INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH

ALIGNING EXPLAINABLE AI WITH USER NEEDS: A PROPOSAL FOR A PREFERENCE-AWARE EXPLANATION FUNCTION

Dr. Lucas M. Hoffmann
Human-Centered AI Research Lab, University of Tübingen, Tübingen, Germany

Dr. Aya El-Masry
Faculty of Information Systems, American University in Cairo, Cairo, Egypt

Published Date: 09 December 2024 // Page no.:- 1-7

ABSTRACT

The rapid advancement and widespread deployment of Artificial Intelligence (AI) models, particularly deep neural networks, have led to remarkable successes across diverse domains. However, the inherent "black-box" nature of many high-performing models poses significant challenges, including a lack of transparency, trust, and accountability. Explainable Artificial Intelligence (XAI) aims to bridge this gap by making AI decisions understandable to humans. While numerous XAI methods have emerged, a crucial aspect often overlooked is the diverse and context-dependent nature of user preferences for explanations. A generic explanation may not suffice for all users or all decision-making scenarios. This article proposes a conceptual framework centered around a mapping function designed to adapt explanation generation to specific user profiles, contextual factors, and AI model characteristics. We review the landscape of XAI, analyze the varying needs of stakeholders, and detail the proposed mapping function's inputs, logic, and outputs. This user-centric approach promises to enhance the utility, trustworthiness, and effectiveness of XAI systems, fostering broader adoption and responsible AI deployment. We conclude by outlining key challenges and future research directions necessary to realize this vision.

Keywords: Explainable artificial intelligence (XAI); user-centered AI; preference-aware explanation; human-AI interaction; personalized explanations; interpretability; decision transparency; adaptive explanation systems; user trust in AI; AI explainability frameworks.

INTRODUCTION

Artificial Intelligence (AI) has permeated nearly every facet of modern life, from healthcare diagnostics to financial trading and autonomous systems. While the predictive power of complex machine learning models, especially deep learning networks, has reached unprecedented levels, their inherent opacity often renders them "black boxes" [3, 39, 47, 50, 91, 107]. This lack of transparency presents significant challenges, including difficulty in debugging model errors [1, 6, 56, 70], ensuring fairness [20, 32, 61, 88, 106], building user trust [19, 29, 59], complying with regulatory requirements [52, 72], and facilitating human understanding and oversight [13, 55]. The absence of explainability can lead to reduced adoption, misapplication, and even ethical dilemmas [9, 10, 45, 62].

To address these critical concerns, the field of Explainable Artificial Intelligence (XAI) has emerged [36, 37, 55]. XAI aims to develop methods and techniques that enable humans to comprehend, trust, and effectively manage AI systems [13]. This involves providing insights into why an AI model made a particular decision, how it arrived at a prediction, or what factors influenced its output. The motivations for XAI are diverse, ranging from debugging and auditing to promoting human learning and compliance [13, 55]. In domains such as healthcare, explainability is crucial for clinical validation and informed decision-making [9, 10, 14, 45, 64, 79, 85]. Similarly, in finance and credit scoring, understanding AI decisions is vital for accountability and regulatory adherence [33, 42, 44, 98].

Despite the proliferation of XAI methods (e.g., saliency maps [4, 5, 15, 28, 116, 117], Layer-wise Relevance Propagation (LRP) [2, 15, 67, 115], LIME [111], SHAP [96], counterfactual explanations [25, 54, 105]), a critical gap persists: explanations are often designed as a "one-size-fits-all" solution, neglecting the heterogeneity of human users and their varying needs [23, 48, 62, 82, 86, 100]. What a developer needs to debug a model differs significantly from what an end-user needs to make a critical decision, or what a regulator requires for compliance. The "problem of ambiguity in XAI" highlights that "explanation" itself is not a monolithic concept [48], and different stakeholders require different kinds of explanations [23, 45, 62, 82, 86].

This article addresses this fundamental challenge by proposing a mapping function as a conceptual framework for generating user-centric explanations in XAI. This function would dynamically select and tailor explanations based on a comprehensive understanding of the user's profile, the specific context of the AI decision, and the

characteristics of the AI model itself. By synthesizing current XAI methodologies with insights from human factors research, we aim to lay the groundwork for more effective, personalized, and ultimately, more impactful explainable AI systems. The remainder of this article is structured as follows: Section 2 reviews related work in XAI and human factors. Section 3 details the proposed mapping function framework. Section 4 discusses key challenges and future research directions, followed by concluding remarks in Section 5.

2. Related Work

The landscape of Explainable AI is rich and diverse, spanning various techniques and theoretical underpinnings. This section provides a concise overview of the prominent XAI methods and crucially, highlights the growing emphasis on human factors in evaluating their effectiveness.

2.1 Defining and Categorizing Explainable AI

The concept of "explainability" in AI is itself complex and subject to multiple interpretations [36, 48, 62, 92, 107, 109]. Generally, XAI aims to make AI models more transparent, interpretable, and understandable [13]. Transparency refers to how the model works internally, interpretability to the degree to which a human can understand the cause and effect of a model's input and output, and understandability to the cognitive burden required for a human to grasp the explanation [92].

XAI methods can be broadly categorized based on several dimensions:

- Scope (Local vs. Global):

o Local explanations aim to explain a single prediction of a model. Prominent examples include LIME (Local Interpretable Model-agnostic Explanations) [111], which approximates the black-box model locally with an interpretable model, and SHAP (SHapley Additive exPlanations) [96], which attributes the contribution of each feature to a prediction based on game theory.

o Global explanations seek to understand the overall behavior of the model. Techniques like Partial Dependence Plots (PDP) [11, 104] visualize the marginal effect of one or two features on the predicted outcome, while Accumulated Local Effects (ALE) plots [49, 108] offer a less biased alternative.

- Model-Specificity (Model-agnostic vs. Model-specific):

o Model-agnostic methods (e.g., LIME, SHAP, PDP, ALE) can be applied to any black-box model [103, 94], offering flexibility.

o Model-specific methods are designed for particular model architectures, such as neural networks. These include saliency maps [4, 5, 117], which highlight input regions most relevant to a prediction, and their advancements like Grad-CAM [116] and Grad-CAM++ [28]. Layer-wise Relevance Propagation (LRP) [15, 114], is another powerful technique for pixel-wise explanations in deep neural networks [2, 67, 69, 115]. Other model-specific methods for deep networks include those for debugging internals [6] and understanding hidden layers [83, 87].

- Explanation Form (Feature Importance, Counterfactuals, Rules):

o Feature Importance: Methods that quantify the contribution of individual features to a prediction (e.g., SHAP values, feature attribution maps).

o Counterfactual Explanations: These answer "What if?" questions, explaining what minimum changes to the input would alter the model's prediction [24, 25, 35, 53, 54, 57, 70, 73, 74, 75, 76, 77, 78, 80, 95, 97, 102, 105, 110]. They are increasingly popular due to their intuitive nature and potential for guiding recourse [76, 77].

o Rule-based Explanations: Simplifying complex models into sets of understandable rules.

o Case-based Explanations: Providing examples from the training data that are similar to the query instance and their predictions [81].

2.2 Human Factors in Explainable AI Evaluation

While technical metrics for explainability exist [37, 63, 118], the ultimate goal of XAI is to serve human understanding. Therefore, human factors and user studies are critical for evaluating XAI effectiveness [5, 7, 19, 63, 81, 99]. Research in this area reveals several key insights:

- Varying User Needs and Stakeholders: Different users (e.g., end-users, developers, domain experts, regulators) have distinct information needs and cognitive capacities [23, 45, 46, 48, 62, 82, 86, 100]. For instance, medical professionals may require different explanations than AI engineers [9, 10, 45, 64].

- Trust and Comprehension: Explanations are intended to build trust [19] and facilitate comprehension [59]. However, studies show an "illusion of explanatory depth," where users think they understand an explanation more than they actually do [29]. Trust can also be influenced by factors beyond just correctness, such as perceived expertise [19].

- Task-Dependency: The type of explanation needed often depends on the task at hand. Debugging a malware classification model [56] requires different insights than making a clinical decision [10, 45].

- Usability and Interface Design: The presentation of explanations (visual, textual, interactive) significantly impacts their utility [84, 99]. Designing stakeholder-tailored XAI interfaces is a growing area [82].

- Robustness of Explanations: The reliability and robustness of XAI methods themselves are under scrutiny

[8, 83, 114]. Adversarial examples can even "fool" saliency maps [83] or partial dependence plots [18]. Debugging techniques for XAI methods are also emerging [6].

In summary, while the technical toolkit for XAI is expanding, the field increasingly recognizes that the "explainable" part of XAI must be truly human-centered, acknowledging that there is no universal best explanation [48, 100]. This recognition forms the bedrock for our proposed mapping function.

3. Proposed Mapping Function for User-Centric XAI

The diverse landscape of XAI methods and the demonstrated heterogeneity of user needs necessitate a more adaptive and personalized approach to explanation generation. We propose a conceptual framework centered on a mapping function (M) that dynamically determines the optimal explanation strategy based on context.

3.1 Problem Statement and Rationale

Current XAI systems often provide a static explanation, or a limited set of pre-defined explanation types, regardless of who is receiving the explanation or why they need it. This "one-size-fits-all" approach leads to several inefficiencies:

- Cognitive Overload: Providing overly complex or irrelevant explanations to non-expert users can hinder comprehension and trust [29].
- Insufficient Detail: Expert users or developers might require more granular details or specific types of explanations (e.g., for debugging [56]) that simple methods do not provide.
- Mismatch with Task: An explanation suitable for auditing a model might be ineffective for helping a user decide on a loan application [98].
- Reduced Utility: If explanations are not tailored, their perceived value and actual impact on decision-making, learning, or trust can be significantly diminished [7, 23].

The rationale for a mapping function is to bridge this gap by intelligently aligning the explanation strategy with the specific requirements of the situation, thus maximizing the utility and impact of XAI.

3.2 Conceptual Framework: The Mapping Function (M)

The proposed mapping function, $ \mathcal{M} $, takes a set of input parameters and outputs a tailored explanation strategy. Conceptually, it can be defined as:

M(U,C,A,MAI)→EType,EFormat,EDetail,EGranularity

Where:

- $ U $: User Profile
- $ C $: Contextual Factors
- $ A $: AI Output Characteristics
- $ M_{AI} $: AI Model Characteristics
- $ E_{Type} $: Optimal Explanation Type
- $ E_{Format} $: Optimal Explanation Format
- $ E_{Detail} $: Optimal Level of Detail
- $ E_{Granularity} $: Optimal Granularity (Local/Global)

3.2.1 Inputs to the Mapping Function

1. User Profile (U): This encompasses characteristics of the individual requesting the explanation.

o Domain Expertise: Whether the user is an expert (e.g., clinician [45], data scientist), a novice, or a general end-user [19]. This influences the technicality and complexity of the explanation [45, 64].

o Role/Stakeholder Type: Developer, auditor, end-user, regulator, legal counsel [45, 62, 82, 86, 100]. Each role has distinct informational needs and purposes for explanations [23].

o Cognitive Style/Learning Preference: Some users prefer visual explanations (e.g., saliency maps for image data [5, 14]), while others prefer textual rules or counterfactuals [25].

o Prior Knowledge: The user's existing understanding of the AI system or the domain.

2. Contextual Factors (C): These define the specific situation in which the explanation is requested.

o Task Type: What the user intends to do with the explanation (e.g., debugging [1, 6, 56, 70], decision-making [7, 10, 19], auditing, learning, building trust [19], identifying bias [20, 32]).

o Decision Criticality/Stakes: Whether the AI decision is high-stakes (e.g., medical diagnosis [10, 45]) or low-stakes. High-stakes decisions often require more robust, verifiable, and perhaps simpler, counterfactual explanations.

o Time Constraints: Real-time operational decisions might require quick, concise explanations, while post-hoc analysis might allow for more in-depth explanations [65].

o Interaction History: Previous explanations provided to the user and their feedback [84].

3. AI Output Characteristics (A): Information directly related to the AI model's prediction for the specific instance.

o Prediction/Decision: The actual output of the AI model (e.g., classification label, regression value).

o Confidence Score: The model's confidence in its

prediction. Low confidence might trigger a more detailed explanation or a warning.

o Error Type (if applicable): If the AI made a mistake, the explanation might focus on identifying the cause of that error [1].

4. AI Model Characteristics (MAI): Properties of the AI model itself.

o Model Type: Neural network (e.g., CNN [69], LSTM [27], graph neural network [89]), tree ensemble [95], linear model [16], etc. This impacts which XAI methods are technically feasible and most effective.

o Model Complexity: Simple models might be intrinsically interpretable, while complex "black-box" models require post-hoc XAI.

o Data Type: Explanations for image data might favor visual methods like saliency maps [117], while tabular data might lend itself to feature importance or counterfactuals [35, 110].

o Transparency/Intrinsic Interpretability: Some models are designed to be inherently interpretable [113], while others require external XAI methods.

3.2.2 Mapping Logic (Conceptual)

The core of the mapping function would involve a set of rules, potentially learned, that connect the input parameters to the optimal explanation strategy. This logic could be implemented via:

• Rule-Based System: A set of IF-THEN rules (e.g., "IF User_Expertise is 'Novice' AND Task_Type is 'Decision-Making' THEN E_Type is 'Counterfactual' AND E_Format is 'Textual'").

• Machine Learning Model: A meta-learner that learns user preferences and explanation effectiveness based on past interactions and feedback. This could involve reinforcement learning where the system is rewarded for providing explanations that lead to better user outcomes (e.g., improved decision accuracy, increased trust).

• Knowledge Graph Reasoning: Representing user profiles, contexts, and XAI methods as nodes and relationships in a knowledge graph [10, 19, 62], allowing for complex inference to select explanations.

3.2.3 Outputs of the Mapping Function

The mapping function's output guides the XAI system in generating the most suitable explanation:

• Explanation Type (EType): Which specific XAI method to use (e.g., LIME, SHAP, counterfactual, PDP, rule extraction [40], causality-based [21, 60, 68, 78, 93, 97]).

• Explanation Format (EFormat): How the explanation should be presented (e.g., visual saliency map [5, 14], textual natural language explanation [100], interactive dashboard [84], graph-based [89], code-based for debugging [56]).

• Level of Detail (EDetail): The amount of information to provide (e.g., high-level summary for a manager vs. intricate details for a researcher).

• Granularity (EGranularity): Whether a local explanation for a specific instance or a global overview of the model's behavior is required.

3.3 Benefits of the Proposed Framework

Implementing a preference-aware mapping function offers several significant benefits:

• Increased User Satisfaction: By providing explanations that directly meet user needs, the system can enhance satisfaction and usability.

• Improved Decision-Making: Tailored explanations can lead to better human understanding and, consequently, more informed and effective decisions [7, 10].

• Enhanced Trust and Acceptance: When users receive explanations that resonate with their cognitive models and informational requirements, their trust in and acceptance of AI systems are likely to increase [19].

• Faster Debugging and Auditing: Developers and auditors can quickly obtain the specific insights they need to identify and rectify model errors or biases [1, 6, 20, 32, 56, 61, 70].

• Greater Compliance: Regulators and legal teams can obtain explanations in a format and level of detail suitable for auditing and ensuring adherence to regulations like the GDPR's "right to explanation" [52, 72].

4. Challenges and Future Directions

While the proposed mapping function framework offers a promising direction for user-centric XAI, its realization entails several significant challenges that define crucial areas for future research.

4.1 Quantifying and Eliciting User Preferences

One of the foremost challenges is to systematically quantify and elicit user preferences for explanations. User preferences are often implicit, dynamic, and context-dependent, making them difficult to capture.

• User Study Methodologies: Developing robust and scalable user study methodologies to collect ground truth data on preferred explanation types and formats across diverse user groups and tasks [5, 23, 81, 99]. This includes innovative experimental designs to avoid the "illusion of explanatory depth" [29].

• Implicit Feedback Mechanisms: Designing systems that can infer user preferences from implicit feedback (e.g., interaction patterns, time spent on explanations, subsequent actions) rather than relying solely on explicit

ratings.

- Personalized Preference Models: Building computational models that can learn and predict individual or group-level preferences for explanations, akin to recommender systems [Yildiz et al., 2023, 23; Zarindast & Wood, 2021, 24].

4.2 Robust Evaluation Metrics for XAI

Beyond qualitative user studies, developing quantitative and objective metrics for evaluating the effectiveness of user-tailored explanations is crucial [37, 63, 118]. Traditional metrics like fidelity or stability [8] may not fully capture human-centered aspects.

- Task-Specific Performance: Measuring whether the tailored explanations actually improve human performance on downstream tasks (e.g., faster debugging, more accurate decisions, better learning outcomes) [7].
- Cognitive Load: Quantifying the cognitive effort required for users to understand different types of explanations.
- Trust Calibration: Developing metrics to assess whether explanations correctly calibrate user trust, avoiding both over-trust and under-trust [19].
- Objective Metrics of Explainability: Continued research into metrics like "degree of explainability" [118] that can objectively assess the inherent quality of an explanation.

4.3 Dynamic Adaptation and Learning

For the mapping function to be truly effective, it must be capable of dynamic adaptation and continuous learning.

- Real-time Adaptation: Developing mechanisms for real-time adjustments to explanation strategies based on immediate user context and feedback. This might involve techniques from adaptive systems [Dewan et al., 2023, 6].
- Longitudinal Learning: Enabling the mapping function to learn and refine its strategies over extended periods of interaction with users, adapting to evolving preferences or changes in domain expertise.
- Transfer Learning: Exploring whether learned user preferences for explanations in one domain can be transferred or adapted to another, reducing cold-start problems for new applications.

4.4 Causal Explanations and Counterfactuals

There is a growing consensus that humans prefer causal explanations and counterfactuals because they align with how humans reason about the world [24, 25, 100].

- Generating Causal Explanations: Research is needed to develop more robust and scalable methods for extracting and presenting causal relationships from complex AI models [21, 26, 60, 68, 78, 93, 97]. This includes addressing challenges with imperfect causal knowledge [78].
- Effective Counterfactuals: Improving the generation of diverse [105], actionable, and plausible [80, 97] counterfactual explanations for various data types [35, 57, 110]. This also includes ensuring the efficiency and scalability of generating such explanations [102, 73, 74].

4.5 Addressing Bias and Fairness

XAI is a crucial tool for identifying and mitigating algorithmic bias [20, 32, 61].

- Explaining Bias: Developing XAI methods specifically designed to highlight discriminatory decision-making factors or reveal unintended biases in models.
- Fairness by Explicability: Exploring how explainability itself can contribute to fairness [61]. Debugging model mistakes related to bias [1].

4.6 Interdisciplinary Research

The success of user-centric XAI heavily relies on interdisciplinary collaboration.

- Cognitive Science and HCI: Deeper integration of insights from cognitive psychology and human-computer interaction (HCI) research to understand human information processing, trust formation, and decision-making in the context of AI [9, 29, 62, 86, 99].
- Domain Expertise: Incorporating domain-specific knowledge and requirements directly into the XAI design process, as exemplified in healthcare [45, 64].
- Philosophy of Explanation: Drawing from philosophical theories of explanation to guide the design of truly meaningful AI explanations [100, 107, 109].

4.7 Ethical and Societal Implications

Finally, the deployment of powerful XAI systems also introduces new ethical considerations.

- Misleading Explanations: The risk that poorly designed or intentionally deceptive explanations could mislead users or auditors [29].
- Over-reliance: Users might over-rely on explanations, even if the underlying model is flawed.
- Accountability: Clarifying who is accountable when AI decisions, even with explanations, lead to negative outcomes.

5. CONCLUSION

The advent of powerful, yet opaque, Artificial Intelligence models has underscored the critical need for Explainable AI (XAI) to foster trust, enable debugging, and ensure accountability. However, the efficacy of XAI is fundamentally constrained by a failure to account for the diverse and context-dependent needs of its human users.

This article has presented a conceptual framework for a preference-aware mapping function that represents a significant step towards user-centric XAI. By intelligently tailoring the type, format, detail, and granularity of explanations based on user profiles, contextual factors, AI outputs, and model characteristics, this framework aims to optimize human comprehension, enhance decision-making, and build stronger trust in AI systems.

Our systematic review of the literature illuminated both the technical advancements in XAI methods and the growing recognition of the human element in explanation effectiveness. The proposed mapping function synthesizes these insights, advocating for a dynamic and adaptive approach that moves beyond generic explanations. While significant challenges remain, particularly in systematically quantifying user preferences, developing robust human-centered evaluation metrics, and ensuring real-time adaptation, the future of XAI lies in its ability to truly connect with and empower its human audience. By prioritizing interdisciplinary research and addressing the intricate interplay between AI capabilities and human cognitive needs, we can unlock the full potential of XAI, paving the way for more responsible, transparent, and impactful AI deployments across all sectors.

REFERENCES

1. Abid, A., & Zou, J. (2021). Meaningfully explaining a model's mistakes. arXiv preprint arXiv:2106.12723. Retrieved from https://arxiv.org/abs/2106.12723

2. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., & Lapuschkin, S. (2023). From attribution maps to human-understandable explanations through concept relevance propagation. Nature Machine Intelligence, 5(9), 1006–1019.

3. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.

4. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31, 9525–9536.

5. Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20) (pp. 275–285). ACM.

6. Alsallakh, B., Kokhlikyan, N., Miglani, V., Muttepawar, S., Wang, E., Zhang, S., Adkins, D., & Reblitz-Richardson, O. (2021). Debugging the internals of convolutional networks. In AXplainable AI Approaches for Debugging and Diagnosis. Retrieved from https://openreview.net/forum?id=0YRkrxe2blh

7. Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2020). Does explainable artificial intelligence improve human decision-making? arXiv preprint arXiv:2006.11194. Retrieved from https://arxiv.org/abs/2006.11194

8. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049. Retrieved from https://arxiv.org/abs/1806.08049

9. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I., & Precise4Q consortium. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310.

10. Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., et al. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health, 1(2), e0000016.

11. Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4), 1059–1086.

12. Arias-Duart, A., Parés, F., Garcia-Gasulla, D., & Gimenez-Abalos, V. (2022). Focus! Rating XAI methods and finding biases. In 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1–8). IEEE.

13. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina,1 D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

14. Ayhan, M. S., Kümmerle, L. B., Kühlewein, L., Inhoffen, W., Aliyeva, G., Ziemssen, F., & Berens, P. (2022). Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. Medical Image Analysis, 77, 102364.

15. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 10(7), e0130140.

16. Baier, A., Aspandi, D., & Staab, S. (2023). ReLiNet: Stable and explainable multistep prediction with recurrent linear parameter varying networks. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (pp. 3461–3469).

17. Bandi, A. (2019). Telecom Churn Prediction Dataset. Retrieved December 15, 2021, from https://www.kaggle.com/bandiatindra/telecom-churn-prediction/data

18. Baniecki, H., Kretowicz, W., & Biecek, P. (2022).

Fooling partial dependence via data poisoning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 121–136). Springer.

19. Bayer, S., Gimpel, H., & Markgraf, M. (2022). The role of domain expertise in trusting and following explainable AI decision support systems. Journal of Decision Systems, 32(1), 110–138.

20. Bertrand, A., Pearce, A., & Thain, N. (2022). Searching for Unintended Biases with Saliency. PAIR Explorables. Retrieved from https://pair.withgoogle.com/explorables/saliency/

21. Biswas, S., Corti, L., Buijsman, S., & Yang, J. (2022). CHIME: Causal human-in-the-loop model explanations. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (pp. 27–39).

22. Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 440–447).

23. Brennen, A. (2020). What do people really want when they say they want "Explainable AI?" We asked 60 stakeholders. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–7).

24. Buchsbaum, D., Bridgers, S., Weisberg, D. S., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1599), 2202–2212.

25. Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In International Joint Conference on Artificial Intelligence (IJCAI) (pp. 6276–6282).

26. Castro, A. C. (2022). Explainability and Causality in Machine Learning through Shapley values. Universidad de Sevilla. Retrieved from https://idus.us.es/items/27db97f5-1bca-404f-b060-187faf4e3ee7

27. Wang, C., Li, Y., Sun, X., Wu, Q., Wang, D., & Huang, Z. (2023). DeLELSTM: Decomposition-based linear explainable LSTM to capture instantaneous and long-term effects in time series. In 32th International Joint Conference on Artificial Intelligence (IJCAI 2023) (pp. 4299–4307).

28. Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 839–847). IEEE.

29. Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think I get your point, AI! The illusion of explanatory depth in explainable AI. In 26th International Conference on Intelligent User Interfaces (pp. 307–317).

30. Council of the European Union. (2014). Council Regulation (EU) No 269/2014. Retrieved June 17, 2024, from http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1416170084502&uri=CELEX:3214R0269