Comparing User Perception of Explanations Developed with XAI Methods

Comparing User Satisfaction of Explanations Developed with XAI Methods

Jonathan Aechtner, Lena Cabrera, Dennis Katwal, Pierre Onghena, Diego Penroz Valenzuela and Anna Wilbik
Department of Data Science and Knowledge Engineering (DKE)
Maastricht University, Maastricht, The Netherlands
{j.aechtner,l.cabreraperez,d.katwal,p.onghena,d.penrozvalenzuela}@student.maastrichtuniversity.nl,
a.wilbik@maastrichtuniversity.nl

***Abstract*—Artificial Intelligence (AI) has gained notable momentum, culminating in the rise of intelligent machines that deliver unprecedented levels of performance in many application sectors across the field. In recent years, the sophistication of these systems has increased to an extent where almost no human intervention is required for their deployment. A crucial feature for the practical deployment of AI-powered systems in critical decision-making processes is the ability to understand how these systems derive their decisions. Accordingly, the AI community is confronted with the barrier of explaining the reasoning behind machine-made decisions. Paradigms underlying this problem fall within the field of eXplainable AI (XAI). Research in this field has introduced various methods to shed light into black box models such as deep neural networks. While local explanation methods explain the reasoning behind an output for a single decision, global explanations aim to describe the general behaviour of a model, i.e. for all decisions. This paper investigates users' perceptions of local and global explanations generated with popular XAI methods — LIME, SHAP, and PDP — by conducting a survey to find which of the explanations are preferred by different users. Meanwhile, two hypotheses are tested: first, explanations increase users' trust in a system, and second, AI novices prefer local over global explanations. The results show that explanations from PDP achieved the best user evaluation among the considered XAI methods.**

I. Introduction

In today's Information Age, information and data has become a commodity that is easily accessible and quickly and widely disseminated. As a response, many sectors in the industry and society have embraced new information technologies, including artificial intelligence (AI), to facilitate and improve decision-making processes. Nowadays, intelligent machines endowed with learning, reasoning and adaption capabilities are achieving unprecedented levels of performance when solving increasingly complex problems [1]. For instance, in the healthcare domain, analysis of data can drive improvements in care quality and efficiency, earlier disease detection, or fraud detection [2], [3]. As automated discovery of patterns in large amounts of data is a core component of many activities, AI is applied in a growing number of diverse areas such as computational biology, law and finance [4]. In some areas, the sophistication of AI-powered systems has increased to an extent where almost no human intervention is required for deployment [1]. The wide-spread use of intelligent systems and automation has many benefits but is also coupled with significant challenges. Whenever decisions derived from such systems affect humans' lives, there is an emerging need to understand how such decisions are derived by AI techniques [1]. While the very first AI systems were easily interpretable, the growing complexity of today's systems, which heavily rely on deep neural networks, makes it hard for humans to understand their inner workings [4]. As intransparent black box machine learning (ML) models are increasingly being deployed, there is a growing danger on making automated decisions that are not justifiable, legitimate, or simply do not allow obtaining detailed explanations of their rationale [1].

Recently, this has motivated many research efforts in the emerging field of e**x**plainable **a**rtificial **i**ntelligence (XAI) to focus on the development of models, methods, and interfaces that are understandable to human users by offering a means of introspection or some notion of explanation [5]. While most works focus on computational approaches to provide explanations, only limited research efforts have assessed the quality of explanations based on their evaluation by users [5]. Usually, the AI or ML community focuses on functional evaluation to investigate technical feasibility of different methods [5], however, it remains an open question how to determine a formal definition of a correct or best explanation to perform a systematic and rigid evaluation of XAI methods. While Adadi et al. [6] found that only 5% of surveyed papers evaluate XAI methods and quantify their relevance, Nunes et al. [7] state that 78% of analyzed papers on explanations in decision support systems lack structured evaluations. Other works have addressed the design and conduction of explanation evaluation in XAI. Gilpin et al. [8] conduct a survey including explainable methods for deep neural networks and categorize evaluation approaches based on different stages of the ML deployment process. On the other hand, Yang et al. [9] propose a framework consisting of multiple levels to evaluate explanations.

Regarding the investigation of *best* explanations, some studies rely on user evaluation, namely users' subjective opinions expressed in surveys or interviews. Different types of measurements have been proposed, measuring user satisfaction [10], acceptance [11], trust [12] or the goodness of an explanation [13]. Miller et al. [14] state that humans are more likely to accept explanations that are consistent with their prior beliefs. Moreover, simpler or more generalizable explanations are often preferred by users.

The aim of this paper is to perform a systematic and rigid

user evaluation and comparison of different XAI methods. More precisely, we perform a benchmark study of a set of selected XAI methods surveying students with and without any expertise in the field of AI and ML to investigate *which of the explanations provided by different XAI methods is preferred by these users*. Moreover, while addressing this research question, we want to test two hypothesis. The first hypothesis addresses the trust in a intelligent system. It has been shown that if users are able to assess the reliability of a system based on their own perception of system accuracy, the resulting trust in the system leads to higher reliance on the system by the user [15]. Therefore, we hypothesize that *explanations increase users' trust in a system* (Hypothesis A). Secondly, we hypothesize that a specific category of XAI methods is preferred by users. Generally, there is a distinction between local and global XAI methods. Whereas local explanations explain the reasoning behind an output for an individual instance or a user query, global explanations aim to describe the general behaviour of a model, i.e. for all instances. Regarding familiarity of a user with a topic, the literature suggests that local explanations are thought of as less overwhelming to AI novices [15], as people with little to no AI knowledge. Therefore, we want to test the hypothesis that *AI novices prefer local over global explanations* (Hypothesis B).

The remainder of this paper is structured as follows: Section II describes the methodology followed and introduces the survey design. Section III presents the survey results and Section IV concludes this paper with a summary and some remarks on future research directions.

II. Methodology

In this work, we followed an experimental design methodology. Accordingly, we designed a questionnaire guided by six key design decisions (D1 - D6):

- **D1** Define a suitable target audience to assess different aspects of explanations produced by XAI methods.
- **D2** Define a use case interesting for the target audience and suitable to perform a comparison of different XAI methods, and select a data set that depicts this use case.
- **D3** Select XAI methods to be assessed by the target audience.
- **D4** Select a black box ML model that will be explained.
- **D5** Define aspects of the explanations generated by selected XAI methods to be evaluated by the target audience.
- **D6** Decide how the identified evaluation criteria (D5) can be assessed by the target audience in a questionnaire.

A. Target User (D1)

For this study, we surveyed university students. We subdivide them into AI novices and AI experts according to their background and thus their knowledge of AI stated in the questionnaire. We define novices as students with little to no expertise in AI and ML, whereas AI experts have a more profound background or specialization in AI. Details on this categorization are given in Section III-A.

We targeted the user group *university students* based on two important criteria. First, we presume that students are knowledgeable in critical thinking which is a requirement for any type of assessment. Secondly, we assume that the majority of students are in their twenties. Thus, all of them are likely to be familiar with AI products and have at least a broad idea of the capabilities of AI and use cases for its deployment. To include people with and without a technical background, enables us to evaluate explanations by a wider audience. Additionally, the specified target users are very accessible to us since we as students ourselves belong to the target audience. Students are also frequently part of a research environment and are more open to answer surveys and participate without the implementation of strong incentives. Hence, this design decision is made to reach a large population with the survey and maximize the number of participants.

B. Use Case (D2)

To test the quality of generated explanations through the questionnaire, we aimed to present a use case with questions in a contextual setting. In a survey, a greater number of participants results in more insightful and meaningful conclusions derived from the survey. Typically, not every person inquired answers a questionnaire. Therefore, we belief that a use case relevant and related to the target user can positively impact the number of participants. Consequently, we specifically address a use case that is interesting and understandable to the target user group. We chose to emulate a plausible implementation of an AI system used in the *admission process of students for graduate schools*. Specifically, the system predicts whether a potential student is to be admitted into a graduate program. For this use case, we used the "Graduate Admission" data set [16]. The data set contains students' acceptance rates into highly ranked graduate programs based on their performance in the Graduate Record Examinations test (GRE score), their cumulative grade point average (CGPA), letter of recommendation score, statement of purpose score, research experience, and the rating of their university.

C. XAI Methods (D3)

With the emerging need for XAI, many XAI methods have been proposed [1]. In our study, we decided to focus only on model-agnostic XAI methods, which can generate explanations for any type of black box model. Model-agnostic methods can be further distinguished into local and global methods. In our study, we included four XAI methods: two local methods (LIME, local SHAP) and two global methods (global SHAP, PDP). This selection allows us to investigate whether there is a user preference towards local or global methods. Two of the most widely applied XAI methods in recent research on the topic of explainability and interpretability are LIME [17] and SHAP [18]. Due to their great popularity, we included both LIME and SHAP in the study.

LIME[1], short for ***l**ocal **i**nterpretable **m**odel-agnostic **e**xplanations*, focuses on training a local surrogate model to approximate the prediction of the underlying black box model

[1]https://github.com/marcotcr/lime

as closely as possible for a single instance. In contrast to the black box model itself, these local surrogate models are transparent or explainable. Therefore, these explanations are used to explain individual predictions of the black box model.

SHAP[2] is short for ***sh**apley **a**dditive ex**p**lanations*, another model-agnostic post-hoc explanation approach. The goal of SHAP is to explain the prediction of an instance by computing the contribution of each feature to the prediction. In order to do this, the SHAP explanation method computes Shapley values that originate from coalitional game theory.

PDP[3], short for ***p**artial **d**ependence **p**lot*, visualizes a model's decision boundary as a function of a specific input feature. The plots enable users to gain some insight about the model's average behavior as values for the different features change [4].

Since SHAP can be utilized as a local and a global technique, we decided to include both in the benchmark study. Hence, in terms of local methods, we wanted to compare LIME and SHAP. To make an equivalent comparison for global methods, we decided to include SHAP and PDP.

D. Black Box Model (D4)

Since we aimed to include only model-agnostic XAI methods in our study, we considered to generate explanations for different black box models. To limit the scope of this study, it was decided to include only one model since this allowed for the direct comparison of generated explanations, without taking different model performances into account. In our study, we used a random forest model [19] that achieved an accuracy of 86.6%

E. Evaluation Criteria (D5)

In accordance with the evaluation approach proposed by Hoffman et al. [20], we decided to use five selected aspects of broadly understood goodness of explanation: *understandability*, *usefulness*, *trustworthiness*, *informativeness*, and *satisfaction*. All five have been part of multiple research works revolving around XAI evaluation [5], [20], [21], and can be used for evaluation by asking the following questions:

- **Understandability**: From the explanation, does the user understand how the model makes a decision?
- **Usefulness**: Is the explanation useful to the user, to make better decisions or to perform an action?
- **Trustworthiness**: Does the explanation increase the user's trust in the model?
- **Informativeness**: Does the explanation provide sufficient information to explain how the model makes decisions?
- **Satisfaction**: Does the explanation of the model satisfy the user?

The perceived fulfilment of each criteria is measured by a user's agreement or disagreement indicated through his or her rating on a 7-point likert scale.

[2]https://github.com/slundberg/shap

[3]https://github.com/SauceCat/PDPbox

F. Questionnaire Design (D6)

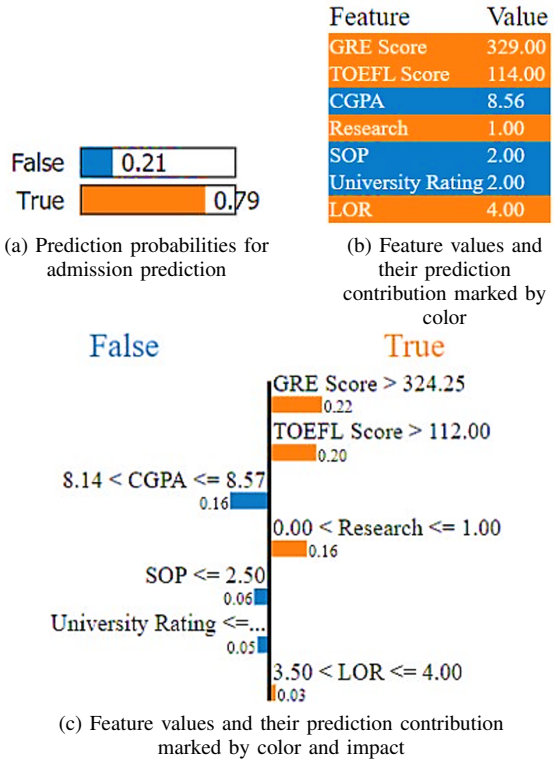
The designed questionnaire is composed of five individual sections. The first section focuses on background questions that are required to differentiate AI novices and AI experts. The remaining four sections focus on the four XAI methods, each comprising a short story or scenario which sets the narrative, and, therefore, helps contextualize the explanation that follows. Each story was created to fit a plausible scenario in which a model's explanation could naturally be included.

For instance, Figure 1 provides information relevant to the story and Figure 2 shows the visual explanation obtained with the LIME method for the following scenario: *"Mary has recently finished her undergraduate program and has begun to think about whether she would like to immediately enrol in a graduate program or look for a job instead. She would be willing to commit the time to apply for a graduate program if the odds of being accepted were favourable. Mary had the feeling that her high GRE scores and glowing letter of recommendation would make up for her poor GPA. To help her decision-making process, she decided to reach out to an education consultancy that could help her identify her prospects of being accepted for a graduate program. The education consultancy used an AI system based on historic data to evaluate the chance of students being accepted. She was asked to provide the following information in order to receive an evaluation."*

	Scale	Mary´s Score
GRE Score	0 - 340	329
TOEFL Score	0 - 120	114
University Rating	0 - 5	2
Statement of Purpose Strength (SOP)	0 - 5	2
Recommendation Letter Strength (LOR)	0 - 5	4
CGPA	1 - 10	8.56
Research Experience	0 or 1	1

Fig. 1: Information about the fictional character Mary appearing in the scenario presented in Section II-F.

The intention was to engage the participants with the subsequent explanation by giving additional contextual information. The stories were kept to a maximum of 7-8 sentences to keep the questionnaire short time-wise while still providing some relatable context. Transitioning from one scenario to the next navigated participants through the survey, naturally introducing them to new explanations one by one while keeping them engaged until the end. For all XAI methods, the explanations were presented in a visual form as generated by the individual methods. We tried to keep the visualizations as close to the default visuals produced by the XAI method as possible. However, we made small modifications, namely added a small description to the plot axes in the original visualization. The additional information was considered necessary to ensure an evaluation of the explanations with as little ambiguity as possible, especially since we targeted people with potentially no technical knowledge. This type of modification can be seen



(a) Prediction probabilities for admission prediction

(b) Feature values and their prediction contribution marked by color

(c) Feature values and their prediction contribution marked by color and impact

Fig. 2: LIME explanation generated for the scenario outlined in Section II-F.

as an incentive which, demonstrated by Dieber et al. [21], proved to remove uncertainties. The targeted time to complete the survey was a maximum of 15 minutes in total.

G. Survey Process

The questionnaire was implemented in Google Forms. The first version of the questionnaire was then sent to five different students. The students were asked to fill out the questionnaire and to provide critical feedback. The intention behind this "trial run" was to get an impression about how the questionnaire was perceived by the target users, if any wording or visuals were incomprehensible, and whether the intended completion time of 15 minutes was a realistic estimation. The received feedback was evaluated and based on the findings the questionnaire was improved. For instance, some of the feedback addressed a lack of additional information accompanying the explanations to ensure that the explanation could be understood, especially by people who had no technical background or no prior AI knowledge. Afterwards the questionnaire was distributed through multiple channels, including personal contacts such as friends, fellow students and students from other faculties at Maastricht University, as well as subject mailing lists, and contacts of academic teachers. The survey was conducted within a time span of three weeks during which 60 target users participated.

III. Results

A. Respondents' Background

In order to determine the level of experience in the field of AI, the respondents were asked in the first section of the questionnaire about the number of courses related to AI (e.g., "Introduction to Machine Learning" or "Pattern Recognition") they participated in, including all courses on concepts associated to machines simulating human behavior such as planning, learning, and reasoning. The distribution of respondents' answers is illustrated in Figure 3.

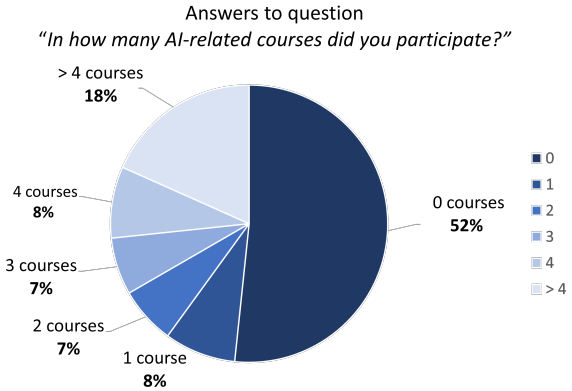


Fig. 3: Distribution of number of AI-related courses attended among 60 respondents.

Based on the number of attended courses, the respondents were divided into two groups: AI novices and AI experts. An AI novice was identified as a person who followed at most one course related to AI. Respondents who took more than one course related to AI during their studies were categorized as AI experts. In this way, it is possible to make a distinction between people with different backgrounds and presumably different characteristics, which in turn potentially influence their perception and interpretation of the explanations to be evaluated. Based on the answers, 60% of participants were labeled AI novices (at most one AI related course) and 40% AI experts. Moreover, an additional background question revealed that 40% of the students have previously heard of or acknowledge the existence of explainable AI methods.

B. Respondents' Evaluation of the XAI Methods

In the last four sections of the questionnaire, respondents assessed the XAI methods — LIME, SHAP local and global, and PDP — in the context of the presented scenarios. In Table I, we can observe the repeating pattern of SHAP performing poorly in comparison to LIME and PDP. Overall, PDP always performs best, followed by LIME, whereas SHAP local and global perform worst. A difference between SHAP local and global is observed in the spread of responses across the likert scale represented by the standard deviation. Evaluation scores for SHAP global are more widely spread compared to SHAP local. This reveals more extreme responses of both agreement and disagreement for SHAP global than for SHAP local. In

TABLE I: Mean score on 7-point likert scale with standard deviation for all evaluation criteria.

	Understandability	Usefulness	Trust	Informativeness	Satisfaction
LIME	4.77 ±1.61	4.79 ±1.49	4.74 ±1.66	4.33 ±1.74	4.08 ±1.68
SHAP (local)	4.03 ±1.61	3.90 ±1.53	3.83 ±1.55	3.37 ±1.59	3.50 ±1.47
SHAP (global)	4.00 ±1.85	3.77 ±1.93	3.85 ±2.02	3.54 ±1.78	3.50 ±1.89
PDP	**5.28** ±1.59	**5.25** ±1.64	**4.84** ±1.79	**5.10** ±1.60	**5.08** ±1.64

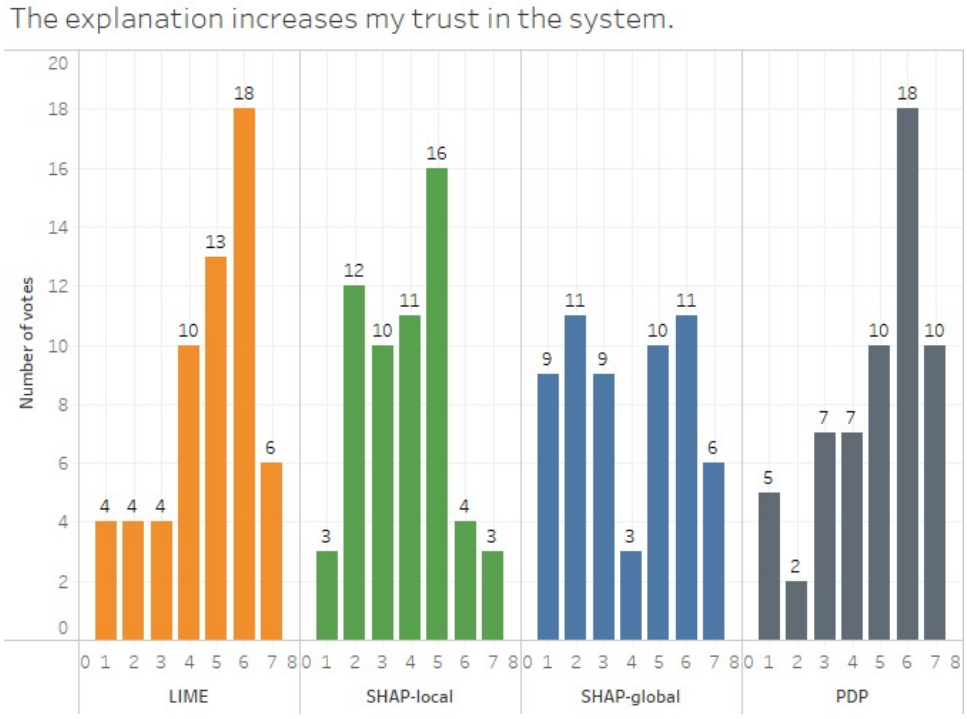


Fig. 4: Evaluation of *trust* evaluation criterion for all XAI methods on 7-point likert scale.

comparison, the standard deviations for both PDP and LIME are smaller and relatively similar.

C. Users' Trust in a Explained System (Hypothesis A)

Hypothesis A is that *explanations increase users' trust in a system*. The intuition behind this hypothesis is that a ML model is expected to be trusted more by students when its prediction is complemented with an explanation. As stated by Ribeiro et al. [17], trust is crucial for effective human interaction with AI systems.

TABLE II: Mean score on 7-point likert scale with standard deviation for *trust* evaluation criterion.

	Mean
LIME	**4.74** ±1.66
SHAP (local)	3.83 ±1.55
SHAP (global)	3.85 ±2.02
PDP	**4.84** ±1.79

The results in Table II seem to indicate that the explanations provided by PDP and LIME increase the trust in the AI system. In contrast, the SHAP methods received a lower degree of confidence which is below the neutral score of 4 on the 7-point likert scale.

When the distribution of responses for trust is examined in the histograms of Figure 4, the indication is confirmed. Results for both LIME and PDP show distributions skewed to the right. This illustrates that the majority of participants agree to the explanation increasing their trust to some degree. In comparison, a worse performance of SHAP can be observed, with SHAP local receiving responses centered around neutrality and SHAP global receiving both high agreement and disagreement from participants. The high variance for SHAP global suggests a divergence of participants' opinion regarding increased trust resulting from the explanation. Probable causes for this variance may be attributed to the complexity of visualization that leads to confusion for both AI novices as well as experts.

It can be concluded that when coupling predictions with explanations from LIME or PDP, users' trust in the AI system is marginally increased. Despite an increase in trust being revealed, it is not as significant as expected. In the case of SHAP, a score lower than the neutral score is observable, indicating that respondents tend to slightly disagree that explanations

increase their trust in a system.

D. User's Preference for Local vs. Global XAI (Hypothesis B)

Hypothesis B is that *AI novices prefer local over global explanations*. As local explanations aim to explain a model's reasoning behind the results for an individual user query, Mohseni et al. [15] suggest that this type of explanation is thought to be "less overwhelming for novices". Therefore, we investigate whether this hypothesis holds according to our target users' evaluation, or not.

Responses of AI experts were not included in this particular analysis, as the hypothesis specifically addresses the preference of AI novices. Hence, only the responses from the 36 AI novice respondents were included. In order to test the hypothesis, we decided to assess each XAI method, local and global, with regard to each of the five evaluation criteria, to find significant differences in preference for AI novice users. The results are illustrated in Table III.

TABLE III: AI novices' evaluation (mean likert scores) of local and global methods with Welch's t-test indicating a significant difference in mean scores if p-value < 0.05.

	Local	Global	P-value (Welch)
Understandability	4.43	**4.56**	0.69
Usefulness	4.31	**4.40**	0.77
Trustworthiness	**4.46**	4.45	0.98
Informativeness	4.09	**4.29**	0.53
Satisfaction	3.91	**4.21**	0.35

A preference is indicated by higher means in either of the first two columns of the table. The results disprove the hypothesis, revealing no indication of a preference of AI novices for local methods over global ones considering all evaluation criteria. In all cases, with the exception of trustworthiness, the mean score for global methods is greater than for local methods.

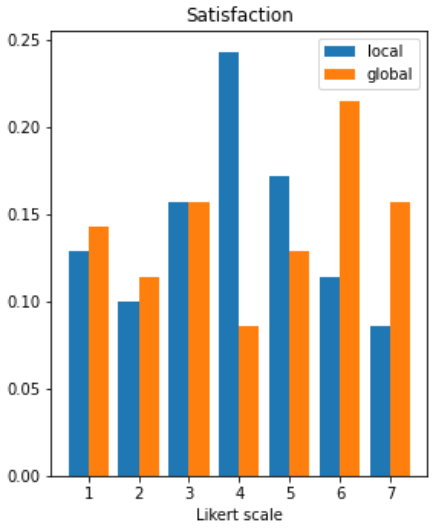


Fig. 5: AI novices' evaluation local and global methods regarding *satisfaction* (proportion of responses on y-axis)

Looking at the distribution of scoring for the satisfaction criteria among all novices, as shown in Figure 5, this observation seems to be confirmed. The blue bars indicate the scoring for local methods, whereas the orange colored bars mark the performance of global methods. The majority of AI novices' assessments of local methods are gathered in the middle of the likert scale. Contrary to this, the distribution for global assessments seems to be tilted slightly toward the right revealing a higher satisfaction evaluation compared against local methods. This diverging characteristic of the evaluation distribution is most visible for the satisfaction criteria.

E. Other findings

Apart from the explicitly stated research question regarding user preference and the two hypotheses, our analysis of the results revealed additional and surprising findings. For instance, we observed a difference in perception of methods between AI novices and AI experts. This can be deduced from Figure 6 in which the results for SHAP local and SHAP global are compared. For local explanations, ratings from both AI novices and experts are centered around the neutral score. For global explanations, a difference between novices' and experts' rating is clearly visible. Specifically, AI novices give this explanation a negative rating, whereas AI experts tend to give more positive feedback.

In general, both target groups seem to recognize the complexity of SHAP local, which describes a single instance on one dimension. The results obtained for SHAP global support the suggestion from Mohseni et al. [15] of AI novice users potentially being overwhelmed by global explanations. Therefore, these findings highlight the need for explanations tailored to the respective audience.

IV. Conclusion

In this paper, we studied the perception of explanations generated by different XAI methods to explain the reasoning behind AI systems' decision-making. The emerging need for XAI stems from the assumption that XAI will play a fundamental role in the further spread and future deployment of AI systems. In our comparison between different XAI methods — LIME, SHAP local and global, and PDP — PDP performed best over all included evaluation criteria with the majority of responses showing fulfilment of the criteria. Second best performed LIME, showing less agreement and more neutrality with fulfilling individual criteria, while SHAP local and global performed the worst with more responses of neutrality and disagreement.

A closer look at the criteria of *trust* did not suffice to make a conclusion about a significant increase of trust related to the provided explanations. Regarding the hypothesis that AI novices prefer local or global methods there were no significant differences, thus, no preference was identified.

In the future, we would like to study the differences in results by research conducted with a narrower definition of AI experts. Current results indicate that PDP is preferred by our

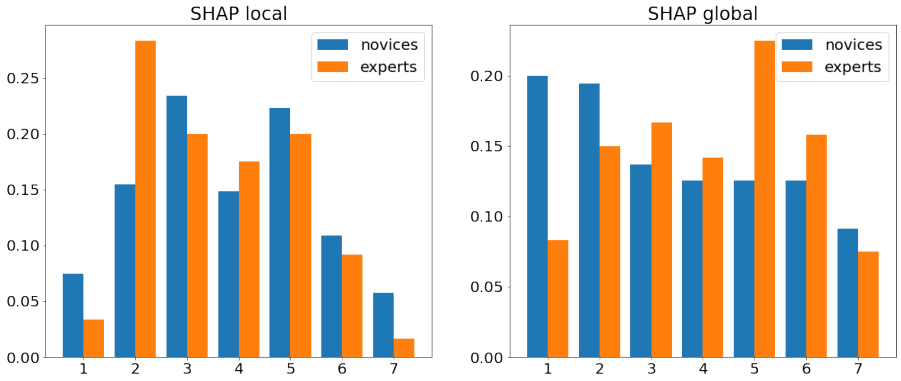


Fig. 6: AI novices' and AI experts' evaluation of SHAP local and global (proportion of responses on y-axis)

users. It may be also interesting to learn why, specifically, this method received better scores than the other XAI methods.

REFERENCES

[1] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[2] Institute for Health Technology Transformation, "Transforming health care through big data: Strategies for leveraging big data in the health care industry," 2013.

[3] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, vol. 2, no. 1, pp. 1–10, 2014.

[4] V. Belle and I. Papantonis, "Principles and Practice of Explainable Machine Learning," *arXiv preprint arXiv:2009.11698*, 2020.

[5] M. Chromik and M. Schuessler, "A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI." in *ExSS-ATEC@ IUI*, 2020.

[6] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[7] I. Nunes and D. Jannach, "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems," *User Modeling and User-Adapted Interaction*, vol. 27, no. 3, pp. 393–444, 2017.

[8] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[9] F. Yang, M. Du, and X. Hu, "Evaluating Explanation Without Ground Truth in Interpretable Machine Learning," *arXiv preprint arXiv:1907.06831*, 2019.

[10] M. Bilgic and R. J. Mooney, "Explaining Recommendations: Satisfaction vs. Promotion," in *Beyond Personalization Workshop, IUI*, vol. 5, 2005, p. 153.

[11] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining Collaborative Filtering Recommendations," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 2000, pp. 241–250.

[12] J. Zhou, Z. Li, H. Hu, K. Yu, F. Chen, Z. Li, and Y. Wang, "Effects of Influence on User Trust in Predictive Decision Making," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[13] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl, "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 81–87.

[14] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[15] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *arXiv preprint arXiv:1811.11839*, 2018.

[16] M. S. Acharya, A. Armaan, and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE, 2019, pp. 1–5.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[18] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[19] L. Breiman, "Random Forests," *UC Berkeley TR567*, 1999.

[20] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[21] J. Dieber and S. Kirrane, "Why Model Why? Assessing the Strengths and Limitations of LIME," *arXiv preprint arXiv:2012.00093*, 2020.