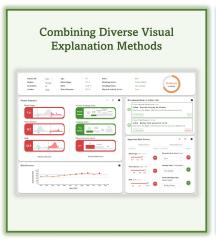
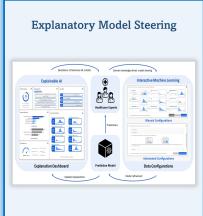
Towards Directive Explanations: Crafting Explainable Al Systems for Actionable Human-Al Interactions

Aditya Bhattacharya aditya.bhattacharya@kuleuven.be KU Leuven Leuven, Belgium





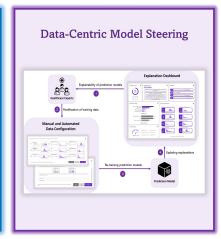


Figure 1: This research focuses on crafting Explainable AI systems for actionable human-AI interactions. Considering the current progress, the author conducted three research studies: (1) Combining diverse visual XAI methods in an explanation dashboard, (2) Explanatory model steering with domain experts through a healthcare-focused XAI system, and (3) Data-centric model steering through manual and automated configuration of training data.

ABSTRACT

With Artificial Intelligence (AI) becoming ubiquitous in every application domain, the need for explanations is paramount to enhance transparency and trust among non-technical users. Despite the potential shown by Explainable AI (XAI) for enhancing understanding of complex AI systems, most XAI methods are designed for technical AI experts rather than non-technical consumers. Consequently, such explanations are overwhelmingly complex and seldom guide users in achieving their desired predicted outcomes. This paper presents ongoing research for crafting XAI systems tailored to guide users in achieving desired outcomes through improved human-AI interactions. This paper highlights the research objectives and methods, key takeaways and implications learned from user studies. It outlines open questions and challenges for enhanced human-AI collaboration, which the author aims to address in future work.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

© 2024 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-0331-7/24/05.

https://doi.org/10.1145/3613905.3638177

CCS CONCEPTS

• Human-centered computing → Human computer interaction (HCI); Interaction design; • Computing methodologies → Artificial intelligence.

KEYWORDS

Explainable AI, Interactive Machine Learning, Explanatory Interactive Learning, Domain-Expert-AI Collaboration

ACM Reference Format:

Aditya Bhattacharya. 2024. Towards Directive Explanations: Crafting Explainable AI Systems for Actionable Human-AI Interactions. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613905.3638177

1 RESEARCH BACKGROUND AND MOTIVATION

The utilisation of Artificial Intelligence (AI) systems has grown significantly in the past few years across diverse domains such as medical [13, 14, 37], finance [10, 12, 15], legal [44, 51, 54] and others [2, 6, 36]. Despite the success of AI systems across various applications, the "black-box" nature of AI models has raised several concerns related to lack of transparency [6, 29, 34] and appropriate trust [27, 47], particularly when predicted outcomes are biased, unfair, incorrect or misguiding [11, 22, 33]. Consequently, the field

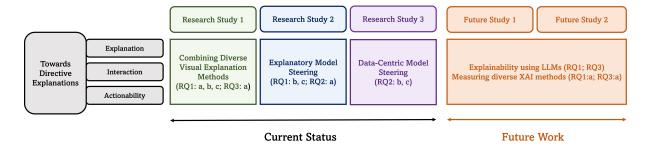


Figure 2: Overview of the current research progress and planned next steps to address research questions

of Explainable AI (XAI) has gained much focus from AI practitioners as explanations can potentially help users to develop a clear mental model of such complex algorithms, eventually enhancing their trust for the adoption of AI systems [2, 18].

However, popularly adopted XAI methods [2, 6, 20, 30, 38] are predominantly designed for AI experts, neglecting the needs of a broader community of non-technical consumers of AI [10, 52]. Moreover, prior works have questioned the efficacy of current one-off static explanations for non-expert users, i.e. users who might be experts in a particular application domain but may lack technical AI knowledge [26]. Hence, there is a need for tailoring explanations for non-expert users and making XAI methods more useful, understandable, actionable and trustworthy through user-centric design principles [31].

Additionally, prior works have emphasised the need for directive explanations to non-expert users for enhanced human-AI collaborations [42, 45]. Directive explanations can assist users in obtaining their desired predictions by interacting with the prediction system for an actionable recourse [45]. Instead of static explanations, non-expert users have expressed the need for more interactive explanations to foster understanding and interpretation [1, 21, 24] of complex AI systems. Such explanations can enable them to understand why a certain prediction is generated and also indicate how to obtain their desired predictions without any intervention from AI experts [23, 29, 39, 46, 53]. Along with interactive explanations, prior research in Interactive Machine Learning (IML) has studied the collaborative interactions of users with AI systems and their vital role in model development, fine-tuning, debugging and evaluation [4, 16, 47, 50]. For instance, researchers of EluciDebug [24] have found that explanation-driven interactions can facilitate users to improve prediction models by helping them to identify bugs and issues and enhance their overall mental model of the system.

However, despite various explanation-driven interaction methods discussed in the literature [8, 24, 25, 49, 50], researchers have identified many challenges when non-technical consumers of AI have been involved in the AI prediction process [26, 41]. For example, non-expert users have expressed the need for more active involvement with AI models through interactive explanations of the underlying data rather than static explanations of the prediction algorithms [26].

Moreover, recent works have emphasised the need for Data-Centric AI (DCAI) [32] approaches for model improvement by improving the quality of the training data for building more reliable and trustworthy prediction models [5]. However, prior research has shown that improving the quality of training data is not straightforward without thorough domain knowledge [17]. Domain experts can identify certain types of biases in data to improve the training data quality. For example, instead of model developers, healthcare experts such as doctors or nurses have a better understanding of patient's medical records, enabling them to identify biases and limitations in training data of prediction models for healthcare applications. Consequently, it is vital to capture the domain knowledge of domain experts through interactive explainability systems that give them more control over the training data [4, 19, 49].

2 RESEARCH OBJECTIVES AND METHODS

The author's research focuses on crafting XAI systems that provide directive explanations for non-expert users and allow them to collaborate with the system for more reliable and useful predictions. The following are the primary research objectives:

- **O1.** *Explanation:* The first objective is to compare the utility of different tailored explanation methods to elucidate predictions generated by AI models for non-expert users.
- **O2**. *Interaction:* Second, the author aims to design interaction mechanisms that can be combined with different explanation methods as a basis to develop more powerful XAI systems. The effect of providing different levels of control over explanations, training data and prediction models on user understanding and trust will be investigated.
- **O3**. *Actionability:* Finally, the author will research approaches to guide users in obtaining their desired predictions through interactive explanations and data-centric approaches.

The following high-level research questions further complement the preceding research objectives:

- **RQ1**: What is the effectiveness of different tailored explanation methods for explaining AI models to non-expert users?
 - (a) Are specific explanation methods better than other methods in explaining the outcomes of models?
 - **(b)** How can we combine explanation methods to develop more powerful explanation interfaces?
 - (c) How can explanation methods be tailored to the needs of non-expert users?

RQ2: How can interactive explanations facilitate non-expert users in model steering through data-centric approaches?

- (a) How can different types of explanations assist non-expert users in model steering?
- (b) How can non-expert users use different data-centric model steering approaches to improve prediction models?
- (c) What is the impact of different levels of control for model steering?
- **RQ3**: How can we tailor explanations to generate actionable insights?
 - (a) What is the effectiveness of different explanation methods for actionable insights?
 - (b) How can we design interactions to enhance the actionability of explanations?

To address the above research questions, XAI systems are designed and developed following user-centric principles. Initially, exploratory studies are conducted to collect the needs of the target users for specific application areas. Then, in multiple iterations, low-fidelity prototypes are designed for initial evaluation with the target users. Then, high-fidelity prototypes are developed and evaluated through quantitative, qualitative or mixed-methods user studies. Finally, the data collected from these studies are analysed to extract valuable insights. These insights are instrumental in addressing the research questions and formulating design implications and guidelines for future research.

3 CURRENT STATUS

This section describes the current progress of the author's research. To fulfil the research objectives, three research studies with two healthcare-focused interactive XAI systems have been conducted to date. The author has conducted six user studies in total involving healthcare experts and patients by following a user-centric design and development process.

Research Study 1 - In his first research study [7], the author designed and developed an explanation dashboard for monitoring the risk of diabetes onset. This interactive dashboard provides explanations for a diabetes risk prediction model by combining data-centric, feature importance and example-based local explanations. The dashboard aims to assist healthcare experts like nurses and physicians in monitoring patients and recommending measures to minimise their risk of type-2 diabetes. First, an exploratory user study was conducted through focus group discussion followed by a co-design session with 4 registered nurses from the Community Healthcare Centre dr. Adolf Drolc in Maribor, Slovenia, to collect the user requirements and understand their challenges in monitoring diabetic patients. Next, a qualitative user study was conducted through individual interviews with 11 healthcare experts to evaluate our low-fidelity click-through prototypical dashboard. Finally, a mixed-methods user study was conducted with 45 healthcare experts and 51 diabetic patients to evaluate our high-fidelity web application prototypical explanation dashboard. Through these user studies, the author measured the usefulness, understandability, actionability and trustworthiness of different types of explanations included in the dashboard. The results underscored the importance of the author's representation of data-centric explanations that presented local explanations with a global overview of feature importance and example-based explanations. However, both healthcare experts and patients highlighted the importance of combining the

different types of explanations for recommending risk-mitigating actions, indicating that any limitation of an individual explanation method can be complemented by other methods.

The following are the key contributions of this study:

- It was found that combining different XAI methods is essential to address different dimensions of explainability for a holistic explanation of predictive models. Including only one or a few types of XAI methods can only provide partial explainability.
- The perspectives of healthcare experts and patients were collected to compare the different types of explanations in terms of understandability, usefulness, actionability and trustworthiness. This research collected insights about the perceived importance of directive data-centric explanations within healthcare XAI systems, illuminating a relatively unexplored subject in the XAI literature.
- Design implications for tailoring visually directive explanations for healthcare experts were presented considering the insights captured from this study.

Research Study 2 - In the second research study [8, 9], the author designed and developed an Explanatory Model Steering (EX-MOS) system for healthcare experts. The EXMOS system provides different types of global explanations to support healthcare experts in improving prediction models through manual and automated data configurations. To evaluate this system, two between-subject user studies were conducted: one quantitative study with 70 healthcare experts and another qualitative study with 30 healthcare experts. The impact of diffrent types of global explanations on trust, understandability and model improvement was measured in this research study. The results highlighted the inefficiency of global model-centric explanations for guiding users during the configuration of the training data. Despite the benefits of data-centric global explanations in helping users comprehend the post-configuration system changes, it was found that a hybrid fusion of both datacentric and model-centric global explanations was most effective for steering prediction models.

The following are the key contributions of this study:

- An XAI system that uses different types of datacentric and model-centric global explanations was designed and implemented to assist healthcare experts in sharing their domain knowledge.
- Perspectives of 100 healthcare experts were collected to highlight the importance of multifaceted explanations for a holistic explainability of the system.
- Guidelines for the design and implementation of explanatory model steering systems for healthcare applications were presented considering the insights collected from this study.

Research Study 3 - In the latest research study, the author expanded his research with the EXMOS system to investigate the impact of different levels of control for model steering through datacentric approaches. More specifically, the latest research explored the effectiveness of manual and automated data configuration methods in healthcare-focused XAI systems. A between-subject mixedmethods user study was conducted with 74 healthcare experts to evaluate the effectiveness of different data configuration methods across multiple measures such as understandability, trust, model improvement, explanation goodness and satisfaction, feedback importance and usability and system interactions. It was found that the study participants could significantly improve the performance of the prediction model using the manual configurations instead of the automated method. The study findings highlighted the necessity of a higher involvement of domain experts for enhanced human-AI collaboration.

The following are the key contributions of this study:

- A user-centric design and implementation of different manual and automated data configuration approaches that enabled domain experts to share their domain knowledge and improve prediction models was presented.
- Detailed insights comparing the significance of manual and automated data configuration methods across multiple measures from the perspective of healthcare experts were presented. The results highlight the necessity of granting more control to domain experts for model improvement.
- Design guidelines for manual and automated datacentric collaborative approaches for human-AI interactions were presented considering the insights collected from the user study.

4 FUTURE WORK

In his previous research studies, the author has mostly focused on different types of visual explanations designed to elucidate prediction models built on structured datasets. However, with the recent advancements in Generative AI [48] and Large Language Models (LLMs) [55], he plans to research on designing XAI systems, which include AI models built on unstructured data such as text or images. The author aims to follow a similar research procedure as our previous studies to explore different approaches for actionable and human-friendly explanations. Furthermore, his current research has highlighted the need for more objective measures to evaluate different XAI methods instead of subjective metrics. Thus, as next steps of his current research, the author wants to focus on the following areas:

Explainability using LLMs - The author hypothesises that LLMs can be used to generate more natural and human-friendly explanations and foster more natural interaction approaches with AI models. More specifically, the author wants to include LLM-based chatbots [26, 40, 43] and allow non-expert users to interact with prediction models through conversation-based explanations and

model steering approaches. He proposes to conduct randomised control experiments with non-expert users to measure the effectiveness of LLMs in generating more actionable, causal and contextual explanations. The author would also like to collect more qualitative data to understand the benefits and pitfalls of using LLMs in explanatory model steering.

Measuring diverse XAI methods - More recently, many researchers have questioned the efficacy of current quantitative usercentric metrics to compare different XAI methods [3, 28, 35]. Most of these metrics are highly subjective to the research participants and do not provide a generalised evaluation of different types of XAI methods for other users or applications. Thus, the author's current research studies have raised a vital open question: "how to measure diverse XAI methods in a unified and fair approach as each method elucidates different aspects of an AI system?" Therefore, he proposes to conduct a systematic review of the research literature to compare the benefits and drawbacks of subjective user-centric XAI metrics and objective XAI metrics. After this systematic review, he aims to formulate a unified and fair evaluation approach for inter-system and intra-system comparisons of XAI methods from this research and evaluate this novel approach with XAI researchers to collect their feedback.

ACKNOWLEDGMENTS

The author would like to thank his supervisor, Prof. Katrien Verbert, for her constant support throughout his PhD. This work is supported by research grants: Research Foundation–Flanders (FWO, grant G0A3319N, G0A4923N, G067721N) and KU Leuven Internal Funds (grant C14/21/072). He would like to extend his gratitude to his collaborators Dr. Simone Stumpf (University of Glasgow, UK), Dr. Gregor Stiglic (University of Maribor, Slovenia) and the Augment HCI group, KU Leuven, for their support during this research work.

REFERENCES

- Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156
- [2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6 (2018), 52138– 52160
- [3] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion 99 (2023), 101805. https://doi.org/10.1016/j.inffus.2023.101805
- [4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. AI Magazine 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513
- [5] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3445736
- [6] Aditya Bhattacharya. 2022. Applied Machine Learning Explainability Techniques. In Applied Machine Learning Explainability Techniques. Packt Publishing, Birmingham, UK. https://www.packtpub.com/product/applied-machine-learning-explainability-techniques/9781803246154
- [7] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In Proceedings of the 28th International Conference on Intelligent User

- Interfaces (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 204–219. https://doi.org/10.1145/3581641.3584075
- [8] Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2023. Lessons Learned from EXMOS User Studies: A Technical Report Summarizing Key Takeaways from User Studies Conducted to Evaluate The EXMOS Platform. arXiv:2310.02063 [cs.LG]
- [9] Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2024. EXMOS: Explanatory Model Steering Through Multifaceted Explanations and Data Configurations. arXiv:2402.00491 (2024). https://doi.org/10.48550/arXiv.2402.00491 arXiv:2402.00491 [cs.AI]
- [10] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In 27th International Conference on Intelligent User Interfaces. ACM, Helsinki Finland, 807–819. https://doi.org/10.1145/3490099.3511139
- [11] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:1802.07228 [cs.AI]
- [12] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2021. Explainable Machine Learning in Credit Risk Management. Computational Economics 57 (01 2021). https://doi.org/10.1007/s10614-020-10042-0
- [13] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Sydney, NSW, Australia) (KDD '15). Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613
- [14] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (Feb. 2017), 115–118.
- [15] Gerald Fahner. 2018. Developing Transparent Credit Risk Scorecards More Effectively: An Explainable Artificial Intelligence Approach.
- [16] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In Proceedings of the 8th International Conference on Intelligent User Interfaces (Miami, Florida, USA) (IUI '03). Association for Computing Machinery, New York, NY, USA, 39–45. https://doi.org/10.1145/604045.604056
- [17] Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe. 2020. Fair AI. Business & information systems engineering 62, 4 (2020), 379–384.
- [18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5, Article 93 (aug 2018), 42 pages. https://doi.org/10.1145/3236009
- [19] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131. https://doi.org/10.1007/s40708-016-0042-6
- [20] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. 2021. Explainable Artificial Intelligence Approaches: A Survey. (2021). https://doi.org/10.48550/ARXIV.2101.09429
- [21] Yucheng Jin, Nava Tintarev, Nyi Nyi Htun, and Katrien Verbert. 2020. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction* 30 (04 2020). https://doi.org/10.1007/s11257-019-09247-2
- [22] Michael I. Jordan. 2019. Artificial Intelligence—The Revolution Hasn't Happened Yet. Harvard Data Science Review 1, 1 (jul 1 2019). https://hdsr.mitpress.mit.edu/pub/wot7mkc1.
- [23] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. , 5686-5697 pages. https://doi.org/10.1145/2858036.2858529
- [24] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM, Atlanta Georgia USA, 126–137. https://doi.org/10.1145/2678025.2701399
- [25] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In 2010 IEEE Symposium on Visual Languages and Human-Centric Computing. IEEE, Leganes, Madrid, Spain, 41–48. https://doi.org/10.1109/VLHCC.2010.15
- [26] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. arXiv:2202.01875 [cs.LG]
- [27] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. Computers in Human Behavior 139

- (2023), 107539. https://doi.org/10.1016/j.chb.2022.107539
- [28] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv:2110.10790 [cs.AI]
- [29] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems (CHI '09). Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023
- [30] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs.AI]
- [31] Winston Maxwell and Bruno Dumas. 2023. Meaningful XAI Based on User-Centric Design Methodology. arXiv:2308.13228 [cs.HC]
- [32] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. 2023. DataPerf: Benchmarks for Data-Centric AI Development. arXiv:2207.10062 [cs.LG]
- [33] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607
- [34] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. https://doi.org/10.48550/ARXIV.1706.07269
- [35] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Comput. Surv. 55, 13s, Article 295 (jul 2023), 42 pages. https://doi.org/10.1145/3583558
- [36] I.K. Nti, A.F. Adekoya, B.A. Weyori, et al. 2022. Applications of artificial intelligence in engineering and manufacturing: a systematic review. J Intell Manuf 33 (2022), 1581–1601. https://doi.org/10.1007/s10845-021-01771-6
- [37] Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly. 2020. Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain.. In AICS. 169–180.
- [38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]
- [39] Abdallah Abbas Ritvij Singh, Edward Korot Sara Beqiri, Peter Woodward-Court Robbert Struyven, and Pearse Keane. 2021. Exploring the What-If-Tool as a solution for machine learning explainability in clinical practice. *Invest. Ophthalmol.* Vis. Sci. 2021;62(8):79. (2021).
- [40] Marwa Sallam. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 11, 6 (19 3 2023), 887. https://doi.org/10.3390/ healthcare11060887
- [41] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* 2 (08 2020), 476–486. https://doi.org/10.1038/s42256-020-0212-3
- [42] Ronal Singh, Paul Dourish, Piers Howe, Tim Miller, Liz Sonenberg, Eduardo Velloso, and Frank Vetere. 2021. Directive Explanations for Actionable Explainability in Machine Learning Applications. (2021). https://doi.org/10.48550/ARXIV.2102.
- [43] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. Nature Machine Intelligence (27 Jul 2023). https://doi.org/10.1038/s42256-023-00692-8
- [44] Eduardo Soares and Plamen Angelov. 2019. Fair-by-design explainable models for prediction of recidivism. arXiv:1910.02043 [stat.ML]
- [45] Alexander Spangher and Berk Ustun. 2018. Actionable Recourse in Linear Classification. https://doi.org/10.1145/3287560.3287566
- [46] Thilo Spinner, Udo Schlegel, Hanna Hauptmann, and Mennatallah El-Assady. 2019. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. (07 2019).
- [47] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. Int. Journal of Human-Computer Studies 67, 8 (2009), 639–662.
- [48] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-Based Design. In 27th International Conference on Intelligent User Interfaces (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 212–228. https://doi.org/10. 1145/3490099.3511119

- [49] Stefano Teso, Öznur Alkan, Wolfang Stammer, and Elizabeth Daly. 2022. Leveraging Explanations in Interactive Machine Learning: An Overview. http://arxiv.org/abs/2207.14526 arXiv:2207.14526 [cs].
- [50] Stefano Teso and Kristian Kersting. 2019. Explanatory Interactive Machine Learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (Honolulu, HI, USA) (AIES '19). Association for Computing Machinery, New York, NY, USA, 239–245. https://doi.org/10.1145/3306618.3314293
- [51] Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. 2022. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. arXiv:2005.04176 [stat.ML]
- [52] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk,
- 1–15. https://doi.org/10.1145/3290605.3300831
- [53] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics PP* (08 2019), 1–1. https://doi.org/10.1109/TVCG.2019.2934619
- [54] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2016. Interpretable Classification Models for Recidivism Prediction. Journal of the Royal Statistical Society Series A: Statistics in Society 180, 3 (09 2016), 689–722. https://doi.org/10.1111/rssa.12227 arXiv:https://academic.oup.com/jrsssa/article-pdf/180/3/689/49430770/jrsssa_180_3_689.pdf
- [55] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for Large Language Models: A Survey. arXiv:2309.01029 [cs.CL]