Human-Centered Explainable AI: Creating Explanations that Address Stakeholder Needs

Anton Hummel[1,2]

[1]XITASO GmbH IT & Software Solutions, Austraße 35, 86153 Augsburg, Germany
[2]University Bayreuth, 95440 Bayreuth, Germany

Abstract

Artificial Intelligence (AI) in clinical decision support systems has considerable potential to improve medical care, but its application in clinical practice is still limited. A lack of transparency and human-oversight builds trust and acceptance barriers. It is often claimed that Explainable AI (XAI) is a promising method for overcoming those barriers. However, current XAI methods often fail to meet the diverse needs of different stakeholders. This research proposal aims to address this issue by developing human-centered explanations tailored to the individual requirements of various stakeholders. Therefore, my research employs a design science research approach, implementing and iteratively evaluating multiple promising XAI concepts, such as concept-based or glocal explanations. This approach will identify and refine the most promising methods based on stakeholder feedback. The results will contribute significantly to the development of human-centered explanations, advancing towards more responsible AI in clinical settings.

Keywords

Explainable AI, Human-Centered AI, Stakeholder Needs, Clinical Decision Support Systems

1. Context and Motivation

In a clinical setting, using artificial intelligence (AI)-based clinical decision support system (CDSS) promises to support physicians in making complex decisions and thus improve patient care. However, few of these AI systems are used in practice because they fail due to a lack of interpretability which can lead to trust barriers among physicians [1]. Explainable artificial intelligence (XAI) methods seem to be a promising tool to overcome those barriers. Furthermore, XAI could be pivotal to fostering human-ai-collaboration and thus fostering the human-ai-team performance [2].

Many researchers suggest putting the end-users and their needs at the center of attention when developing XAI [3]. Various groups involved in XAI have distinct interests, and recognizing these differences is important for developing methods that align with their specific needs (in the following called “desiderata”, as suggested by Langer et al. [3]). The success of an XAI method depends on the satisfaction of the stakeholders’ desiderata, which motivates explainability approaches that provide explanatory information to facilitate the stakeholders’ understanding.

Considering the varied levels of expertise among users— ranging from domain expertise to AI expertise— there is a distinct need for personalized explanations. For instance, a senior physician may require a less complex explanation in intensive care units than an assistant physician. At the same time, an AI developer necessitates a different form of explanation altogether. One common issue with current XAI methodologies is their failure to address the diverse desiderata of stakeholders, thereby failing to achieve their intended goals. The challenge lies in delivering the appropriate type of explanation to the right group of users, as suggested by Mohseni et al. [4].

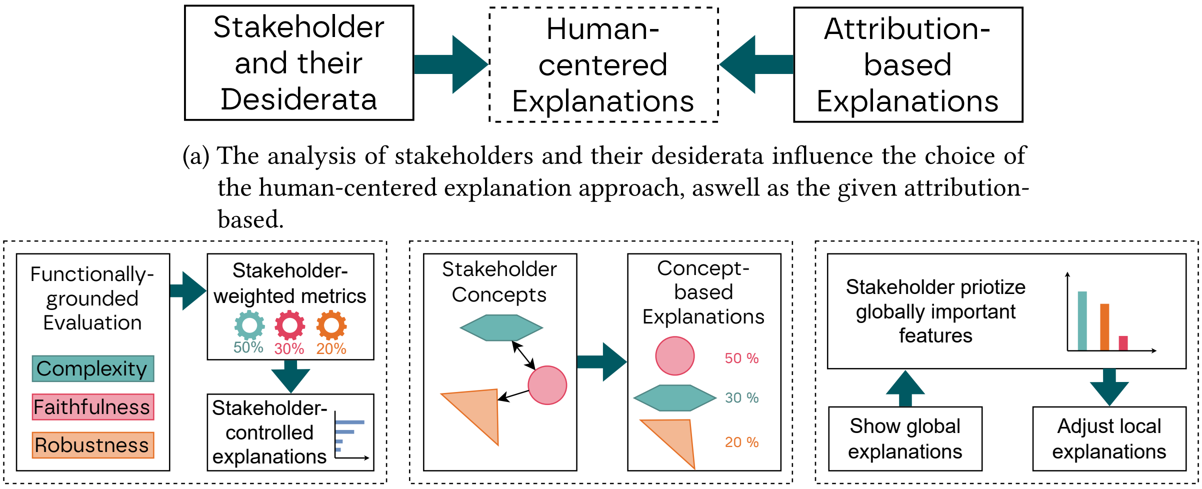
To address these complexities, developing human-centered explanations, also known as “user-centered” or “personalized” explanations, is recommended [5]. “Human-centered” explanations should meet the individual needs of various stakeholders, particularly in high-risk areas like healthcare [6, 7].

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

anton.hummel@xitaso.com (A. Hummel)
0009-0001-6945-3003 (A. Hummel)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

This can be achieved through user-centered design of XAI [8] or by choosing the most suitable technical XAI methodology.


Stakeholder and their Desiderata
Human-centered Explanations
Attribution-based Explanations
Functionally-grounded Evaluation
Complexity
Faithfulness
Robustness
Stakeholder-weighted metrics
50%
30%
20%
Stakeholder-controlled explanations
Stakeholder Concepts
Concept-based Explanations
50 %
30 %
20 %
Stakeholder priotize globally important features
Show global explanations
Adjust local explanations

(a) The analysis of stakeholders and their desiderata influence the choice of the human-centered explanation approach, aswell as the given attribution-based.

(b) User-controlled Explanations.

(c) Concept-based Explanations.

(d) Glocal Explanations.

Figure 1: The concept of human-centered explanations: All starts with the analysis of the stakeholders and their desiderata. These are used to combine them with attribution-based explanations to create human-centered explanations. Therefore, I propose three different promising approaches: User-controlled Explanations (Figure 1b), Concept-based Explanations (Figure 1c), and Glocal Explanations (Figure 1d).

Therefore, my research aims to address this problem in detail. My goal is to achieve AI systems that different stakeholders accept and create an appropriate, reliable interaction between the human and the AI. I will apply, assess, combine, and create new methods with promising approaches to creating human-centered explanations (see Figure 1). The different XAI approaches are described in Section 2. These methods will be applied and investigated in a clinical setup. I assume I can transfer the knowledge to other less-risky domains from the results obtained in this high-risk clinical domain. I am convinced that, especially in high-risk domains like healthcare, we must integrate the socio-technical perspective and the individual stakeholder desiderata to build trustworthy AI.

2. Key Related Work & Background

In this section, I provide a brief overview of related work that provides promising approaches for human-centered XAI and provide the background information for the key human-centered XAI approaches. Hereby, I also highlight papers that take the role of stakeholders and their desiderata in XAI into account. Table 1 summarizes and assigns the related work to main categories.

Before providing more background information about XAI approaches, I highlight the study from Calisto et al. [9] that explores the impact of personalized AI communication on clinical outcomes in breast cancer diagnosis. The study involved 52 varying expertise-level clinicians who used conventional and assertiveness-based AI communication styles to diagnose patient cases. Results showed that personalized AI communication significantly reduced diagnostic time and errors, particularly for less experienced clinicians, without compromising accuracy. The findings highlight the importance of adaptable AI communication to build trust, reduce cognitive load, and streamline clinical workflows, offering valuable insights for designing effective AI systems in high-stakes domains. My work aims to extend these insights in healthcare by evaluating the impact of different XAI concepts. The following paragraphs describe three promising XAI approaches and highlight related work for the concrete approach.

Optimizing Explanation by Explanation Properties. This approach leverages XAI evaluation methods to assess how well explanations meet various XAI goals. Functionally-grounded evaluation

Table 1
The related work can be mainly divided into research that incorporates stakeholders and their individual desiderata, research that provides human-centered XAI approaches and research that focuses on specific domains.

Paper	Stakeholder Desiderata	XAI approach			Domain
		Property-based	Concept-based	Glocal	
Calisto et al. [9]	✓				Healthcare
Decker et al. [10]	✓	✓			
Tadesse et al. [11]	✓	✓			
Das et al. [12]	✓		✓		
Achtibat et al. [13]			✓	✓	
My Work	✓	✓	✓	✓	Healthcare

methods serve as proxies for stakeholder desiderata, enabling the assessment of explanations based on different properties. This information can be used to directly optimize explanations towards those properties or aggregate explanations based on their properties. Users can personalize explanations by weighting different properties according to their preferences. Decker et al. [10] propose a method to enhance the reliability of feature attributions by combining multiple attribution methods to derive optimal convex combinations, improving robustness and faithfulness. Tadesse et al. [11] introduce a direct optimization approach that reliably produces explanations with optimal properties, allowing users to control trade-offs between different properties.

Concept-based Explanations. This approach shifts the focus from feature-level explanations to concept-level explanations, aiming to express AI predictions in semantically human-understandable concepts. This allows for personalization towards stakeholder desiderata through pre-explaining concept elicitation or interactive adjustments. Das et al. [12] introduce the State2Explanation (S2E) framework, which provides concept-based explanations for AI decision-making, enhancing both AI agent learning and end-user understanding. The framework defines criteria for concepts in sequential decision-making and learns a joint embedding model between state-action pairs and concept-based explanations, significantly improving user task performance.

Glocal Explanations. Combining local XAI methods (explaining individual predictions) and global XAI methods (explaining the whole model), this approach enriches explanations semantically, making them more human-centered. Known as "Glocal" explanations, this method matches global model concepts to individual predictions, bringing stakeholders into the loop of AI prediction. Achtibat et al. [13] introduce the Concept Relevance Propagation (CRP) approach, which combines local and global perspectives by extending Layer-wise Relevance Propagation (LRP).

3. Research Goal and Questions

My research aims to develop XAI methods that personalize explanations to various stakeholder desiderata. Hereby, the plan is to use evaluation methods and human-centered concepts that improve the explanations by satisfying the end-user desiderata. Using functionally-grounded evaluation methods creates an initial understanding of the explanation quality. Promising approaches like property-optimized explanations, concept-based explanations, or glocal explanation frameworks should be further investigated and improved. I will apply the different XAI methods in CDSS' with different stakeholders, to assess the methods applicability and human-ai-collaboration.

The research goal leads to the following general research question:

GRQ: *How should AI explanations be created to satisfy various stakeholder desiderata in clinical decision support systems?*

If related to current promising approaches, the following subquestions arise from this general research question:

- **RQ1:** How can functionally-grounded evaluation methods adapt explanations to individual stakeholder desiderata?
- **RQ2:** How can concept-based explanations enhance user understanding by aligning with individual user mental models?
- **RQ3:** How can glocal explanation methods be implemented to enrich local explanations with global model insights and thus satisfy stakeholder desiderata?

4. Research Approach

The primary objective of this research is to develop human-centered explanations for AI systems through the exploration of various Explainable AI (XAI) methodologies. To achieve this, a mixed-methods research approach will be employed, combining design science research (DSR) with qualitative and quantitative techniques to comprehensively address the research questions (see Figure 2).

As a core component of DSR, my research will involve the iterative design and development of XAI approaches as artifacts. The development will be iterative, with continuous feedback integration from stakeholders of varying expertise levels. Through cycles of *awareness of problem*, *suggestion*, *development*, and *evaluation*, the explanations will be progressively aligned with each stakeholder group's cognitive styles and domain-specific knowledge [14]. Each artifact iteration will be evaluated based on its alignment with stakeholder desiderata and its effectiveness in enhancing understanding. Hereby, with each cycle the number of selected XAI approaches will be reduced, to finally focus on one approach.

The single research methods, that are used in the process steps in each DSR cycle are described in the following and are divided into qualitative (see Section 4.1) and quantitative analysis (see Section 4.2) methods.

4.1. Qualitative Analysis

As part of the DSR method, the following qualitative analysis research methods are used:

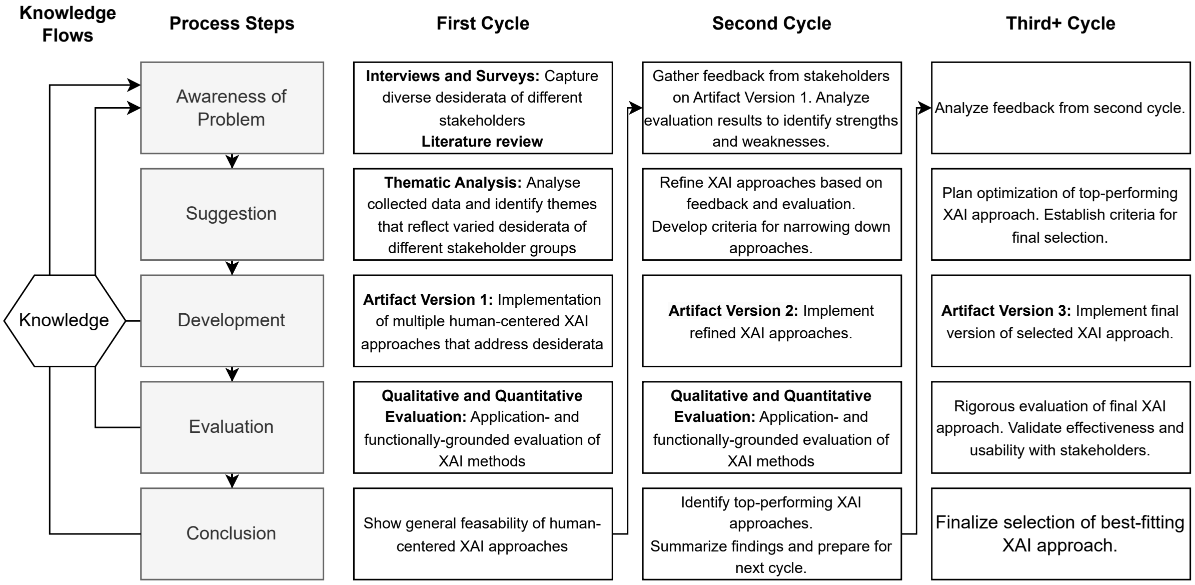
Stakeholder Analysis through Interviews and Surveys. I will be gathering interviews and surveys to capture the diverse desiderata of different stakeholder groups and qualitative data. These will target stakeholders with varying roles, such as decision-makers, developers, and domain experts while considering their respective levels of expertise. This approach will ensure a comprehensive understanding of stakeholder-specific desiderata and mental models regarding AI systems.

Thematic Analysis. The collected data will undergo thematic analysis, focusing on identifying themes that reflect the varied expectations and requirements of different stakeholder groups. This analysis will guide the development of explanation frameworks that are adaptable to different user expertise levels and cognitive styles.

4.2. Quantitative Analysis

Additionally, the following quantitative analysis research methods are used within the DSR cycles:

Experimental Design with Stratified Sampling. Controlled experiments will utilize stratified sampling to ensure representation across stakeholder groups and expertise levels. Participants will engage with various explanations (feature-based, concept-based, and glocal) to assess their impact on understanding, appropriate reliance, and satisfaction across diverse user profiles.



Process Steps	First Cycle	Second Cycle	Third+ Cycle
Awareness of Problem	**Interviews and Surveys:** Capture diverse desiderata of different stakeholders **Literature review**	Gather feedback from stakeholders on Artifact Version 1. Analyze evaluation results to identify strengths and weaknesses.	Analyze feedback from second cycle.
Suggestion	**Thematic Analysis:** Analyse collected data and identify themes that reflect varied desiderata of different stakeholder groups	Refine XAI approaches based on feedback and evaluation. Develop criteria for narrowing down approaches.	Plan optimization of top-performing XAI approach. Establish criteria for final selection.
Development	**Artifact Version 1:** Implementation of multiple human-centered XAI approaches that address desiderata	**Artifact Version 2:** Implement refined XAI approaches.	**Artifact Version 3:** Implement final version of selected XAI approach.
Evaluation	**Qualitative and Quantitative Evaluation:** Application- and functionally-grounded evaluation of XAI methods	**Qualitative and Quantitative Evaluation:** Application- and functionally-grounded evaluation of XAI methods	Rigorous evaluation of final XAI approach. Validate effectiveness and usability with stakeholders.
Conclusion	Show general feasability of human-centered XAI approaches	Identify top-performing XAI approaches. Summarize findings and prepare for next cycle.	Finalize selection of best-fitting XAI approach.

Figure 2: Research Approach: Mixed-methods with design science research as wrapper (adapted from Kuechler and Vaishnavi [14]).

Evaluation Metrics. Functionally-grounded evaluation metrics such as faithfulness, sensitivity, and complexity will be employed to measure explanation quality quantitatively.

4.3. Mixed-Methods Research Approach

The chosen research approach strategically incorporates the diverse needs of multiple stakeholders to answer the research questions and achieve the final contribution. Therefore, the research uses thematic analysis and stratified experimental design to customize explanation approaches to meet stakeholder groups' specific needs and preferences. It aims to assess how technical implementation affects explanation properties and user satisfaction based on stakeholders' expertise levels. Functionally-grounded evaluation methods are used to objectively assess various XAI approaches and optimize explanations to address diverse stakeholder requirements, directly tackling research question RQ1. Here, I assume that functionally-grounded evaluation methods are promising tools for improving the quality and effectiveness of explanations. Additionally, the research seeks to enhance interaction through human-centered explanations by developing frameworks tailored to stakeholders' knowledge and cognitive styles, exploring the alignment of explanations with user desiderata (RQ2), and balancing global and local perspectives (RQ3), ensuring technical robustness and adaptability to various stakeholders. I expect that concept-based explanations improve decision-making in human-AI teams more effectively than feature-based explanations and that glocal explanations provide a more comprehensive understanding of AI models than purely global or local explanations.

5. Preliminary Results

My previous research work has been mainly directed towards gaining domain-specific knowledge in the field of AI in healthcare and attaining a comprehensive understanding of XAI. The foundation of this research was an exhaustive literature review, that included an evaluation of various XAI methodologies to determine their applications and limitations. This examination incorporated an exploration of AI and XAI methods on healthcare applications, identifying critical areas where these technologies can improve clinical outcomes. Furthermore, I started to explore the role of XAI in building trust and improving collaboration between human practitioners and AI systems by participating workshops with clinicians and regular project meetings.

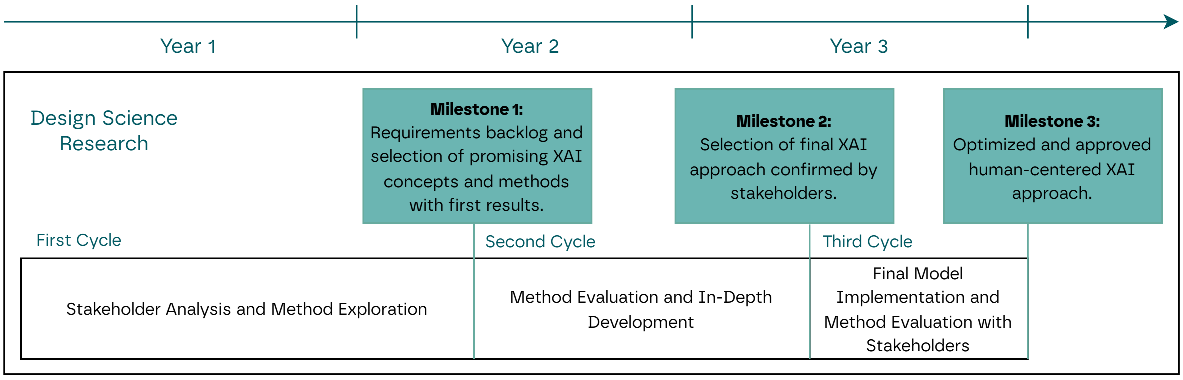

Year 1
Year 2
Year 3
Design Science Research
Milestone 1: Requirements backlog and selection of promising XAI concepts and methods with first results.
Milestone 2: Selection of final XAI approach confirmed by stakeholders.
Milestone 3: Optimized and approved human-centered XAI approach.
First Cycle
Second Cycle
Third Cycle
Stakeholder Analysis and Method Exploration
Method Evaluation and In-Depth Development
Final Model Implementation and Method Evaluation with Stakeholders

Figure 3: Research Roadmap: The previously described DSR approach (see Section 4) is applied to the next three years of my phd roadmap. Thereby, each cycle of the DSR has a clear focus and results in a milestone. From the the beginning of my research the focus shifts from the stakeholder analysis and the exploration of multiple XAI approaches towards the optimization of one XAI approach.

To translate theoretical insights into practical applications, I participated in stakeholder workshops with our research project partners from a university hospital that focused on developing CDSS' specific to the intensive care unit (ICU) environment and the collection of user requirements. These workshops were conducted with practicing physicians, ensuring that the AI systems align with clinical desiderata.

I engaged in a dataset exploration process, initiating research with the MIMIC-IV dataset and subsequently securing access to data from the ICU station at a local university hospital. This work contained data preprocessing, feature engineering, and implementing multivariate time series models to predict the patient's length of stay or mortality.

To further assess the efficacy of XAI methods, I systematically evaluated various attribution-based XAI methods, including SHAP [15], LIME [16] and others. This evaluation was based on a multivariate time series data model to assess explanation properties such as faithfulness, complexity, and sensitivity to effectively measure each XAI method's utility.

Moreover, my research extended into an interdisciplinary examination of the implications of XAI within the context of the EU AI Act, thereby contributing to socio-technical discussions and regulatory analysis. Through these foundational efforts, I have established a robust groundwork for my current research works to advance human-centered explanations of AI predictions in healthcare settings.

6. Research Roadmap and Final Contribution

The research roadmap which will guide the development and validation of human-centered explanations, by highlighting the milestones to address the main research question and its final contribution. As visualized in Figure 3, the DSR method is planned for three years.

Research Milestones. My research is segmented into three main milestones that align with the primary research goal of developing human-centered explanations. Each milestone marks an end of a DSR cycle and thus, concludes the goal of the respective cycle.

1. **First Milestone:** Completed initial development and evaluation of multiple XAI approaches, establishing a foundation for refinement and selection in subsequent cycles. This includes creating a requirements backlog and selecting promising XAI concepts and methods, with initial results to guide further development.
2. **Second Milestone:** Refined and evaluated XAI approaches, identifying promising methods for final selection. This involves confirming the selection of the final XAI approach with stakeholders, ensuring alignment with project goals and stakeholder needs.

3. **Third Milestone:** Successfully selected and optimized the best-fitting XAI approach for implementation. This milestone includes optimizing and approving a human-centered XAI approach, followed by final testing with stakeholders to ensure effectiveness and satisfaction.

Final Contribution. My research identifies and optimizes the most effective human-centered XAI approaches that address various stakeholder desiderata for CDSS'. Through a systematic DSR process, my work refines multiple XAI methods, evaluates their performance, and selects the best-fitting approach. My work not only advances the field of XAI by demonstrating the feasibility and effectiveness of human-centered solutions but also provides a comprehensive framework for future XAI development and implementation, ensuring enhanced interpretability and stakeholder satisfaction across diverse stakeholder groups. Thereby, the interdisciplinary exchange with socio-technical research disciplines is crucial.

7. Conclusion

My research uses a mixed-methods approach to enhance and implement novel XAI approaches that target the desiderata of various stakeholders in clinical decision support systems. Applying a DSR approach, the research incorporates attribution-based explanations to optimize and implement new human-centered XAI methods and evaluate them iteratively with various stakeholders. By this, my research significantly contributes to advancing human-centered XAI, thereby promoting the overall development of human-centered AI.

Acknowledgments

I express my sincere gratitude to my academic supervisor, Niklas Kühl, and my industrial supervisor, Jan-Philipp Steghöfer, for their exceptional guidance and support throughout my research. This research is sponsored by the German Federal Ministry of Education and Research under grant number 16SV9031.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-4o and Grammarly to: Grammar and spelling check, paraphrase and reword. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] E. Nasarian, R. Alizadehsani, U. R. Acharya, K.-L. Tsui, Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework, Information Fusion 108 (2024) 102412. URL: https://www.sciencedirect.com/science/article/pii/S1566253524001908. doi:10.1016/j.inffus.2024.102412.

[2] M. Schemmer, N. Kuehl, C. Benz, A. Bartos, G. Satzger, Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 410–422. URL: https://doi.org/10.1145/3581641.3584066. doi:10.1145/3581641.3584066.

[3] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research, Artificial Intelligence 296 (2021) 103473. URL: http://arxiv.org/abs/2102.07817. doi:10.1016/j.artint.2021.103473, arXiv:2102.07817 [cs].

[4] S. Mohseni, N. Zarei, E. D. Ragan, A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems, ACM Transactions on Interactive Intelligent Systems 11 (2021) 1–45. URL: https://dl.acm.org/doi/10.1145/3387166. doi:10.1145/3387166.

[5] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. URL: https://www.sciencedirect.com/science/article/pii/S1566253524000794. doi:10.1016/j.inffus.2024.102301.

[6] B. Finzel, Human-Centered Explanations: Lessons Learned from Image Classification for Medical and Clinical Decision Making, KI - Künstliche Intelligenz 38 (2024) 157–167. URL: https://doi.org/10.1007/s13218-024-00835-y. doi:10.1007/s13218-024-00835-y.

[7] N. Kühl, Human-centric artificial intelligence: The road ahead., Transfer: Zeitschrift für Kommunikation & Markenmanagement 70 (2024).

[8] I.-C. Jung, K. Schuler, M. Zerlik, S. Grummt, M. Sedlmayr, B. Sedlmayr, Overview of basic design recommendations for user-centered explanation interfaces for AI-based clinical decision support systems: A scoping review, DIGITAL HEALTH 11 (2025) 20552076241308298. URL: https://journals.sagepub.com/doi/10.1177/20552076241308298. doi:10.1177/20552076241308298.

[9] F. M. Calisto, J. M. Abrantes, C. Santiago, N. J. Nunes, J. C. Nascimento, Personalized explanations for clinician-AI interaction in breast imaging diagnosis by adapting communication to expertise levels, International Journal of Human-Computer Studies 197 (2025) 103444. URL: https://linkinghub.elsevier.com/retrieve/pii/S1071581925000011. doi:10.1016/j.ijhcs.2025.103444.

[10] T. Decker, A. R. Bhattarai, J. Gu, V. Tresp, F. Buettner, Provably Better Explanations with Optimized Aggregation of Feature Attributions, 2024. URL: http://arxiv.org/abs/2406.05090. doi:10.48550/arXiv.2406.05090, arXiv:2406.05090 [cs].

[11] H. B. Tadesse, A. Hüyük, W. Pan, F. Doshi-Velez, Directly Optimizing Explanations for Desired Properties, 2024. URL: http://arxiv.org/abs/2410.23880, arXiv:2410.23880.

[12] D. Das, S. Chernova, B. Kim, State2Explanation: Concept-Based Explanations to Benefit Agent Learning and User Understanding, Advances in Neural Information Processing Systems 36 (2023) 67156–67182. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/d4387c37b3b06e55f86eccdb8cd1f829-Abstract-Conference.html.

[13] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From attribution maps to human-understandable explanations through Concept Relevance Propagation, Nature Machine Intelligence 5 (2023) 1006–1019. URL: https://www.nature.com/articles/s42256-023-00711-8. doi:10.1038/s42256-023-00711-8, publisher: Nature Publishing Group.

[14] W. Kuechler, V. Vaishnavi, A Framework for Theory Development in Design Science Research: Multiple Perspectives, Journal of the Association for Information Systems 13 (2012). URL: https://aisel.aisnet.org/jais/vol13/iss6/3. doi:10.17705/1jais.00300.

[15] S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[16] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016, pp. 1135–1144. URL: https://dl.acm.org/doi/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.