to be Included, Explanation Customisation/Information Delivery, and Explanation Interaction/Interaction to be afforded. Each guideline references the explanation contents from the Target Mental Model, except for *explanation request* component.

We decided to not include *explanation request* in the guidelines in consideration of several regulations, such as The European Union's General Data Protection Regulation (GDPR) and European Commission Checklist for Trustworthy Artificial Intelligence (ALTAI)[1]. According to these regulations, explanation should be always provided, by law, to any uses when AI is involved. AI in healthcare was classified as high-risk AI according to White Paper On Artificial Intelligence by European Commission [2], which makes explanation availability even more essential in a healthcare scenario. We therefore removed the "explanation request" option from the design guidelines since, even if desirable from an non-expert users perceptive, would be an unethical and unlawful design choice.

[1]https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust
[2]https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

We then designed a user interface prototype based on the guidelines at Table 2. We explored each guidelines' presentation possibilities and the specific functionalities of the system that could realise them. We decided on a website where the user could carry out breast self-assessment based on screening images from their medical scan portable device. The final prototype was developed after several cycles of feedback between designers, and was then uploaded as a website at (https://retnolaras.github.io/care/).

4. Discussion

Previous research have used the development of mental models to fully explore users' understanding and help the design of transparent AI systems in various contexts [18][33][34], including the research which we adapted this stage-based participatory process from [8]. As mentioned in the Background section, the difference between our mental models with previous research is on what the mental model is about and it's richness. Our mental models can be considered limited, in that they only draw on how the users perceive a prospective AI system but not on how it works in details, in a real life context. Therefore unlike previous research our study cannot provide a detailed understanding of how and why the AI system works in practice [34]. Nonetheless, we successfully distilled different stakeholders insights on explanation of a AI medical support system, and formed them as very detailed mental models. We critically discussed the difference in understanding and perceptions of AI explanation needs, from an expert and non-expert perspective, we discussed issues of explanation modality and interaction, and combined expert and non-expert views in a target mental model. The resulting design guidelines were also contextualised to current practice and health regulations.

The explanation design guidelines we developed were based on critical reflections of Target Mental Model results. There is definitely room for improvements where we can incorporate other AI design guidelines or explanation recommendation to elaborate on our current guidelines. For example, from the Amershi et al.'s guidelines [35]; AI should show contextually relevant information (G4) and mitigate social biases (G6), we could add those guidelines to our guideline Information Delivery: Simple and General (EDG9). Another example, from [32]; suggesting that explanation should be contrastive, could contribute to our guideline for Interaction: Input Comparison (Visualisation). A follow on critical literature review would also help to verify and validate our proposed design guidelines.

5. Limitation and Future Works

There are several limitations of our study that should be addressed in future works. The stage-based participatory process is not complete. The final stage, which evaluates the developed prototype's effectiveness, has not carried out yet. We need to test whether the prototype has reached the design goals and wholly followed the design guidelines. To test if our prototype has reached the design goals, which is to design an explanation that can help user to make a considered trust judgements, we need to assess if there is any change in users' perception and their trust level. To measure the change in user's trust, we plan to use a quantitative measurement instrument [28] in a controlled experiment setting quantitatively measuring the extent to which each of the guidelines realised in the prototype contributed to enable considered trust judgements by non-expert users. In addition, we will conduct a lab-study and interview to get qualitative insight on both the prototype and the design guidelines.

We also acknowledge limitations within the research stages we had conducted. The participants involved were recruited from our personal network, which might limit the views variation in differing opinions. Finally, the explanation design guidelines proposed by this paper have not yet been evaluated, both in the guidelines' applicability across AI medical support systems variety; and the guidelines' clarity. Finally, the prototype we developed only delved into one type of modality, a graphic user interface. How the design guidelines implemented to an audio user interface or a conversational user interface also needs further exploration.

6. Conclusion

In this paper, we successfully applied a stage-based participatory design process to define future design guidelines for trustworthy AI healthcare system explanation interfaces for non-expert users. We developed an Expert Mental Model, User Mental Model, and Target Mental Model of AI medical support system's explanation. These mental models captured the needs and visions of the different stakeholders involved in a human-AI explanation process in a healthcare scenario. We used the developed Target Mental Model to inform a set of 14 explanation design guidelines for the development of trustworthy AI Healthcare System Explanation Interfaces, which specifically catered for non-expert users, while still taking into account medical experts' practice and AI experts' understanding. These guidelines emerged as an outcome of several stages of interviews, feedback from different types of stakeholders, thorough analysis of the current literature, and critical reflections on the insights obtained through the participatory process.

References

[1] D. Gefen, E. Karahanna, D. W. Straub, Trust and tam in online shopping: an integrated model, MIS quarterly 27 (2003) 51–90.

[2] GOV.UK, The future of healthcare: our vision for digital, data and technology in health and care, 2018. (Accessed on 02/10/2019).

[3] A. Alaszewski, Risk, trust and health, 2003.

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

[5] Z. C. Lipton, The doctor just won't accept that!, arXiv preprint arXiv:1711.08037 (2017).

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

[7] M. R. Cohen, J. L. Smetzer, Ismp medication error report analysis: Understanding human over-reliance on technology it's exelan, not exelon crash cart drug mix-up risk with entering a "test order", Hospital pharmacy 52 (2017) 7.

[8] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, H. Hussmann, Bringing transparency design into practice, in: 23rd international conference on intelligent user interfaces, 2018, pp. 211–223.

[9] B. Y. Lim, A. K. Dey, Design of an intelligible mobile context-aware application, in: Proceedings of the 13th international conference on human computer interaction with mobile devices and services, 2011, pp. 157–166.

[10] P. Pu, L. Chen, Trust building with explanation interfaces, in: Proceedings of the 11th international conference on Intelligent user interfaces, ACM, 2006, pp. 93–100.

[11] B. Y. Lim, A. K. Dey, Evaluating intelligibility usage and usefulness in a context-aware application, in: International Conference on Human-Computer Interaction, Springer, 2013, pp. 92–101.

[12] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, in: Advances in Neural Information Processing Systems, 2016, pp. 3504–3512.

[13] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and vascular neurology 2 (2017) 230–243.

[14] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: 2015 International Conference on Healthcare Informatics, IEEE, 2015, pp. 160–169.

[15] Z. Che, S. Purushotham, R. Khemani, Y. Liu, Distilling knowledge from deep networks with applications to healthcare domain, arXiv preprint arXiv:1512.03542 (2015).

[16] J. B. Lyons, G. G. Sadler, K. Koltai, H. Battiste, N. T. Ho, L. C. Hoffmann, D. Smith, W. Johnson, R. Shively, Shaping trust through transparent design: theoretical and experimental guidelines, in: Advances in human factors in robots and unmanned systems, Springer, 2017, pp. 127–136.

[17] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, B. Wielinga, The effects of transparency on trust in and acceptance of a content-based art recommender, User Modeling and User-adapted interaction 18 (2008) 455.

[18] C.-H. Tsai, P. Brusilovsky, Designing explanation interfaces for transparency and beyond., in: IUI Workshops, 2019.

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

[20] R. Buckman, How to break bad news: a guide for health care professionals, JHU Press, 1992.

[21] M. W. Rabow, S. J. Mcphee, Beyond breaking bad news: how to help patients who suffer., Western Journal of Medicine 171 (1999) 260.

[22] W. F. Baile, R. Buckman, R. Lenzi, G. Glober, E. A. Beale, A. P. Kudelka, Spikes—a six-step protocol for delivering bad news: application to the patient with cancer, The oncologist 5 (2000) 302–311.

[23] J. L. Szalma, G. S. Taylor, Individual differences in response to automation: The five factor model of personality., Journal of Experimental Psychology: Applied 17 (2011) 71.

[24] G. Ho, D. Wheatley, C. T. Scialfa, Age differences in trust and reliance of a medication management system, Interacting with Computers 17 (2005) 690–710.

[25] B. M. Yalaza M, İnan A, Male breast cancer, J Breast Health (2016).

[26] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative research in psychology 3 (2006) 77–101.

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

[28] R. Larasati, A. De Liddo, E. Motta, The effect of explanation styles on user's trust, in: Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with IUI 2020, 2020.

[29] P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplements 27 (1990) 247–266.

[30] D. J. Hilton, Conversational processes and causal explanation., Psychological Bulletin 107 (1990) 65.

[31] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gpdr, Harv. JL & Tech.

31 (2017) 841.
[32] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence (2018).
[33] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, 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.
[34] T. Kulesza, S. Stumpf, M. Burnett, I. Kwan, Tell me more?: the effects of mental model soundness on personalizing an intelligent agent, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 1–10.
[35] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al., Guidelines for human-ai interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 2019, p. 3.