User Study on the Effects Explainable AI Visualizations on Non-experts

Sophia Schulze-Weddige(✉) and Thorsten Zylowski

Future Labs, CAS Software AG, CAS-Weg 1-5, 76131 Karlsruhe, Germany
sophia.schulze-weddige@gmx.de, thorsten.zylowski@cas.de

Abstract. Artificial intelligence is drastically changing the process of creating art. However, in art, as in many other domains, algorithms and models are not immune from generating discriminatory and unfair artifacts or decisions. Explainable Artificial Intelligence (XAI) makes it possible to look into the "black box" and to identify biases and discriminatory behaviour. One of the main problems of XAI is that state-of-the-art explanation tools are usually tailored to AI experts. This paper evaluates how intuitively understandable the same tools are to laypeople. By using the prototypical use case of predictive sales, and testing the results with users, the abstract ideas of XAI are transferred to a real-world setting to study its understandability.

Based on our analysis, it can be concluded that explanations are easier to understand if they are presented in a way that is familiar to the users. A presentation in natural language is favorable because it presents facts unambiguously. All relevant information should be accessible in an intuitive manner that avoids sources of misinterpretations. It is desirable to design the system in an interactive way that allows the user to request further details on demand. This makes the system more flexible and adjustable to the use case. The results presented in this paper can guide the development of explainability tools that are adapted to a non-expert audience.

Keywords: Explainable AI · Human-centric AI · User study

1 Introduction

Many facets of art can be created by artificial intelligence, including paintings and literary works, as well as audio and video art. However, these systems can contain biases and show discriminatory behavior. For example, biases were found in AI-based generated art [17]. In addition, there is a large body of work dealing with the classification of art, especially paintings [2,19]. For training such classifiers, datasets are collected that may be biased (e.g. eurocentristic bias, gender bias, etc.). This would result in classifications favoring certain regions or groups. One can easily imagine a classifier that systematically rates European paintings higher (in price, in quality) than paintings from regions that are less strongly

M. Wölfel et al. (Eds.): ArtsIT 2021, LNICST 422, pp. 457–467, 2022.
https://doi.org/10.1007/978-3-030-95531-1_31

represented internationally, simply because the representation of the different regions is unequally distributed in the data. In cases like these, where the decisions affect important aspects of our lives, it is indispensable to understand the underlying decision process to control for fairness and safety. Explainable Artificial Intelligence (XAI) is one way to make the decisions of automated systems transparent. It has been used to uncover cases in which algorithms had been unfair, for example towards protected minorities or women [3]. Detecting these biases is a first and necessary step to eliminate them.

XAI has become very popular in recent years. Many large software companies such as Microsoft[1] and IBM[2] develop tools to access the decision process of AI systems. Novel papers about the topic are being published regularly. But many explanation methods currently available are tailored for people with prior knowledge in machine learning. This makes them difficult to understand for laypeople. The need for human-centric explanations has been reported frequently and attempts have been made to provide them [16].

This paper aims to evaluate out-of-the-box explainability methods in the context of non-expert users. That is, evaluate whether the tools available are suited to explain automated decision processes to people that have no proficiency in AI. A user study is conducted based on a real-world example from the sales industry which leads to precise recommendations for the development of human-centered XAI. There are excellent summaries about XAI [15] and human-centered AI [5], but they fail to provide detailed suggestions because many decisions depend heavily on the use case. By generalizing the insights from the results of this specific use case, this paper aims to provide valuable information which is intended to help other practitioners to make informed choices for their explanation systems.

2 Related Work

XAI aims to clarify how an automated decision is generated. Hereby, many XAI approaches focus on the systems that generate the decision. While it is imperative to accurately portray the decision process, correctness is not sufficient to make explanations understandable to humans [11]. The notion of human-centered XAI puts the human back in the focus of attention. The goal is to provide explanations that appeal to the person using the system. This means the explanations are easy to understand and not misleading. As the users cannot be expected to have prior knowledge about AI, the explanation should be adjusted to the target audience [13].

Explainability includes the ability of humans to understand the explanation. When designing a human-centered system, the first step is to define precisely whom the explanation is aimed at. Then, the goal of the explanation needs to be determined. Some guidelines help in designing such a system [9]. But they state

[1] See Microsoft's toolkit at https://github.com/interpretml/interpret.
[2] See IBM's toolkit at https://github.com/Trusted-AI/AIX360.

that many use case-specific decisions need to be made. As the use cases can be distinct, it is difficult to give general best practices.

One way to design an ideal system for a specific use case follows the sociotechnological approach described by Ehsan and Riedl [5]. They state that the social and the technical part of a human-centered system co-evolve in an iterative process. They describe a cycle of altering the system to the needs of the user and evaluating the effect. One example for a use case-specific implementation is called "Glass Box" [16]. They developed a chat- or voice-based interactive dialogue for the loan application data set [4]. In their study, participants are presented with counterfactual explanations for why their loan application was rejected. If they are not contemptuous with the answer they can ask follow-up and what-if questions. More research on use case-specific implementations for human-centered XAI is needed to develop a deeper understanding of how to make explanations understandable for humans.

3 The User Study

A user study is conducted to evaluate different explainability tools based on a real-world example in predictive sales which aims to learn from past sales outcomes to predict customer behavior in the future. The explanation methods evaluated in this study are based on a classifier that predicts the status of a lead. A lead is the collection of data about an individual customer. The data set contains 440 variables and comprises roughly 10000 leads. A random forest classifier was used for the classification task. An averaged accuracy of 0.9 was yielded and a precision, recall, and F1-score of 0.91.

An online application is developed in which the participants can choose between three different example leads. One of the example leads is lost, the second is won and the third is very close to the decision boundary. The examples are selected randomly in the interval of their prediction, namely [0, 0.5), [0.49, 0.5), [0.5, 1]. For each example, the participant can further choose between five visualizations which are generated with the SHAP [10], LIME [12] and alibi explain [18] python packages. They are chosen based on popularity and type of explanation. Only instance-based explanations are included, as users are usually more interested in understanding information regarding themselves, rather than aim to comprehend the underlying model behavior in detail [1].

3.1 The Explanation Tools

LIME provides local explanations for individual predictions by fitting interpretable models to input-output pairs. These pairs are generated by randomly perturbing the instance at question to depict its local surroundings. The perturbed instances are then used as an input to the model to calculate their outputs. The generated samples are weighted by their distance to the instance in question. LIMEs visualization is divided into three parts. On the left, the prediction probability for the different classes is depicted. In the center, the feature

importance of the ten most important features is shown. On the right, the corresponding values of these features can be found. Colors indicate whether the influence of that feature is positive (orange) or negative (blue).

Fig. 1. Sample explanation from the custom application with LIME. (Color figure online)

Lundberg and Lee developed an explanation method that is based on Shapley values from cooperative game theory [10]. Shapley values distribute the surplus in a cooperative game between the contributors depending on their influence on the outcome [8]. In the context of XAI, the features can be interpreted as the players and the prediction as the outcome. Hence, Shapley values are a measure for the contribution of the individual features to the overall prediction.

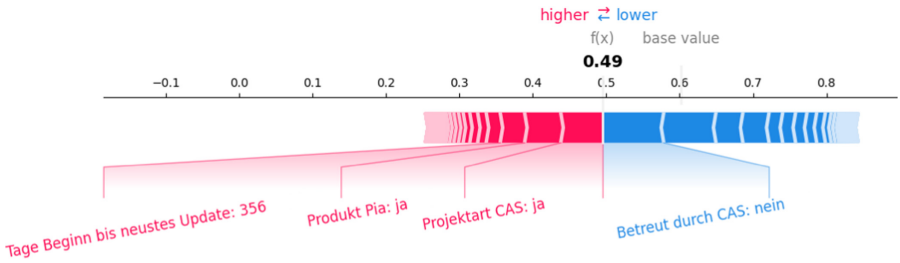

Fig. 2. Sample force plot with SHAP.

Three visualizations from the SHAP package are used in this study. All of them show the features with the highest Shapley values. The first one uses the analogy to a force that is pushing the prediction to its final value to visualize the effect of individual variables (Fig. 2). Variables with a positive effect push the prediction higher up the number strip, which is indicated by arrows to the right. Variables with a negative effect push the prediction to the left, which means lowering the prediction value. The forces are at an equilibrium in the final prediction value. The width of the arrows indicates the strength of the

effect. Variables with an effect of at least 5% are written out together with the corresponding value.

In the second visualization, Shapley values are expressed in a bar plot (Fig. 3). The bar plot is bi-directional, which means the bars start at 0 in the center and point either to the left for negative values or to the right for positive ones. Additionally, the bars are color-coded with blue for negative and red for positive effects. The length of the bar indicates the magnitude of the Shapley value. The y-axis shows the names of the variables and the corresponding values of the instance.

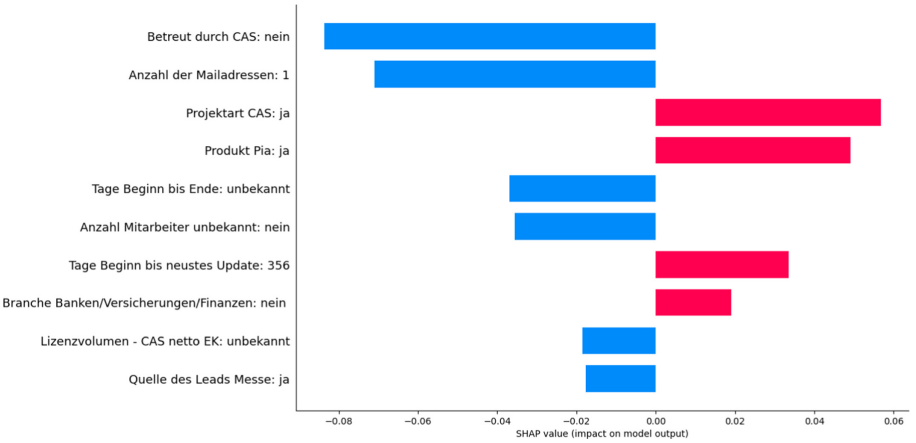

Fig. 3. Sample bar plot with SHAP.

Thirdly, the same information can be depicted in a decision plot (Fig. 4). This type of plot shows a decision path that can be followed from bottom to top, where it ends at the final prediction. Which variable is being considered can be seen on the y-axis. The x-axis shows the current prediction value. When the line moves to the left, it denotes a negative effect in the corresponding variable. When it moves to the right, the effect is positive.

Lastly, explanations are presented as counterfactual examples. By making explicit what would have to change in the input in order to yield a different output, they not only provide information on the reasons behind a decision but also on how to alter it in the future. This information is highly valuable for humans who usually ask for why something happened *rather* than something else [11]. Moreover, counterfactual explanations are easy to understand because they are presented in natural language which is "the most accessible modality of explanation" [5]. The changes are presented in a bullet point list stating what could be done to improve the prediction of the instance. Further, the magnitude of the change is shown in brackets. A bullet point list could look like this:

- Source of the lead is **not** trade fair (0.1)
- Project type **is** partner (0.07)

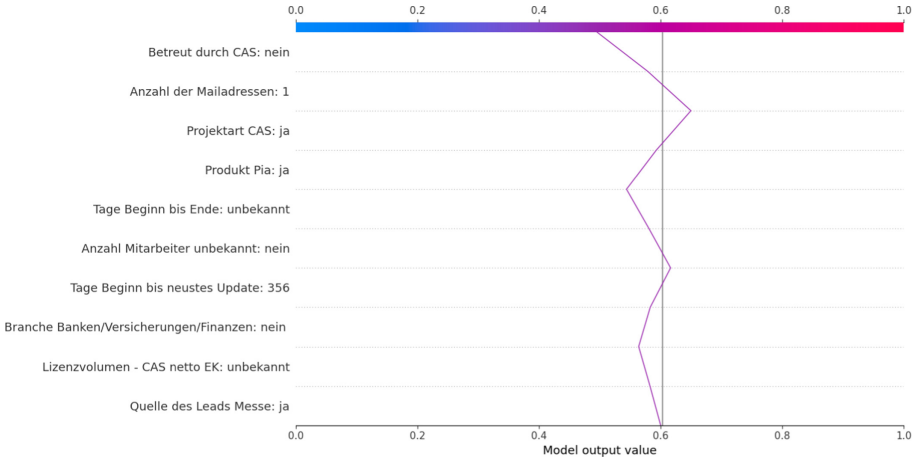

Fig. 4. Sample decision plot with SHAP.

3.2 Study Design

In the user study, participants first make themselves comfortable with the use of the online application by exploring it freely. Then they answer six questions. The necessary information to answer them can be found in the application. The participants are not instructed to search for the answers in a specific way but are free to use the application to their liking. This includes choosing which and how many methods to consult before answering each question. On the one hand, the questions aim to evaluate whether the participants can extract the relevant information from the application. On the other hand, the questions help to see which explainability methods are preferably used to find the information. After each question, participants indicate which methods they used for their answer. It is possible to select multiple methods if more than one were considered. During the whole study, participants are asked to describe their train of thought and opinion about the methods. At the end, five statements from the system causability scale (SCS) [7] are used to inquire the opinion about the different explainability methods. The agreement to those statements is measured on a five-item Likert scale [14].

1. I understood the explanations within the context of my work.
2. I did not need support to understand the explanations.
3. I was able to use the explanations with my knowledge base.
4. I think that most people would learn to understand the explanations very quickly.
5. I did not need more references in the explanations: e.g., medical guidelines, regulations.

4 Results

Fifteen employees from CAS Software AG participated in the user study, five of which work in the sales department, five in the research department and the remaining five in consulting, development, or product management. Neither of the participants is an expert in AI or has seen the explainability tools before, except for one who briefly encountered SHAP. The interviews lasted between 20 and 50 min with an average of 30 min per participant. The results of the Shapiro Wilk normality test show that the data from the SCS is not normally distributed for the bar plot ($p = 0.0003$) and the counterfactual explanation ($p = 0.01$). Thus, a non-parametric method for the evaluation of the ratings is used. The Wilcoxon signed rank test for dependent samples is conducted pair-wise for all explainability methods.

Table 1. Table of summary statistics. This table shows the means, standard deviations and the SCS scores for the five explainability methods.

	Bar plot	Decision plot	Force plot	LIME	Counterfactuals
Mean	4.41	3.49	3.29	3.07	4.08
Std	1.05	1.28	1.47	1.29	1.21
SCS score	88	70	66	61	82

The mean values and standard deviations can be found in Table 1. As the mean rankings for the bar plot ($\mu = 4.41$) and the counterfactual explanations ($\mu = 4.08$) are unequivocally higher than for the decision plot ($\mu = 3.49$), the force plot ($\mu = 3.29$) and the LIME plot ($\mu = 3.07$), a one-sided hypothesis test is conducted. For the decision, force, and LIME plot the effect is not as clear. Therefore, a two-sided test was conducted to compare these three among each other.

Nine out of the fifteen participants gave the bar plot the highest rating and another four participants gave it the second-highest rating. From the remaining six participants, four gave the highest score to the counterfactual explanations and two to the decision plot. All participants except for two have a mean rating higher than 4.0 for the bar plot. For the counterfactual explanations, it is all but four.

The results show that the bar plot has a significantly higher rating than the decision plot ($p = 0.0062$), the force plot ($p = 0.0011$), and the LIME plot ($p = 0.0003$). The counterfactual explanations have a significantly higher rating than the LIME plot ($p = 0.002$). No significant differences can be found between the ratings of the decision, force, and lime plot. For an overview of all the results see Table 2.

After each question, the participants indicated which methods they used to find the relevant information. The bar plot was used most frequently. It was

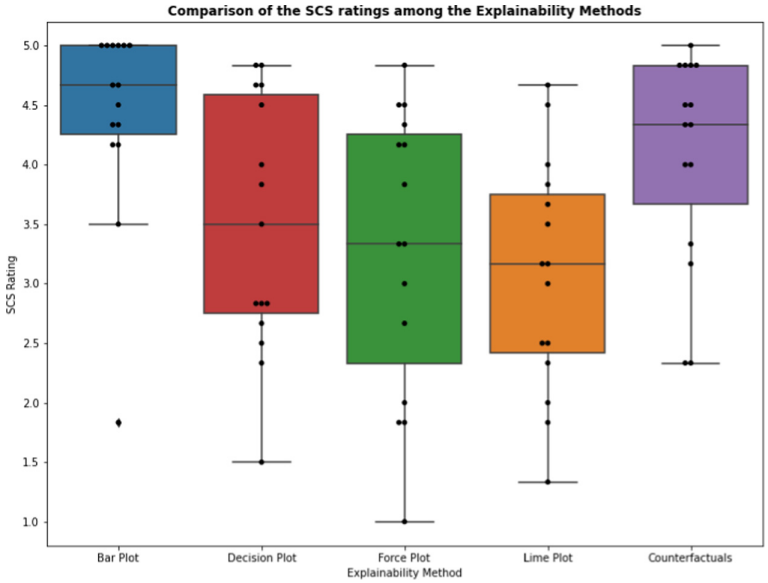

Fig. 5. Box plots for the SCS ratings. The box plots show the median and the first and third quartile, as well as the minimum and maximum values. The ratings of each participant for each plot are displayed as a black dot.

Table 2. Results of the Wilcoxon signed-rank test. This table shows the p-values calculated by the Wilcoxon signed-rank test. A p-value lower than 0.01 is considered significant and shown in bold.

	Counterfactuals	Decision plot	Force plot	LIME
Bar plot	0.103	**0.006**	**0.001**	**0.0004**
Counterfactuals		0.069	0.028	**0.002**
Decision plot			0.599	0.256
Force plot				0.495

involved in answering one of the questions in 44 cases. The second most frequently used method was the counterfactual explanations with 34 cases, followed by the decision plot with 27 cases, the LIME plot with 24 cases, and lastly the force plot with 16 cases.

The results for the SCS ratings match with the statement of the participant during the user study. All of the participants talked positively about the bar plot and five explicitly stated it as their favorite method. Ten participants noted that they had trouble understanding the decision plot and six said that the force plot has too little information as it only shows a small number of variables. Moreover, nine participants said that they had trouble understanding the variables.

5 Discussion and Conclusion

The analysis of the SCS ratings clearly shows that the bar plot is the favorite explainability method in this study. It has a significantly higher rating than the decision, force, and LIME plot and it received the highest overall ranking for almost two-thirds of the participants. Further, it has been used most frequently during the study. It appears that many users are familiar with bar plots, which shows that most people prefer visualizations that they already know. This can be confirmed by the second highest ranking method, the counterfactual examples. These are easy to understand as they are presented in text form. The familiarity helps to focus on the relevant aspects. It makes sense to pursue using simple, familiar plots if possible. In cases where a more complex display of information is needed, precise and easy clarifications should be provided. In many applications, it is not feasible to carry out a tutorial or lengthy introduction. Hence, the application must be self-sufficient in its usage.

Although simplicity is favorable to see all information at a glance, relevant information such as the direction of the plot (decision plot and force plot) and the values on the axes (bar plot) should be made clearly visible. Relevant clarifications about the interpretation of the plot should be integrated directly into the plot. Important aspects of the plot should be highlighted. This particularly holds for parts of the plots that might lead to misinterpretations. For example, the effect size of the bar plot might be overestimated because the bars are scaled by the width of the plot. This makes it difficult to grasp the real effect size from the plot. Directing the focus to the numeric value of the effect by increasing the font size of the x-axis or adding the values next to the plots, reduces the chances of misinterpretation. The counterfactual explanations could be enhanced by transforming the bullet points into more natural sentences and presenting precise actionable suggestions for improvement. In general, user studies are highly relevant to detect pitfalls like these and should be conducted to build an explanation system that matches the target group.

Moreover, the possibility for interaction between the user and the application is highly favorable. One way to allow for interaction is to provide details on demand. Following Grice's maxims, only information that is relevant to the situation and not known yet should be displayed [6]. This can vary between instances. Thus, it facilitates understandability if the user can ask for clarification if necessary but is not overwhelmed with details at the beginning. Further, integrating interactions to the application makes it more flexible and adjustable to the user's needs. That means the same system can be used by different people in a way that suits them. One way to anticipate interactions is in the form of interactive plots. For example, hovering over variable names to receive further elaborations or linking sources where additional help can be found.

In addition, participants asked for more details about the variables. Especially possible value instantiations are of interest. Nine participants explicitly stated that they had trouble understanding the variables. An explanation regarding the variable names and possible values could facilitate the understanding. The variables are the basis of the explanations. If they are not clear to the user

it makes the interpretation of the explanations difficult even if the effect was understood correctly.

All in all, in order to yield higher understandability the explanation methods should be adjusted to the use case and the target audience. But some general considerations can guide the design choices. Simple and familiar visualizations are preferable over complex and detailed ones. Possible sources of misinterpretation should be detected and the visualizations should aim to direct the focus to relevant information to avoid them. Moreover, allowing for interaction between the user and the system increases flexibility and enhances the user experience. Following these suggestions, explanation tools can be used to reveal the underlying decision process of an algorithm to non-experts in AI art, as well as various other domains.

References

1. Arya, V., et al.: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques (2019). arXiv preprint arXiv:1909.03012
2. Cetinic, E., Grgic, S.: Genre classification of paintings. In: 2016 International Symposium ELMAR, pp. 201–204 (2016). https://doi.org/10.1109/ELMAR.2016.7731786
3. Chiusi, F.: Report: automated society 2020. J. Chem. Inf. Model. **110**(9), 1689–1699 (2017)
4. Dua, D., Graff, C.: UCI machine learning repository (2017). https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
5. Ehsan, U., Riedl, M.O.: Human-centered explainable AI: towards a reflective sociotechnical approach. In: International Conference on Human-Computer Interaction, pp. 449–466 (2020). http://arxiv.org/abs/2002.01092
6. Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Speech Acts, Syntax and Semantics, vol. 3, pp. 41–58. Academic Press, New York (1975)
7. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the system causability scale (SCS): comparing human and machine explanations. KI - Kunstliche Intelligenz **34**(2), 193–198 (2020)
8. Kuhn, H.W., Tucker, A.W.: Contributions to the Theory of Games (AM-28), Vol. II. Annals of Mathematics Studies, Princeton University Press (2016). https://books.google.de/books?id=Pd3TCwAAQBAJ
9. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. Conf. Human Factors Comput. Syst. - Proc. (2020). https://doi.org/10.1145/3313831.3376590
10. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **2017**(Section 2), 4766–4775 (2017)
11. Miller, T.: Explanation in artificial intelligence insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). arXiv:1706.07269
12. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier (2016)
13. Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of user-centered explainable AI. In: CEUR Workshop Proceedings, Vol. 2327 (2019)
14. Robinson, J.: Likert Scale, pp. 3620–3621. Springer, Netherlands, Dordrecht (2014). https://doi.org/10.1007/978-94-007-0753-5

15. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: a review of methods and applications. Proc. IEEE **109**(3), 247–278 (2021). https://doi.org/10.1109/JPROC.2021.3060483
16. Sokol, K., Flach, P.A.: Glass-box: explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In: IJCAI, pp. 5868–5870 (2018)
17. Srinivasan, R., Uchino, K.: Biases in generative art - a causal look from the lens of art history (2021)
18. Van Looveren, A., Klaise, J.: Interpretable counterfactual explanations guided by prototypes (2019). arXiv preprint arXiv:1907.02584
19. Zujovic, J., Gandy, L., Friedman, S., Pardo, B., Pappas, T.N.: Classifying paintings by artistic genre: an analysis of features classifiers. In: 2009 IEEE International Workshop on Multimedia Signal Processing, pp. 1–5 (2009). https://doi.org/10.1109/MMSP.2009.5293271