Towards Automated Human-Centered Recommendation of Explainable AI Solutions*

Nils Ole Breuer[1,2,*], Sahin Albayrak[1,2]

[1]GT-ARC gGmbH, Ernst-Reuter-Platz 7, 10587 Berlin
[2]DAI-Labor, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin

Abstract

Finding a suitable XAI method from the many XAI possibilities for a specific use case is a non-trivial task. There are recommendation algorithms for recommending XAI methods. However, these are often based solely on the underlying ML model's technical characteristics and do not consider the needs of the target group. Also, these systems often recommend only off-the-shelf XAI methods, frequently failing to achieve the desired explanatory goal. We therefore introduce an automated recommendation framework that tackles both of these problems. On the one hand, we created a low-threshold process in which the needs of the target group can be captured in natural language. On the other hand, we recommend *XAI Solutions* that include both a suitable XAI method and actionable human-centered design guidelines, which describe how the explanation should be adjusted to be useful for the target group. Our recommendation framework consists of a customized GPT that offers suitable *XAI Solutions* based on the given design principles and an XAI database. We evaluate our recommendation framework in two real-world scenarios. The evaluation shows that it can generate human-centered *XAI solutions* that meet the needs of the target group.

Keywords

Explainable AI, XAI, Recommendation framework, Interactive explanations

1. Introduction

The advent of accessible AI technology presents a significant opportunity for individuals without a technical background to realize their ideas for AI systems. While this democratization of AI signifies a beneficial advancement, as it fosters the creation of new interdisciplinary AI systems incorporating a diverse range of ideas, it also introduces risks. At the same time, it emphasizes the need for explainable and transparent AI. This underlines the central role of explainable AI (XAI) in ensuring transparency and control.

Despite the potential benefits of XAI, there are two challenges for applying XAI methods. On the one hand, it is very difficult to select the most suitable method for a specific use case from the large and confusing range of XAI methods. On the other hand, off-the-shelf XAI methods are usually not easily applicable, as explanations for AI systems usually have to be generated in a very context-specific way and the explanations have to be adapted to the specific target group

Multimodal, Affective and Interactive eXplainable AI Workshop, 27th European Conference on Artificial Intelligence

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.
nils.breuer@gt-arc.com (N. O. Breuer)

to be interpretable. Consequently, the laudable intentions to make the AI system as transparent and explainable as possible usually fail because an unsuitable XAI method was selected or the target group does not understand the explanation, as the helpfulness in specific use cases is questionable [1, 2, 3, 4, 5, 6].

Our analysis reveals two problem areas in the practical application of XAI methods: first the selection of an XAI method and second the adaptation of the explanation to a specific target group. There are solutions for both problem areas. Recommendation algorithms exist to support the search process for a suitable XAI method. These recommendations are mostly based on the technical details of the ML model and not on the audience to whom the explanation is directed [7, 8]. Additionally, there is an awareness among researchers in the XAI field that explanations should be tailored to the audience to be understandable and useful. The research direction of human-centered XAI (HCXAI) addresses this problem and has already produced useful concepts [9, 10, 11]. However, these concepts have rarely been formulated as practical actionable guidelines so that they can be quickly and easily applied for an explanation.

In this work, we aim to connect and address these two problems with a new human-centered recommendation framework for *XAI Solutions*. The objective of our recommendation framework is twofold. First, to enable practitioners without extensive experience in XAI to easily access a suitable *XAI Solution*. This requires a low-threshold and interactive process to gather relevant information. For this, we implemented a natural language-based procedure to collect information about the target group, i.e., the explainees, and also the technical characteristics of the AI model to be explained. Second, we want to recommend *XAI Solutions* which we define not only as an equivalent for an off-the-shelf XAI method but rather we use the information about the target group to generate a combination of a suitable XAI method and actionable human-centered design guidelines that are based on HCXAI concepts to make the explanation as understandable as possible for the target group. For this, we formulate human-centered design principles based on theories of HCXAI and social sciences. Additionally, we build an XAI database with the most relevant XAI methods.

We create a custom GPT for the recommendation process, using the options provided by OpenAI's ChatGPT platform. The custom GPT uses knowledge from our human-centered design principles and knowledge from our XAI database to recommend a custom *XAI Solution* according to our definition.

In summary, our work provides the following contributions:

- We formulate human-centered design principles that are based on socio-cognitive theory and HCXAI [12, 13] with which explanations of XAI methods can be easily adapted.
- We propose an automated process that allows a wide range of individuals, including non-XAI-experts to receive recommendations for an *XAI Solution* for their use case, grounded in human-centered principles.
- We create a custom GPT that can use pre-defined knowledge to generate customized *XAI Solutions* for a specific target group.

2. Related Work

2.1. Human-Centered XAI (HCXAI)

Many explanations of XAI approaches currently rely on an algorithm-centric perspective and are therefore based on the intuitions and explanatory objectives of XAI researchers. These perspectives and explanatory objectives diverge significantly from the requirements of layperson-friendly explanations [11]. This leads to a dissonance between the theory of XAI methods and their practical application [14]. For this rationale, researchers posit that in crafting explanations for AI systems, primacy should be accorded to the human recipient for whom the explanation is intended [15]. This concept led to the research field of human-centered XAI (HCXAI).

The idea that many HCXAI scientists pursue is that explanations will be more understandable and useful if they correspond to the social and cognitive processes of human beings. This human-centered perspective also builds on the article of Miller [12], which examines socio-cognitive theories and XAI through extensive analysis.

HCXAI has developed many theories and principles to date, as can be seen from the sheer number of recently published articles [10, 11, 16, 17, 18, 9].

These theories and principles have already been put into practice in some studies [19, 20, 21, 22, 23].

2.2. XAI Method Recommendation

The problem of selecting a suitable XAI method is ubiquitous in the literature. The usual approach is to use one of the many available XAI libraries [24, 25, 26] and then select a method that fits the model architecture. Another option is to make a selection of XAI methods and then analyze them using technical evaluation methods and select the "best" one [8, 27]. These processes require a high level of expertise and are also demanding for data scientists and machine learning engineers.

Based on interviews with data scientist Retzlaff et al. [28] introduce a decision tree-based approach for selecting the most suitable XAI method for a specific use case. The decision tree enables data scientists to understand the tradeoffs between different XAI methods and also shows the user how methods can be combined to ensure the best possible use.

One of the few works that aim to automate the recommendation process is the *AutoXAI* framework by Cugny et al. [7]. In this framework, a user can specify different context variables, technical data about the ML model, and the data set. The *XAI Question Bank* from Liao et al. [16] is then used to select a suitable XAI method. In addition Caro-Martínez et al. [29] created a holistic platform for the recommendation of personalized XAI experiences based on a case-based reasoning approach with an ontology. For Caro-Martínez et al. [29] an explanation experience consists of a solution to an explainability problem and an evaluation of the proposed solution.

Studies that focus more on the needs of the target group such as [30, 31] offer helpful analyses of the target group and also conceptual approaches to what should be considered to recommend an XAI method in a human-centered way. The actual matching and recommendation process in these studies are left to future work.

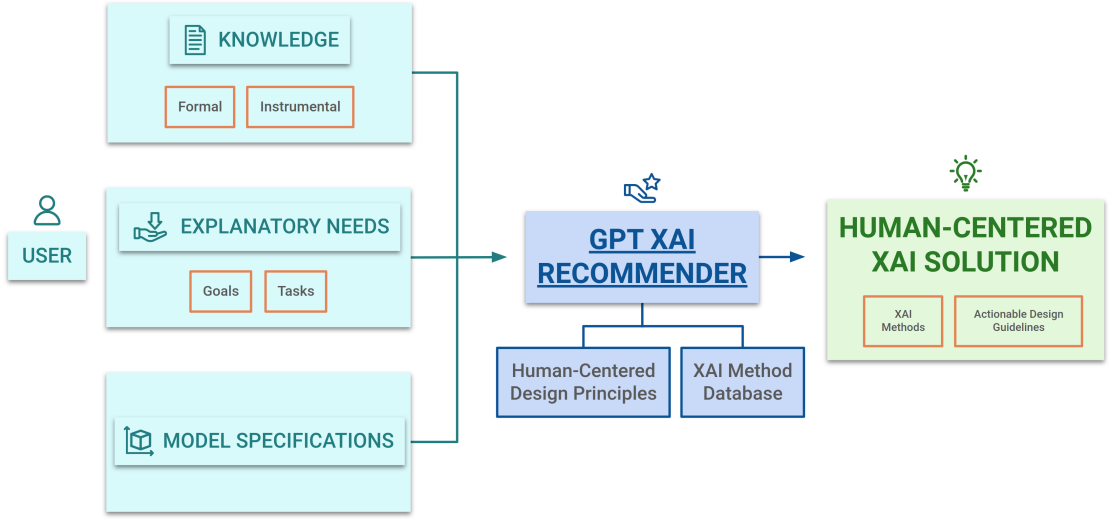

Figure 1: Our Recommendation Framework: On the left side is the input (information about the target group) that the user provides to the recommendation GPT (turquoise), in the middle is the custom recommendation GPT (blue) that uses our formulated actionable human-centered design principles (3.1) and the XAI method database (3.3, and on the right the final *XAI Solution*.

3. Recommending Human-Centered XAI Solutions

In this section, we delve deeper into the individual components of our recommendation framework. First, we explain our human-centered design principles, then discuss how we gather information about the target group of the *XAI solution*, how we map this information to XAI methods, and finally, how the individual components are processed through our recommendation procedure. Figure 1 shows an overview of all the components of our recommendation framework.

3.1. Human-Centered Design Principles

As already described in the introduction, existing XAI recommendation algorithms [7] are limited to simply recommending the raw XAI methods, mostly based on the technical characteristics of the AI application. Evaluation and user studies have shown that these are often misleading and do not lead to a causal understanding of the ML model [1, 6]. That is why we refer to *XAI Solutions* in this paper and not just *XAI methods*. As stated above we define an *XAI Solution* as a comprehensive recommendation that encompasses not only an XAI method but also actionable human-centered design guidelines, which can be used to tailor the explanation to benefit the target group, leading to a deep understanding of the behavior of the ML algorithm.

From the insights of human-centered XAI research [11] and social sciences [12], we derive design principles for four categories [13, 32, 23] that are relevant for understanding an explanation. The *design principles* are therefore a broad collection of theories, which are then adapted by the recommender to form *actionable guidelines* for the explainee target group. The guidelines for the four categories can then be used to easily tailor the explanation to be more interpretable

for the target group.

- **Communication.** Social science theories indicate that explaining is a form of social dialogue [12]. User studies in the XAI domain also show that explainees primarily seek some form of social dialogue in explanations or prefer explanations to be supplemented with verbal descriptions [17].
- **Interactivity.** Studies have shown that the mere presentation of diagrams as explanations is an obstacle to interpretability, for both experts and non-experts [33]. Research in HCXAI indicates that it's preferable to design explanations for an AI system as an interactive process in which the behavior of the algorithm can be understood [5, 34].
- **Selectivity / Complexity.** The social sciences have shown that explanations are selective, which means that the explainer only selects the most important and relevant causes that are necessary to form an explanation [12]. To apply this to explainable AI, a way should be found to query the needs of the consumer of the explanation before the explanation is generated in order to select the most important parts of the explanation [22].
- **Customizability.** Studies have shown that if the person receiving the explanation has the opportunity to personalize the explanation to their level of knowledge, mental model, and preferences, it can have a positive impact on the understanding of the explanation [35]. These personalizations could be, on the one hand, that the complexity or the form of presentation can be adapted independently.

The full design principles for all categories can be found in Appendix A. These design principles form one of the knowledge files (see Figure 1, blue box) that the recommendation GPT uses to generate the actionable design guidelines of the *XAI Solution*.

3.2. Background Knowledge and Explanatory Needs of the Target Group

As can be seen from Figure 1, the recommendation GPT receives information about the background knowledge and the explanatory needs of the target group as input.

Background Knowledge In many studies where a target group is characterized to create more personalized explanations, strict stakeholder groups are defined, each with specific needs for an explanation [36, 37]. However, this rigid classification is outdated due to the widespread use of AI applications in society. Therefore, for our recommendation framework, we use a more detailed analysis to identify the needs of the target group. To do this, we reference the *expertise* definition of [38], which was adapted within a framework for XAI purposes by [31]. Following the framework [31] we decompose expertise into types of knowledge in specific contexts. The types of knowledge relate to *formal* knowledge, e.g. familiarity with theories, *instrumental* knowledge, e.g., programming experience, and *personal* knowledge, e.g., information that a person knows from the media. To retrieve this information about the explainee group we formulate open-ended questions that the user of the recommendation framework should answer in natural language. We provide exemplary answers for each question to guide the answers in the right direction. All questions can be found in the Appendix B.1. By allowing users themselves to describe the target group in their language, we generate a much more accurate

description of the target group. However, through our example questions, we still guide them in a specific direction that aligns with our design principles.

Explanatory Needs Further important information about the target group includes the needs they have for an explanation. Explanatory needs can be divided into *tasks* that the target group wants to solve with the explanation and *goals* that the target group wants to achieve with an explanation [31]. *Tasks* correspond to low-level questions that the target group wants an answer for, e.g., what features are most important for the ML model? *Goals* are more high-level objectives that the target group wants to achieve, e.g., ensure that the ML model complies with regulations and laws. Just as for the collection of background information, we formulated open questions that the user of the recommendation framework should answer in their own words. The questions and sample answers can be found in the Appendix B.2.

3.3. XAI Database

In addition to the actionable human-centered design guidelines, XAI methods are also part of our *XAI Solution*. For this purpose, we created an XAI database with common XAI methods. The database defines for each method technical characteristics with which the method is compatible. Beyond that, we also match goals and tasks to the XAI methods so that the explanatory needs of the target group can be used to find a suitable XAI method. For example, the above-mentioned task (*What features are most important for the ML model?*) is assigned to a feature attribution method.

3.4. Recommendation Procedure

Due to the high context specificity, the individual requirements of the target group, and also the many different types of ML models, it is very challenging to create a structured selection process of XAI methods and human-centered design principles. Many of our attempts with structural databases failed because we always identified a new use case with a new alternative ML model in combination with specific requirements of a target group for which no suitable XAI method in combination with well-founded design guidelines could be found. This is why we have opted for a more open approach. This approach is reflected in the way the information is collected, namely in the form of guided free text.

To match the freely formulated explanatory needs and background knowledge of the target group with actionable human-centered design guidelines and a suitable XAI method from our XAI database, we utilize the remarkable association possibilities and few-shot learning capabilities of OpenAI's large language model GPT-4 [39]. We create a custom *XAI Solution* recommendation GPT. For this, we formulated a detailed instruction prompt explaining the recommender's procedure, and also we provided few-shot prompting examples. In Figure 1 it can be seen that we additionally provide the recommender GPT with two so-called "knowledge files": Our Human-Centered Design Principles (compare Section 3.1) and our XAI Database (compare Section 3.3). As can be seen in Figure 1 the recommendation GPT considers the explanatory needs, and the background knowledge of the target group. Furthermore, it also uses technical aspects of the ML model if they are mentioned in the answers of the user. Even

though the information retrieval process is structured as a question-answering scheme where the users of the framework formulate the knowledge and explanatory needs of the target group by answering questions the recommender GPT does not act as a chat agent. It rather receives all of the answers of the information retrieval process concatenated as one text section. Based on that input it then uses the "knowledge files" to generate an individual *XAI Solution* consisting of actionable human-centered design guidelines and an XAI method. The detailed instruction prompt can be found in Appendix C.

4. Evaluation of XAI Solutions

We evaluate the capabilities of our recommendation GPT with two use cases. As there is still no standardized methodology for evaluating the output of LLM, we have evaluated the XAI solutions of our recommendation GPT in three different ways. Human evaluation is still the gold standard for assessing the quality of LLM output. Because conducting an extensive human-evaluation study is time-consuming we limited ourselves to a qualitative analysis of the generated *XAI solutions*. In addition, we use two automated evaluation metrics to assess the quality of the design guidelines of the XAI solutions. An embedding-based metric for semantic textual similarity called SemScore [40]. And a LLM-based evaluation method called G-Eval [41].

4.1. Use Case 1: Oncology AI

The first scenario that we analyzed is that of a senior doctor working in the field of oncology. Recently, the hospital where she works decided to install a new type of AI algorithm to help detect early breast cancer. However, there are concerns among doctors as they have heard of other medical algorithms that have also produced negative aspects. The doctors would therefore like to have an explanation for the algorithm's decision so that they can evaluate whether the explanation is flawed and at the same time create confidence in working with the algorithm. The whole information retrieval process can be viewed in the Appendix D.1 in addition to the *XAI Solution* provided by our recommendation GPT. In Figure 2 an excerpt of the information retrieval process for the formal knowledge of the target group is visualized.

4.2. Use Case 2: Extreme Weather AI

In the second scenario, a group of ML engineers is searching for an explanation for an ML algorithm with which extreme weather events can be predicted based on climate data. The problem is that the engineers do not have any formal and instrumental knowledge about the weather data. Therefore, they want to use the explanation to improve the model and also understand the climate data in more detail. Again the whole fictive interaction and the *XAI Solution* can be found in Appendix D.2. In Figure 3 we visualized the information retrieval process of the instrumental knowledge and the explanatory needs of the target group.

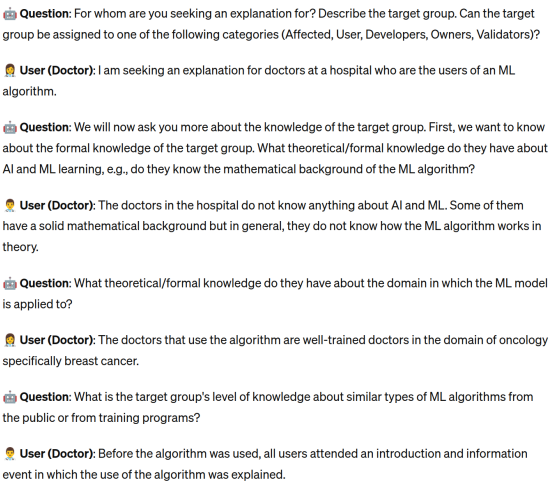

Figure 2: Information retrieval of formal knowledge of target group *doctors*.

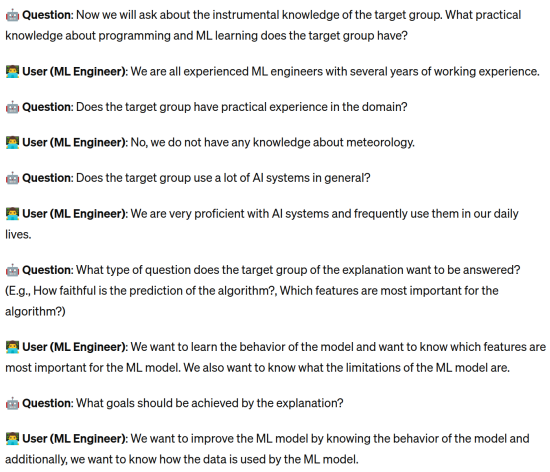

Figure 3: Information retrieval of instrumental knowledge and explanatory needs target group *ML engineers*.

4.3. Qualitative Analysis

The XAI Solution (Appendix D.1, Figure 4) shows that for each of the four explanatory categories, the recommender GPT gives an actionable design recommendation. The recommendation follows our definitions and design principles which we extracted from HCXAI research. For example, the recommendation is to use clear and concise language that integrates medical terms for the explanation. This corresponds to our rationale for using language as a supporting medium for visual explanations. This shows that the recommender GPT uses the provided knowledge file to generate the design recommendation. The selected XAI methods also meet the needs of the target group. Counterfactual explanation methods are especially often used to create a causal understanding of an ML model. Since the goal of the explanation for the target

group is to evaluate how trustworthy the algorithm is, a counterfactual explanation fits well, as it allows decision characteristics to be recognized. These characteristics can then be compared with the physicians' medical expertise to decide whether the prediction is trustworthy or not. From a technical perspective, both recommended XAI methods can also be used for the ML model.

As for use case 1 the *XAI Solution* (Appendix D.2, Figure 5) shows some favorable properties for use case 2 that are based on the needs and goals of the target group. For example, one goal is to understand how the model uses climate data to make its prediction. The XAI Solution, therefore, recommends adding an interactive component to the explanation where the engineers can construct different scenarios and understand how different climate features impact the prediction of the ML model. Additionally, the XAI Solution suggests that the explanation can use highly specific technical language and statistical concepts because the target group consists of ML engineers with several years of working experience. Another objective of the target group is that they want to know which features are most relevant for the ML model, hence the XAI methods recommended in the XAI Solution are suitable, as two of them are feature attribution methods.

4.4. Semantic Similarity Evaluation

We also use two automated evaluation metrics so that the evaluation does not depend solely on our, possibly biased, perception. SemScore [40] compares the semantic content of a model's output and a reference text using embeddings. This method fits well because we want to compare if the design guidelines confirm our proposed human-centered design principles. SemSocre computes a correlation value between the embedding of the model output and the reference text. G-Eval [41] is a framework that uses Chain-of-Thought prompting to assess the quality of LLM output based on some evaluation criteria, in our case semantic similarity on a scale from 1 to 5. Meta evaluations show high human alignment values for both metrics.

Use Case \ Metric	SemScore	G-Eval
Oncology AI	0.555	4.38
Extreme Weather AI	0.675	4.625

Table 1
Results of both semantic similarity evaluation methods SemScore [40] and G-Eval [41]. SemScore has a scale from -1 to 1 and G-Eval has a scale from 1 to 5.

Table 1 shows the results of both metrics SemScore and G-Eval for our two use cases. SemScores are between -1 and 1 where a value close to one means high semantic similarity. G-Eval is an evaluation metric where an LLM assigns a score between 1 and 5 where 5 indicates high semantic similarity. The evaluation results show that we get moderately good correlation scores (*0.555* and *0.675*) with SemScore. This means that the recommended design guidelines of the recommendation GPT align positively with our human-centered design principles. Also, the scores of the G-Eval framework show positive values for the semantical similarity criteria. Table 1 shows the mean values (*4.38* and *4.63*) of 20 trials of the G-Eval framework with the same LLM as a backbone.

5. Discussion

In this study, we embarked on developing a recommendation framework that finds and suggests human-centered *XAI Solutions* for individual use cases. For that, we first had to introduce our notion of *XAI Solution* which is a combination of conventional XAI methods and actionable human-centered design guidelines with which these methods can be tailored to a specific target group. To recommend the design guidelines we formulated human-centered design principles based on research, drawing on studies from human-centered XAI [11] and socio-cognitive analysis [12]. These principles describe how explanations should be created for different target groups so that they are as understandable as possible. Through the formulation of our design principles, we can recommend actionable guidelines.

Because a human-centered *XAI Solution* is highly individual and context-specific, it is almost impossible to implement a hard-coded and structured approach with database queries or decision trees. This conclusion has also been reached by numerous predecessors [30, 31, 7] who have made the theoretical formulations but were unable to implement a recommendation algorithm. Hence we decided to utilize the impressive association capabilities of the large language model GPT-4. We build a customized recommendation GPT that uses among others "knowledge files" as information bases to generate its recommendation.

We evaluated the results of the recommendation framework for two use cases with three evaluation metrics. The qualitative analysis of the XAI solutions shows that the recommender GPT is capable of using the information about the target group to formulate actionable human-centered design guidelines that correspond to the design principles we formulated. The recommendation framework grasps the important parts of the target group information and matches it to the design principles, for example, for the target group *ML engineer* it recommends that statistical concepts and machine learning concepts should be used for the explanation. In contrast, for the target group *doctor*, it is recommended to avoid technical terms of machine learning but rather to convey the explanation in medical jargon. Furthermore, if possible, the explanation should not be centered around ML theories, as the information is not helpful and rather counterproductive for them. The positive direction suggested by our qualitative analysis is also backed by the quantitative evaluation metrics. We focused the automated evaluation metrics on the criteria of semantic similarity. The rationale behind that choice is that we want to assess whether the generated human-centered design guidelines semantically follow our design principles, i.e. have the same meaning. The evaluation with both metrics shows positive results which means that even though both evaluation techniques are based on completely different methodologies our recommendation GPT produces trustworthy actionable design guidelines which are in agreement with our design principles. Furthermore, we argue that the framework pays attention to the needs of the target group when selecting a suitable XAI method. In both of the use cases, the recommender provides suitable help on which XAI method should be used to achieve the explanatory goal. For example, the group of *ML engineer* wants to know which features are most important for the ML model to use this information to improve the ML model. The recommendation framework rightfully suggests using feature attribution methods that can provide precisely this information.

Another distinguishing feature of our recommendation framework is that it is language-based. This opens up the possibility for people who are not necessarily familiar with AI to describe

their needs in their own words. Natural language enables a whole new level of customization and users can describe their use cases much more precisely than approaches that only offer a limited choice for specific use cases [7, 20, 42].

Our recommendation framework thus addresses both issues introduced in the introduction regarding XAI. It assists in selecting an appropriate XAI method, considering not only the technical characteristics of the ML model but also selecting the method based on the knowledge level and explanation goals of the target group. Furthermore, the recommendation framework provides recommendations and guidelines based on well-established theories of HCXAI, on how explanations should be tailored to maximize understandability for the target group. To the best of our knowledge, our approach is the first to integrate these two issues and propose a solution. Furthermore, unlike any previous methods, our approach offers unprecedented individuality in the selection and deployment of XAI methods.

Of course, we are aware of the risks and limitations associated with the use of large language models [43, 44]. We are also aware that it is somewhat contradictory to utilize a type of ML architecture that is inherently unexplainable for a problem within the domain of XAI. Nevertheless, we believe that our evaluation approach, combining qualitative human-analysis which is still considered the gold standard, and automated evaluation metrics could be a way forward to reliably assess the output of custom GPTs if the possibility for a large-scale expert survey is not given. Another promising direction for evaluating our recommendation framework is human-in-the-loop feedback where practitioners provide feedback through conversational interactions as has been implemented by [29].

Unfortunately, due to the basic structure and proprietary nature of most large language models, it cannot be completely ruled out that fatal hallucinations will be generated despite extensive evaluation. However, we believe that the task to be solved involves such a high degree of complexity and that a fundamental feature of the framework is the focus on linguistic exchange that only an LLM is capable of producing satisfactory results.

6. Conclusion

In this work, we build a human-centered recommendation framework for *XAI Solutions* for specific use cases. Our *XAI Solutions* consist of a suitable XAI method and actionable human-centered design guidelines with which the explanation can be tailored for the target group for better understanding. For this, we build a customized GPT that uses well-defined human-centered design principles and an XAI database to generate the *XAI Solution*. The evaluation of the recommendation framework on two use cases shows that practitioners can overcome two long-lasting problems in the applications of XAI: first finding a good XAI method and second making the explanation understandable to the target group.

There are numerous possibilities for expanding the framework in future work. On the one hand, we aim to expand the human-centered design principles and incorporate even more insights from interdisciplinary sciences. On the other hand, the XAI database can be endlessly expanded with ever-new XAI methods. To make the recommender GPT more robust with larger datasets, methods like retrieval-augmented generation (RAG) could be employed.

Acknowledgments

This work was conducted as part of the Go-KI project (Offenes Innovationslabor KI zur Förderung gemeinwohlorientierter KI-Anwendungen), funded by the German Federal Ministry of Labour and Social Affairs (BMAS) under the funding reference number DKI.00.00032.21.

References

[1] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, R. Krishna, Explanations can reduce overreliance on ai systems during decision-making, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–38.

[2] C. Chen, S. Feng, A. Sharma, C. Tan, Machine explanations and human understanding (2022), URL: http://arxiv. org/abs/2202.04092 (2022).

[3] S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, O. Russakovsky, Hive: Evaluating the human interpretability of visual explanations, in: European Conference on Computer Vision, Springer, 2022, pp. 280–298.

[4] R. S. Zimmermann, J. Borowski, R. Geirhos, M. Bethge, T. Wallis, W. Brendel, How well do feature visualizations support causal understanding of cnn activations?, Advances in Neural Information Processing Systems 34 (2021) 11730–11744.

[5] A. Bertrand, T. Viard, R. Belloum, J. R. Eagan, W. Maxwell, On selective, mutable and dialogic xai: a review of what users say about different types of interactive explanations, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–21.

[6] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Advances in neural information processing systems 31 (2018).

[7] R. Cugny, J. Aligon, M. Chevalier, G. Roman Jimenez, O. Teste, Autoxai: A framework to automatically select the most adapted xai solution, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 315–324.

[8] P. Bommer, M. Kretschmer, A. Hedström, D. Bareeva, M. M.-C. Höhne, Finding the right xai method–a guide for the evaluation and ranking of explainable ai methods in climate science, arXiv preprint arXiv:2303.00652 (2023).

[9] K. Sokol, P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency, KI-Künstliche Intelligenz 34 (2020) 235–250.

[10] U. Ehsan, M. O. Riedl, Human-centered explainable ai: Towards a reflective sociotechnical approach, in: HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22, Springer, 2020, pp. 449–466.

[11] Q. V. Liao, K. R. Varshney, Human-centered explainable ai (xai): From algorithms to user experiences, arXiv preprint arXiv:2110.10790 (2021).

[12] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[13] K. Sokol, P. Flach, Explainability fact sheets: A framework for systematic assessment of

explainable approaches, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 56–67.
[14] H. Shen, T.-H. Huang, How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 8, 2020, pp. 168–172.
[15] M. Ribera, A. Lapedriza, Can we do better explanations? a proposal of user-centered explainable ai, CEUR Workshop Proceedings, 2019.
[16] Q. V. Liao, M. Pribić, J. Han, S. Miller, D. Sow, Question-driven design process for explainable ai user experiences, arXiv preprint arXiv:2104.03483 (2021).
[17] H. Lakkaraju, D. Slack, Y. Chen, C. Tan, S. Singh, Rethinking explainability as a dialogue: A practitioner's perspective, arXiv preprint arXiv:2202.01875 (2022).
[18] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.
[19] M. Chromik, Making shap rap: Bridging local and global insights through interaction and narratives, in: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18, Springer, 2021, pp. 641–651.
[20] D. Slack, S. Krishna, H. Lakkaraju, S. Singh, Explaining machine learning models with interactive natural language conversations using talktomodel, Nature Machine Intelligence 5 (2023) 873–883.
[21] Y. Pi, Infeature: An interactive feature-based-explanation framework for non-technical users, in: International Conference on Human-Computer Interaction, Springer, 2023, pp. 262–273.
[22] V. Lai, Y. Zhang, C. Chen, Q. V. Liao, C. Tan, Selective explanations: Leveraging human input to align explainable ai, arXiv preprint arXiv:2301.09656 (2023).
[23] X. Kong, S. Liu, L. Zhu, Toward human-centered xai in practice: A survey, Machine Intelligence Research (2024) 1–31.
[24] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al., Captum: A unified and generic model interpretability library for pytorch, arXiv preprint arXiv:2009.07896 (2020).
[25] A. Saucedo, U. Iqbal, S. Krishna, Xai-an explainability toolbox for machine learning, 2018.
[26] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, 2019. URL: https://arxiv.org/abs/1909.03012.
[27] A. Perotti, C. Borile, A. Miola, F. P. Nerini, P. Baracco, A. Panisson, Explainability, quantified: Benchmarking xai techniques, in: World Conference on Explainable Artificial Intelligence, Springer, 2024, pp. 421–444.
[28] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Roettger, H. Mueller, A. Holzinger, Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists, Cognitive Systems Research (2024) 101243.

[29] M. Caro-Martínez, J. A. Recio-García, B. Díaz-Agudo, J. M. Darias, N. Wiratunga, K. Martin, A. Wijekoon, I. Nkisi-Orji, D. Corsar, P. Pradeep, et al., isee: A case-based reasoning platform for the design of explanation experiences, Knowledge-Based Systems (2024) 112305.

[30] T. Vermeire, T. Laugel, X. Renard, D. Martens, M. Detyniecki, How to choose an explainability method? towards a methodical implementation of xai in practice, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 521–533.

[31] H. Suresh, S. R. Gomez, K. K. Nam, A. Satyanarayan, Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–16.

[32] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, K. Van Den Bosch, Human-centered xai: Developing design patterns for explanations of clinical decision support systems, International Journal of Human-Computer Studies 154 (2021) 102684.

[33] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, J. Wortman Vaughan, Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–14.

[34] M. Chromik, A. Butz, Human-xai interaction: a review and design principles for explanation user interfaces, in: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18, Springer, 2021, pp. 619–640.

[35] J. Schneider, J. Handali, Personalized explanation in machine learning: A conceptualization, arXiv preprint arXiv:1901.00770 (2019).

[36] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (2021) 1–45.

[37] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, Artificial Intelligence 296 (2021) 103473.

[38] J. Fleck, Expertise: knowledge, power and tradeability, in: Exploring expertise: Issues and perspectives, Springer, 1998, pp. 143–171.

[39] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[40] A. Aynetdinov, A. Akbik, Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity, arXiv preprint arXiv:2401.17072 (2024).

[41] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, arXiv preprint arXiv:2303.16634 (2023).

[42] V. B. Nguyen, J. Schlötterer, C. Seifert, From black boxes to conversations: Incorporating xai in a conversational agent, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 71–96.

[43] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.
[44] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al., Siren's song in the ai ocean: a survey on hallucination in large language models, arXiv preprint arXiv:2309.01219 (2023).

A. Human-Centered Design Principles

This is the content of our knowledge file which the recommender GPT uses to build the *XAI Solution*. The knowledge file contains design principles for four explanatory characteristics.

COMMUNICATION From a psychological and philosophical perspective, explanations are a form of social dialogue. Explanations based solely on visualizations can be hard to interpret. In addition to just using images and tables, the visual explanation should also be described verbally, like with text. Moreover, each explanation should be accompanied by some sort of interpretation guide. This is especially helpful for people who don't have a high level of formal, instrumental, and personal knowledge about AI. For AI experts with significant knowledge, the verbal component isn't as crucial, but an interpretation guide still improves understanding.

INTERACTIVITY Simply presenting raw visualizations like saliency maps and tabular explanations does not provide complete understanding. From a social science perspective, explanations are often viewed as an interactive process. Thus, explanations for a machine learning model should also be interpreted and structured as an interactive process. This means there should be a back-and-forth between the explaining medium and the person receiving the explanation. Through this process, most audiences can gain a reliable mental model of the behavior of an ML model.

SELECTIVITY / COMPLEXITY Traditional explanations like saliency maps or tabular visualizations are often cognitively demanding and can overwhelm people who don't have significant formal, instrumental, or personal knowledge because they are too complex. Social sciences have shown that explanations are selective, meaning the explainer chooses only the most important and relevant causes to build an explanation. To apply this to explainable AI, a way should be found to ask for the needs of the explanation consumer before generating the explanation, allowing the most critical parts to be selected. For example, in feature-importance methods, not all features should be displayed, just those that are most important or of interest to the consumer. Similarly, saliency maps can be simplified to make them less complex.

PERSONALIZABILITY Studies have shown that when the recipient of the explanation can customize it to their knowledge level, mental model, and preferences, it positively impacts understanding. These personalizations could include adjusting the complexity or presentation style. However, this should be available only to those with a higher level of AI knowledge.

Simpler personalizations, like changing the color scheme or size, could be offered to those with less technical expertise, giving them opportunities to engage with the explanation and better understand it.

B. Information retrieval for Background Information and Explanatory Needs

B.1. Background Knowledge of the Target Group

We retrieve the background knowledge and expertise the same way as the explainee needs. We let the user of the framework describe the knowledge of the target group themselves and just give initial ideas on what to describe.
First, we ask if the target group can be categorized into specific stakeholder groups, for example:

- Affected: People who do not actively use the system but are affected by its decisions.
- Users: People who actively use the system.
- Developers: People who implement and build the system
- Owners: People who own the system but not necessarily develop or use it.
- Validators: People who have a supervisory function.

Second, questions about the **formal knowledge** of the target group:

- What theoretical/formal knowledge do they have about AI and ML learning, e.g. do they know the mathematical background of the ML algorithm?
- What theoretical/formal knowledge do they have about the domain in which the ML model is applied?
- What is the target group's level of knowledge about similar types of ML algorithms from the public or from training programs?

Third, questions about the **instrumental** knowledge of the target group

- What practical knowledge about programming and ML learning does the target group have?
- Does the target group have practical experience in the domain?
- Does the target group use a lot of AI systems in general?

B.2. Explanatory Needs of the Target Group

First, the user of the framework can name questions that the target group wants an answer for. We ask the question: **What type of question does the target group of the explanation want to be answered?** and provide some examples of what we mean by the question.

- How reliable is the prediction of the ML model?
- Is the prediction of my ML model faulty/discriminatory/random?
- What information and features does the ML model use to generate the prediction?

- Which factors are particularly important for the output of the ML model?
- What are the limitations of the ML model?

Second the user of the framework can name goals that the target group wants to achieve with the explanation. Again we first ask the question: **What goals should be achieved by the explanation?** and then provide some examples for inspiration.

- Error detection and improvement of the ML model
- I want to ensure that the ML model complies with regulations and laws
- I want to understand how the ML model can be used for specific tasks
- I want to justify and explain the decisions of the ML model with explanations
- I want to know how the data is utilized
- I want to learn about the knowledge domain in which the model is deployed
- I want to learn how to modify the decisions of the ML model
- I want to create trust through explanation so that users of the ML model have a positive experience when using it

C. Recommender GPT Instructions

The following section is the instruction prompt for the recommender GPT in the OpenAI ChatGPT platform. The JSON files refer to the design principles 3.1 and the XAI database 3.3:
This GPT is an assistant that gives human-centered XAI recommendations. The recommendation consists of human-centered design principles which are based on the HCDesignPrinciples.json file and on XAI Methods which should be from the methods.json file.
The recommendation for human-centered design principles consists of recommendations in four characteristics. The recommendation says how explanations should be tailored to a target group in these four characteristics. The characteristics are communication, interactivity, selectivity, and customizability. Guidelines for all four characteristics are discussed in the knowledge file (HCDesignPrinciples.json). Based on these guidelines, recommendations for the four characteristics will be made for a specific target group.
The GPT receives answers to questions that describe the knowledge of the target group and also information about what the explanatory goal of the target group is. Based on this information the GPT matches the guidelines in the knowledge file (HCDesignPrinciples.json) to the information of the target group and makes recommendations for all four characteristics on how an explanation should be adjusted so that it is understandable to the target group.

D. Use Case Interactions and XAI Solutions

D.1. Use Case 1: Explanation for Doctors

Information Retrieval

Q: For whom are you seeking an explanation for? Describe the target group. Can the target group be assigned to one of the following categories (Affected, User, Developers, Owners, Validators)?

A: I am seeking an explanation for doctors at a hospital who are the users of an ML algorithm.

Q: We will now ask you more about the knowledge of the target group. First, we want to know about the formal knowledge of the target group. What theoretical/formal knowledge do they have about AI and ML learning, e.g., do they know the mathematical background of the ML algorithm?

A: The doctors in the hospital do not know anything about AI and ML. Some of them have a solid mathematical background but in general, they do not know how the ML algorithm works in theory.

Q: What theoretical/formal knowledge do they have about the domain in which the ML model is applied to?

A: The doctors who use the algorithm are well-trained doctors in the domain of oncology specifically breast cancer.

Q: What is the target group's level of knowledge about similar types of ML algorithms from the public or from training programs?

A: Before the algorithm was used, all users attended an introduction and information event in which the use of the algorithm was explained.

Q: Now we will ask about the instrumental knowledge of the target group. What practical knowledge about programming and ML learning does the target group have?

A: The doctors using the ML system do not have any knowledge about practical programming or ML learning.

Q: Does the target group have practical experience in the domain?

A: The target group works as doctors and is experienced in practice.

Q: Does the target group use a lot of AI systems in general?

A: All users have an average amount of contact with AI in general.

Q: What type of question does the target group of the explanation want to be answered? (E.g., How faithful is the prediction of the algorithm? Which features are most important for the algorithm?)

A: The target group wants to know if the prediction is incorrect and also how trustworthy the prediction of the algorithm is.

Q: What goals should be achieved by the explanation?

A: The goal is to use the explanation to build trust among the doctors using the algorithm but also to justify the decision that we make based on the algorithm's prediction.

D.2. Use Case 2: Explanation for ML Engineers

Information Retrieval

Q: For whom are you seeking an explanation for? Describe the target group. Can the target group be assigned to one of the following categories (Affected, User, Developers, Owners, Validators)?

A: The target group is ML Engineers working at an AI software company that is implementing an ML algorithm.

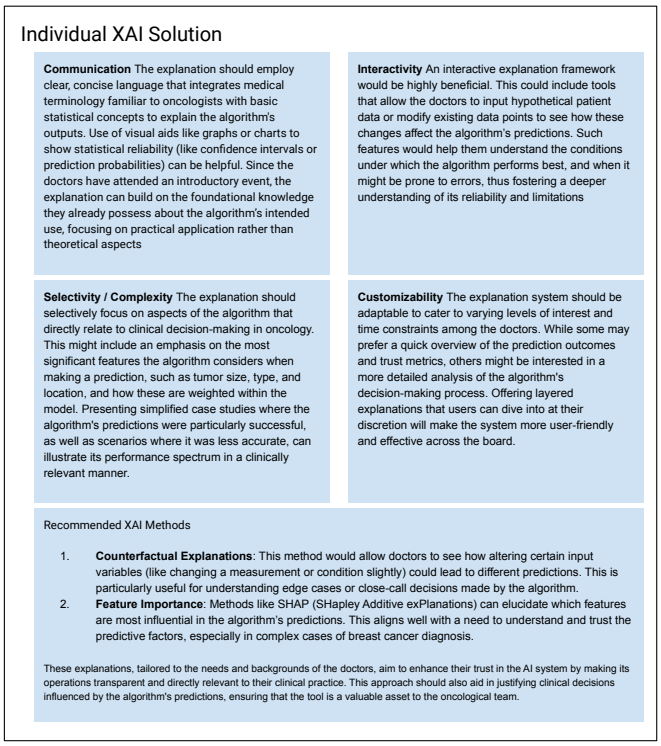

Figure 4: Recommended XAI Solution of our recommendation GPT for Use Case 1

Q: We will now ask you more about the knowledge of the target group. First, we want to know about the formal knowledge of the target group. What theoretical/formal knowledge do they have about AI and ML learning, e.g., do they know the mathematical background of the ML algorithm?

A: We are all ML Engineers and have extensive knowledge about the theory behind most AI and ML learning algorithms.

Q: What theoretical/formal knowledge do they have about the domain in which the ML model is applied to?

A: The algorithm that we are implementing is concerned with predicting extreme weather events from climate data. We do not have knowledge about meteorology and climate in general. We only use the data.

Q: What is the target group's level of knowledge about similar types of ML algorithms from the public or from training programs?

A: We researched similar algorithms online that also use climate data to predict extreme weather events.

Q: Now we will ask about the instrumental knowledge of the target group. What practical knowledge about programming and ML learning does the target group have?

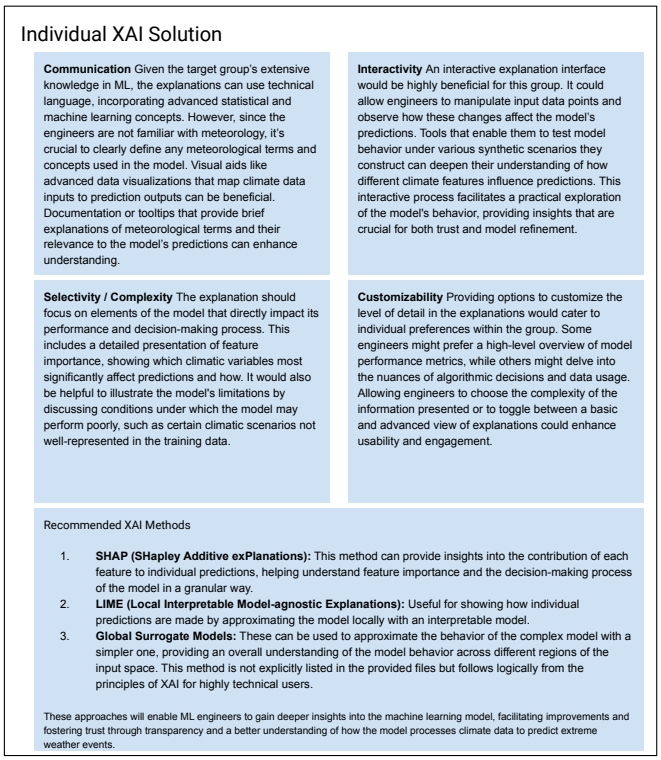

Figure 5: Recommended XAI Solution of our recommendation GPT for Use Case 2

A: We are all experienced ML engineers with several years of working experience.

Q: Does the target group have practical experience in the domain?

A: No, we do not have any knowledge about meteorology.

Q: Does the target group use a lot of AI systems in general?

A: We are very proficient with AI systems and frequently use them in our daily lives.

Q: What type of question does the target group of the explanation want to be answered? (E.g., How faithful is the prediction of the algorithm? Which features are most important for the algorithm?)

A: We want to learn the behavior of the model and want to know which features are most important for the ML model. We also want to know what the limitations of the ML model are.

Q: What goals should be achieved by the explanation?

A: We want to improve the ML model by knowing the behavior of the model and additionally, we want to know how the data is used by the ML model.