Towards Participatory Design Spaces for Explainable AI Interfaces in Expert Domains

Henrik Mucha[1], Sebastian Robert[1], Rüdiger Breitschwerdt[2], Michael Fellmann[3]

[1] Fraunhofer IOSB, Fraunhoferstraße 1, 76131 Karlsruhe, Germany
[2] Wilhelm Büchner Hochschule, Hilpertstraße 31, 64295 Darmstadt, Germany
[3] Universität Rostock, Albert-Einstein-Straße 22, 18059 Rostock, Germany
henrik.mucha@iosb.fraunhofer.de
sebastian.robert@iosb.fraunhofer.de
ruediger.breitschwerdt@wb-fernstudium.de
michael.fellmann@uni-rostock.de

Abstract. In this position paper, we lay out an approach to use participatory and co-design methodology to explore how users perceive and interact with explanations of artificially intelligent decision support systems. We describe how we intend to construct bottom-up participatory design spaces to systematically inform the design of interactive explanations in Human-AI interaction.

Keywords: Explainable AI (XAI), Participatory Design, Design Spaces.

1 Introduction

Expert domains where decision making comes with high risk and is subject to liability need Explainable AI (XAI). In fact, these domains such as the medical one need representations of AI explanations that speak the language of their users. Hence, AI explanations should be designed in a human-centered way. Despite recent efforts [4-17], the interaction design of explanatory interfaces and the resulting behavior of the users of such intelligent interactive systems is currently under-researched, especially from a perspective of human-centered design. While the underlying technology of AI explanations is constantly and eagerly pushed forward, the way we present its result to the actual users has not been sufficiently studied yet. There exists a substantial body of research from the social sciences that provides valuable insights for designing human-centered AI explanations (see [5] for a detailed account). However, how to turn theory into practice, i.e., into working and testable interfaces has still to be explored. We propose to use established concepts from Design, Human-Computer Interaction, and Design Science Research to identify patterns and best practices for designing useful and usable [19] Human-AI interactions. Our goal is to establish a framework and a methodology to construct design spaces for XAI interfaces that are informed by user-generated design solutions. Design spaces are a well-established concept from Design Science Research [1,8,9] and a powerful tool to inform design decisions in technological development processes. We propose that by engaging in participatory design activities with the actual domain experts, i.e., the users of a system, we may better understand

how humans perceive AI recommendations and then derive design patterns that inform the designers of such systems in a human-centered way. In the remainder of this paper, we will lay out how this idea can work as a research approach.

2 Research Opportunity and Goal

2.1 Human-centered Explanations

In order to be considered capable to serve as a tool in expert domains, artificial intelligence (AI) must be interpretable, i.e., intelligent systems that derive conclusions and recommendations from machine learning processes must be able to explain their behavior in a way that is understandable for their users. Understandable or interpretable explanations are a prerequisite for the formation of trust which again is key for technology acceptance and adoption. Trust is gained by being honest and truthful, i.e., by making clear what led someone or something to do or say X and not Y. This is done through explanations. An explanation is usually defined as an answer to a why-question [5]. Explanations are interactive in nature; hence, interactivity is crucial for XAI [14]: There are two actors in explanations, the explainer and the explainee. In order for these two actors to interact with each other, a common ground must be established. Consequently, explanations must be designed to meet the needs of the individuals or the group of people they address. In other words, the explainer must speak the language of the explainee and therefore, AI explanations must be designed in a human-centered way.

2.2 Related Work

Given that we present a position paper, we only briefly touch upon research we regard as motivational for our own approach. Miller [5] provides a comprehensive study on explanations in AI. The paper summarizes insights from the social sciences and puts them into perspective for AI researchers. He focuses on the questions: What are explanations? Which models and theoretical frameworks are worth looking at? He sums up his major findings in four statements:

1. *Explanations are contrastive*
2. *Explanations are selected (in a biased manner)*
3. *Probabilities probably don't matter*
4. *Explanations are social*

Further, Miller identifies a research opportunity relevant to our work: "as far as the authors are aware, there are currently no studies that look at the cognitive biases of humans as a way to select explanations from a set of causes". Ribera and Lapedriza [12] follow up on Miller and proceed to think about what user-centered XAI could be. They suggest establishing a user model that differentiates between (1) developers and AI researchers, (2) domain experts and (3) lay users. Weld [18] sketches a vision for building interactive explanation systems, stressing the point that interactivity will be key for explanations to work and we add: Especially in expert domains. Ming et al. [6]

present a concept and actual interface showing how explanations may look like on a practical level and as interactive systems.

2.3 The Case for Participatory Design Spaces in Human-AI Interaction

The related work has three implications: First, there seems to be a consensus that AI explanations have to be interactive. Second, explanations have to speak the language of the respective user. Third, there is a valid need to investigate how actual users of (expert) systems perceive and understand the representations of AI explanations (i.e., user interfaces consisting of e.g., texts, diagrams, charts, illustrations, etc.).
We argue that this research gap can be addressed by exploring people's attitudes and behavior through co-design [13] or participatory design [1,16] , i.e., by going through design activities together [7,8]

In other words, and formulated as a research question: *How do different representations of AI explanations, i.e., different explanatory user interfaces, affect the decision making of (expert) users taking into account their cognitive biases?*

The research goal is to develop a framework and a methodology to establish user touchpoints along the XAI development process. These will take form as so-called participatory design spaces. The latter is a term we intend to coin in order to describe design spaces that are constructed from design patterns observed in user-generated design solutions. By formalizing user perceptions and opinions as design spaces, we should ultimately be able to provide valid guidelines for more informed design decisions in the new field of designing explanations in Human-AI Interaction.

2.4 Research Approach: Bottom-up Participatory Design Spaces

We propose to derive design patterns from user-generated design artifacts of representing (interactive) AI explanations to inform designers of Human-AI interactions. To this end, we collect user-generated design solutions by engaging with users in participatory design activities. This means that we create design artefacts together with users that seek to optimize existing representations of AI explanations from the subjective point of view of these users. This is typically done by asking three questions with regard to an explanation:

1. What do you see?
2. What is good and what is bad about it?
3. What would you do differently?

The latter is manifested and made concrete in the form of e.g., sketches. From this collection we will then derive patterns and further evaluate them e.g., through crowdsourcing and online-evaluation tools. Eventually, we can then summarize and formalize these patterns as design spaces.

Typically design spaces are constructed top-down, i.e., by reviewing relevant literature.[1] We propose to construct bottom-up design spaces by collecting data from engaging in a combination of behavioral experiment and design activity with users. In our paper on Workbook Sprints [8] we describe how this may take shape in practice.

In sum, we seek to establish a research framework that allows to describe evidence-based design heuristics for decision support systems focusing on health. Ideally, the methodology could also be applied to other expert domains due to the underlying systematic. However, this research approach comes with a number of challenges and opportunities, summarized in Table 1.

Table 1. Challenges and Opportunities of Participatory Design Spaces for XAI Representations

Challenges	Opportunities
• Differential Expertise [1] • Domain specific problems that are hard to evaluate, e.g., by lack of large cohorts • Availability of domain experts • Incentives for participation • Prototyping decision support • Managing and updating the design space(s)	• Empirically informed design decisions • Traceability of design decisions • Little effort for covered decision problems • Common ground for focused evaluation

3 Conclusion and Future Work

In this position paper we have motivated the need for participatory design spaces as an approach to inform the human-centered design of Human-AI Interaction focusing on XAI in expert domains. We highlighted existing work to emphasize the current relevance of this matter. We continued to describe how our research agenda can succeed on a methodological level and through experiments. Currently, our approach takes shape in an empirical study where we focus on user's perceptions and optimization proposals for Local Interpretable Model-Agnostic Explanations (LIME) for a general estimation task. The study has two parts, one co-design part in the form of remote workshops and the second as a large-scale online survey. We consider this endeavor a proof of concept to, if successful, be transferred to our main domain of interest, which is medical decision support. We like to bring the knowledge and expertise of HCI methodology to the discourse on XAI and follow the call for inter-disciplinary research efforts on machine behavior [11]. Thus, we hope to make a valuable contribution to inform the design of usable, useful, and trustworthy intelligent systems.

[1] Schaub's design space for privacy notices [13] is a good example for this process from a different research field.

References

1. Baumer, E. P.: Toward human-centered algorithm design. Big Data & Society, 4(2), (2017)
2. Bødker, S., Kyng, M.: Participatory design that matters—Facing the big issues. ACM Transactions on Computer-Human Interaction (TOCHI), 25(1), 1-31. (2018).
3. Hevner A., R., March, S. T., Park, J., Ram, S.: Design science in information systems research. MIS quarterly (2004), 75–105. (2004).
4. Liao, Q. V., Gruen, D., & Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-15). (2020).
5. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. (2019).
6. Ming, Y., Qu, H., & Bertini, E.: Rulematrix: Visualizing and understanding classifiers with rules. IEEE transactions on visualization and computer graphics, 25(1), 342-352.Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location. (2018).
7. Mucha, H. Robert, S.: Emerging Perspectives on Medical Decision Support: Co-Designing XAI. Fair & Responsible AI Workshop @ CHI2020. (2020).
8. Mucha, H., Mevißen, D., Robert, S., Jacobi, R., Meyer, K., Heusler, W., Arztmann, D.: Co-Design Futures for AI and Space: A Workbook Sprint. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-8). (2020).
9. Offermann, P., Levina, O., Schönherr, M., Bub, U.: Outline of a design science research process. In: Vaishnavi, V.K., Purao, S. (eds.) Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology (DESRIST'09). Malvern/ PA, USA, article 7. ACM, New York. (2009).
10. Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S.: A design science research methodology for information systems research. Journal of management information systems 24, 3 (2007), 45–77. (2007).
11. Rahwan, I., Cebrian, M., Obradovich, N., et al.: Machine behaviour. Nature, 568(7753), 477-486. (2019).
12. Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of user-centered explainable AI. In IUI Workshops. (2019)
13. Sanders, E. B. N., Stappers, P. J.: Co-creation and the new landscapes of design. Co-design, 4(1), 5-18. (2008).
14. Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., ... & Kersting, K.: Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, *2*(8), 476-486. (2020).
15. Schaub, F., Balebako, R., Durity, A. L., Cranor, L. F.: A design space for effective privacy notices. In Eleventh Symposium On Usable Privacy and Security (SOUPS). (2015).
16. Simonsen, J., Robertson, T. (Eds.).: Routledge international handbook of participatory design. Routledge. (2012).
17. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y.: Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-15). (2019).
18. Weld, D. S., Bansal, G.: The challenge of crafting intelligible intelligence. Communications of the ACM, 62(6), 70-79. (2019).
19. Xu, W.: Toward human-centered AI: a perspective from human-computer interaction. interactions, 26(4), 42-46. (2019).