Perceptions of Explainable AI: how presentation is content

Perceptions of Explainable AI: How Presentation is Content

Fabio Michele Russo
IMT School for Advanced Studies Lucca
fabio.russo@imtlucca.it

1. Introduction

Software is made for people. Although there is software that interfaces only with other software, in the end, when programs automatically are adjusting parameters of the energy plant so that it does not overheat, they do that so the plant does not shut down completely, cutting off power to human activities. Many such examples can be found.

Ultimately, the purpose of any software is to make life better for humans [1], [2]. It is therefore around *software design* that I focus this contribution. Software design is an intentional process of creating computer applications to meet specific needs. At its core, design is about problem solving. It starts with a requirement - a need, a challenge, a goal - and it consists in the development of a solution that integrates functionality and usability [3], [4].

2. Background and related work

Explainability (*XAI*) is the capacity to extract from machine learning (*ML*) predictors the reasons for their predictions. It can mean improving the ML algorithms themselves or making *model-agnostic* algorithms of explainability that can *ex-post* extract explanations from the model and its prediction. Each explanation is run on a specific data point (*local explainers*) or on the overall model behavior (*global explainers*).

The field concerning presentations of XAI is fairly young, with studies on visual analytics presentations for explainable deep learning algorithms emerging only around 2015-2016 [§5.6, 5]. With such a new field, sufficient attention to how exactly presentation modalities shape understanding has been limited. My approach is to design novel presentations of XAI and evaluate them against existing state of the art presentations [6].

There is, however, an important corpus of literature on XAI itself, including work on popular algorithms like *SHAP* and recent studies questioning its efficacy [7], [8], [9].

3. Framework

The approach I propose is based on the research question of whether the presentation format of an AI explanation significantly affects how users perceive, understand and utilize the explanation itself. In pursuing answers to this question, my framework consists of the following steps:

1. Collect the state of the art in XAI presentation to end users.
2. Design and implement new XAI presentations with goals of usability, comparability, efficacy, realism and innovation.
3. Investigate users needs through pre-registered human studies. For each *interface* data will be collected on key features of knowledge acquired (*understanding*), relevance of said knowledge in improving the human's capability to solve problems (*actionability*) and variations in human *trust* towards the automated decision support system.
4. Analyze results on quantitative and qualitative metrics.

Key dimensions under consideration are: visual vs. textual explanations; traditional/static vs. contestable presentations; technical complexity and completeness vs. simplified analogies. All these dimensions will be evaluated relative to the user expertise levels and evaluation will be done employing XAI-specific metrics, psychological scales, statistical significance analyses and qualitative considerations. Finally, as a computer scientist myself, this work is being conducted inter-disciplinarily with psychologists and it is my desire to involve interested researchers from other disciplines.

4. Case study

In choosing the main case study I aim for one that is of interest to the general public and that requires no specialist knowledge. Therefore, I chose the domain of personal finance: a scenario where a person requests a small loan (50€-1000€) through a banking smartphone app. I investigate automated decisions of granting or not granting the loan, taken by an ML predictor and based on the applicant’s characteristics and financial history.

This approach allows for a broad participant pool and addresses a popular problem involving features clear to non-experts. Participants will be able to watch and interactively explore the behavior of the predictor through the XAI-powered interface, with the study gathering data on how understanding, actionability and trust change when users are presented with different forms and presentations of explanations: rule-based as graphical decision trees or textual descriptions; counterfactuals; feature importances presented as numbers, graphs or textual descriptions, all presented through different designs fostering distinct user experiences.

5. Implications and future work

This approach will show how different presentations of the same underlying explanation lead to different outcomes in human understanding and problem-solving capacity, which are the ultimate goals of XAI. I will also investigate trust, recognizing that it is not necessarily a positive quality for an ML predictor. In fact, an ML predictor that fosters misplaced trust is a source of concern.

There are ethical considerations regarding how presentations of the same information affect users. When left unchecked, presentation can create deep biases in the user, while shielding the practitioners from criticism because of the correct content that is present underneath the facade. When used by unprepared or malicious practitioners, presentation can be a dangerous potential source of manipulation.

This work has repercussions for regulators and legal scholars. Current regulations require "clear and meaningful" explanations of automatic decisions in high-risk cases [10]. As understanding of human perception of explainability grows, so can a corpus of best practices as well as our awareness of particular points of vulnerability of XAI systems when employed for critical human decisions.

Moreover, this work contributes to philosophical discussions: how we should balance the need to foster correct and relevant understanding of ML systems with concerns of scientific paternalism and how we, researchers in computer science, must join a conversation on this point, one that is conducted trans-disciplinarily. Finally, questions about the human concept of trust are key: is trust towards the computer system the same as trust towards its designers, or are we measuring two different things?

6. Conclusion

This abstract proposes a perspective and a research framework that will show how presentations of XAI fundamentally shape user understanding, actionability and trust. By developing new presentations of XAI and comparing them with existing ones, the work will provide empirical evidence on how different interfaces affect these dimensions. The findings hold implications across disciplinary borders spanning from research to industrial and regulatory fields. Ultimately, this work acknowledges the ethical dimension of presentation choices in XAI and aims to ensure that explainable systems truly serve their human users, aligning with the core principle that software is made for people.

7. Acknowledgements

I would like to thank my colleague Alice Andrea Chinaia for copy editing.

8. References

---

1. Friedman, B., & Kahn, P. H. (2003). "Human values, ethics, and design." The human-computer interaction handbook, 1177-1201.
2. Shneiderman, B. (2022). "Human-centered AI." Oxford University Press.
3. Nielsen, J. (1994). "Usability Engineering." Morgan Kaufmann.
4. Garrett, J.J. (2010). "The Elements of User Experience: User-Centered Design for the Web and Beyond." New Riders.
5. La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R. and Angelini, M. (2023), State of the Art of Visual Analytics for eXplainable Deep Learning. Computer Graphics Forum, 42: 319-355. https://doi.org/10.1111/cgf.14733
6. Liao, Q. V., & Varshney, K. R. (2023). "Human-centered explainable AI (XAI): From algorithms to user experiences." ACM Computing Surveys.
7. Lundberg, Scott M and Lee, Su-In, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, volume 30, 2017.

https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

8. Xuanxiang Huang, Joao Marques-Silva, On the failings of Shapley values for explainability, International Journal of Approximate Reasoning, Volume 171, 2024, 109112, ISSN 0888-613X, https://doi.org/10.1016/j.ijar.2023.109112. https://www.sciencedirect.com/science/article/pii/S0888613X23002438
9. Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). "Problems with Shapley-value-based explanations as feature importance measures." Proceedings of the 37th International Conference on Machine Learning, PMLR 119:5668-5679.
10. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) http://data.europa.eu/eli/reg/2024/1689/oj