Perceptions of Explainable AI: how presentation is content

Fabio Michele Russo[a,*]

[a]IMT School for Advanced Studies Lucca
ORCID (Fabio Michele Russo): https://orcid.org/0009-0003-4138-8822

Abstract. This extended abstract presents the author's current research in Explainable Artificial Intelligence (*XAI*). I begin by offering a perspective on research in this field and a summary of my contributions to date. Next, I present the questions being investigated in this work, the adopted research framework, some of this work's implications across disciplinary borders and directions for remaining work. This extended abstract draws from work which I also described in an extended abstract presented at *BRIO - Bias, Risk and Opacity in AI*, 1 July 2025, Milan, Italy.

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 [2], [12]. It is therefore around software design for Explainable AI that we 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 [10], [3].

2 Short background and related works

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 [7]. With such a new field, sufficient attention to how exactly presentation modalities shape understanding has been limited. The approach presented in this work consists in designing novel presentations of XAI and evaluating them against existing state of the art presentations [8].

There is of course an important corpus of literature on XAI itself, including work on popular algorithms like *SHAP* [9] and recent studies questioning its efficacy [5], [6].

3 Published contributions

In pursuit of a XAI algorithm that could enable the design of a real-time explanation system for image data in real-world use cases, in Russo et al., *Explainable AI in Time-Sensitive Scenarios: Prefetched Offline Explanation Model* [11] we have developed an algorithm (POEM) for XAI over image data, building on an existing algorithm (ABELE [4]) and improving by a factor of 8 to 10 its execution time. Moreover, POEM features completely new explanation generation algorithms - specifically for exemplars and counterexemplars. These explanations are more diverse and more plausible: these properties are observable in the design of the generation algorithms and have been verified experimentally by us. Since diversity and plausibility are sometimes antagonistic in explanation generation, improving on both we believe is a particularly good result.

4 Current framework

The approach we 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, the 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 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. Understanding, actionability and trust are *dimensions of interest*.
4. Analyze results on quantitative and qualitative metrics.

Key variables under consideration are: visual vs. textual explanations; traditional/static vs. contestable presentations; technical complexity and completeness vs. simplified analogies. All these will be evaluated relative to the user's expertise levels and by employing XAI-specific metrics, psychological scales, statistical significance

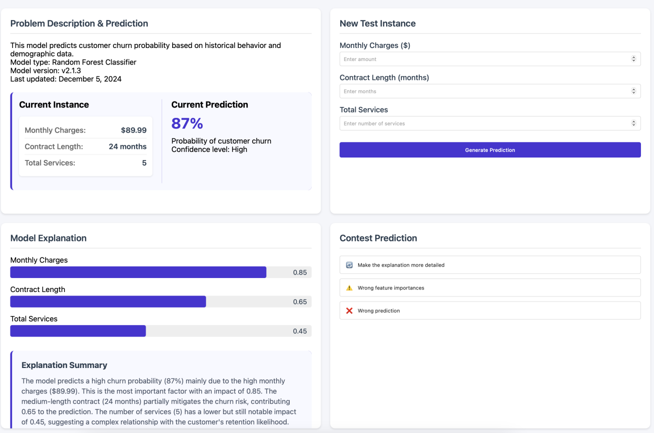
[*] Email: fabio.russo@imtlucca.it.

analyses and qualitative considerations. Finally, as a computer scientist myself, this work is being conducted inter-disciplinarily with psychologists and it is our desire to involve interested researchers from other disciplines.

5 Case study

In choosing the main case study we aim for one that is of interest to the general public and that requires no specialist knowledge. Therefore, we chose the domain of personal finance: a scenario where a person requests a small loan (50€-1000€) through a banking smartphone app. We 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 different XAI-powered interfaces, with the study gathering data on how understanding, actionability and trust change when users are presented with different interfaces, that include different forms and presentations of explanations as well as different avenues of interaction. Such forms of explanations are: 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. The first novel presentation of XAI that will be part of the experiment is introduced in Figure 1.



Problem Description & Prediction

This model predicts customer churn probability based on historical behavior and demographic data.
Model type: Random Forest Classifier
Model version: v2.1.3
Last updated: December 5, 2024

Current Instance

Monthly Charges:	$89.99
Contract Length:	24 months
Total Services:	5

Current Prediction

87%

Probability of customer churn
Confidence level: High

Model Explanation

Monthly Charges	0.85
Contract Length	0.65
Total Services	0.45

Explanation Summary

The model predicts a high churn probability (87%) mainly due to the high monthly charges ($89.99). This is the most important factor with an impact of 0.85. The medium-length contract (24 months) partially mitigates the churn risk, contributing 0.65 to the prediction. The number of services (5) has a lower but still notable impact of 0.45, suggesting a complex relationship with the customer's retention likelihood.

Figure 1. The first of the developed new presentations of XAI, codenamed *Anemone*. This presentation focuses on user exploration of the model's behavior: first, the user obtains a prediction with explanation on the data point of interest (in the left two quadrants); second, in the top right quadrant, the user can modify some of the input features - marked as "explorable" in the model's configuration - and get a new prediction with explanation, which is shown compared to the original one. This way, with an experiment involving A/B style studies, we want to understand the impact of this kind of exploration on the dimensions of interest. This approach also recognizes the fact that, from a security or ethical standpoint, some of the features in the problem space should not be freely explored, and allows them to be marked as "non explorable".

6 Implications

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. We 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 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 [1]. 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?

7 Future work

This experiment will lead to considerations on human perceptions of XAI that will be crucial in understanding human needs and requirements of the same. By letting user requirements lead the work, the next piece of research will consist in adapting the novel XAI algorithm that we already developed and ensuring that it can fit as well as possible with human requirements. Points of improvement to be considered are: speed and applicability to very high-dimensional datasets. However, more improvements may need to be made, improvements that will become clear once the current work reaches its conclusion with the identification of critical user requirements fostering understanding, actionability and affecting trust.

8 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.

References

[1] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence*. 2024.

[2] B. Friedman and P. H. Kahn. Human values, ethics, and design. In *The human-computer interaction handbook*, pages 1177–1201. 2003.

[3] J. Garrett. *The Elements of User Experience: User-Centered Design for the Web and Beyond*. New Riders, 2010.

[4] R. Guidotti, A. Monreale, S. Matwin, and D. Pedreschi. Black box explanation by learning image exemplars in the latent feature space. In *ECML/PKDD*, volume 11906, pages 189–205, 2019.

[5] X. Huang and J. Marques-Silva. On the failings of shapley values for explainability. *International Journal of Approximate Reasoning*, 171: 109112, 2024.

[6] I. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5668–5679, 2020.

[7] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42: 319–355, 2023.

[8] Q. V. Liao and K. R. Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *ACM Computing Surveys*, 2023.

[9] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[10] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, 1994.

[11] F. M. Russo, C. Metta, A. Monreale, S. Rinzivillo, and F. Pinelli. Explainable ai in time-sensitive scenarios: Prefetched offline explanation model. In D. Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, and F. Naretto, editors, *Discovery Science*, pages 167–182, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-78980-9.

[12] B. Shneiderman. *Human-centered AI*. Oxford University Press, 2022.