What does it mean to explain?
A user-centered study on AI explainability

Lingxue YANG [1], Hongrun WANG [1] and Léa A. DELERIS [1]

[1] BNP Paribas, France
name.lastname@bnpparibas.com

Abstract. One frequent concern associated with the development of AI models is their perceived lack of transparency. Consequently, the AI academic community has been active in exploring mathematical approaches that can increase the explainability of models. However, ensuring explainability thoroughly in the real world remains an open question. Indeed, besides data scientists, a variety of users is involved in the model lifecycle with varying motivations and backgrounds. In this paper, we sought to better characterize these explanations needs. Specifically, we conducted a user research study within a large institution that routinely develops and deploys AI model. Our analysis led to the identification of five explanation focuses and three standard user profiles that together enable to better describe what explainability means in real life. We also propose a mapping between explanation focuses and a set of existing explainability approaches as a way to link the user view and AI-born techniques.

Keywords: Explainable AI, Explainability, User research, User centered design.

1 Introduction

1.1 Background

Support from Artificial Intelligence (AI) and in particular machine learning algorithm has become pervasive in our personal and professional lives. However, one frequent concern is the perceived lack of transparency of such models, in particular when they are used in a context that can have a material influence on our lives, such as health, recruitment, legal or banking settings. It has thus become essential to further develop the capability to explain such advanced models.

In fact, the AI academic community has been active in exploring mathematical approaches that can increase the explainability of models (e.g., LIME [1], Shapley value [2], counterfactual explanations [3]). However, such efforts have been predominantly based on computer scientists' perceptions of what constitutes an explanation. This may produce a gap between explainability techniques and what it means to explain to real users [4] [5].

In this study, our objective is not to further develop technical solutions but rather (i) to take a user-centered perspective in defining what it means to explain AI models

and (ii) to study the fit and alignment of existing explainability methods in practice, i.e., with real users. To further anchor our analysis in real-world setting, we have focused specifically on AI models used within a financial institution though a large part of our analysis is not sector specific.

We start by reviewing explainability techniques. We observe that the current methods are suited for a specific population i.e., data scientists and that the needs of other stakeholders have not necessarily been considered. The main contribution of the paper stems from the insights derived from our user study focused on the variety of user needs for understanding and explaining AI based on the diversified view of stakeholders within a large financial institution. We organized the study in two phases: individual user interviews and group workshops. In the end, our analysis led us to define (i) three user profiles who need explanations with different motivations in different contexts (ii) five "Explanation focuses", which corresponds to specific explanation concerns of users and (iii) a mapping between explanation focuses and explainability methods.

2 Related work

In this section, we review some key concepts of explainable artificial intelligence (definition and needs) and summarize what has been done from computer science and human computer interface (HCI) academic communities.

2.1 XAI definition

Defense Advanced Research Projects Agency (DARPA) initiated explainable artificial intelligence (XAI) program in 2017 [6] to address a machine's inability to explain its thoughts and actions to human users. They then introduced XAI with the goal of enabling users to understand, trust, and effectively manage this emerging generation of artificial intelligence models. Ever since, there has been a surge of interest in the research on XAI both in the Artificial Intelligence (AI) and HCI communities. As a consequence, multiple related words have emerged seeking to provide an additional level of specification about the underlying needs and concepts, in particular interpretability, explainability and intelligibility [5], [7]–[12]. For example, Doshi and Kim [8] define interpretability as "the ability to explain or to present in understandable terms to a human". Separately, Gilpin et al. believe that interpretability is not sufficient for human to understand black-box models. They propose to define explainability as "models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions." [13], which goes beyond interpretability. Liao et al. consider explainability to be everything that makes machine learning (ML) models transparent and understandable to humans [5]. Recently, Arrieta et al [12] clarified some related concepts including understandability, comprehensibility, interpretability, explainability and transparency. They propose that transparency, interpretability and comprehensibility be merged together into

understandability (or intelligibility), which measures the degree to which a human can understand a decision made by a model.

From these definitions, we observe that there is not yet clear consensus on their individual definition and they are sometimes used interchangeably [14]–[16]. However, these definitions share implicitly one common factors: the importance of considering the recipient of the explanation, the one who asks questions about the AI models and receives the explanations. The purpose of this paper being focused on real-world feedback rather than concepts, we do not seek to contribute directly to the semantic discussion of the nuances among all those terms. In the remainder of the paper, to avoid ambiguity, we chose to use explainability when we talk about the explanations that users need.

2.2 XAI techniques

The computer science community has been actively exploring approaches to improve the explainability of algorithms as it can constitute a clear barrier – among others – to the broad deployment of artificial intelligence solutions to end users. In that spirit, methods have been developed to enable people to derive insights into the functioning of an AI solution.

Several taxonomies have been proposed for users to understand the diverse forms of explanations that are available and the questions that each can address [12], [14], [17]–[19]. First, an explanation can either be static or interactive depending on whether the response can be changed according to the feedback from the user. One example of interactive explanation can be a dialog, for instance through a Chabot. Nonetheless, the vast majority of the literature focuses on static explanations [14]. We can also define explanations according to how they are generated [8]. Explanations can be an intrinsic part of the model, in the sense that there is no need for an additional model to generate the explanation. Such models are deemed by nature transparent and easy to interpret for most of the users. The most common examples are short decision trees or sparse linear models [20]. They are often referred to as white box AI models or transparent models by opposition to Black box models such as random forest and deep neural networks. Recently, computer scientists have also worked on self-explaining neural networks, which consists on modifying the architecture of a network to make it interpretable (Explainability by design) [9], [10]. By contrast to intrinsic explanations, a post-hoc explanation is based on applying an additional AI method on the initial model (the decisions are already made) [1], [2], [21]–[25]. Typically, post-hoc explanation approaches can be used on all kinds of AI models, i.e., they are model agnostic, which gives them some sort of universality. Finally, explanations can be either global or local. Global explanations seek to describe the behavior of the entire model [2], [21]–[23], while local explanations provide explanations for single prediction[1], [2], [24], [25].

The majority of explainability methods work either by analyzing the contribution of input features to the model outputs, we call them feature based explanations or by analyzing the instances that were introduced in the model we call them example based explanations[26]–[28]. An important consideration at this point is that the methods

mentioned have predominantly come from the computer science community and are therefore biased towards computer scientists needs for explanation. Existing solutions for XAI are mostly developed in the format of a python package, usually using one or several explainability techniques (Eli5[1], AIX360[2], xai[3], ethik[4]). They thus tend to be better suited for explanation to AI experts or for debugging purposes for technical practitioners (e.g., data scientists). A concrete example could be the self-explaining neural network, while data scientists can read into the additional explainer layer of the neural network to learn the contributions of input features, this type of information are not interpretable for other users [9].

In this paper, our ambition is not to develop new explainability techniques but rather get insights of users' need on this topic especially those who were less served by the current solutions. However, we sought to map the existing realm of such techniques with our findings, so as to better understand what aspects of user-needs for explanations are well-covered and which ones may require further investigation.

2.3 XAI HCI approaches

Given that the way of representing explanations of AI models depends on the recipients as well as their needs, it is essential to work on XAI from user-centered approach. In this section, we reviewed the main contributions to the concept of explainable AI in the HCI community around (i) developing a better understanding of the need for explainability, (ii) enabling visual analytics for XAI to facilitate the users to understand the overall model behavior and input feature behavior and (iii) studying theoretical constructs for XAI from social science.

Regarding the needs for explainability. Many studies have explored what prompts the need for explainability [13], [29], [30], for instance *improvement and optimization of the system*, identification of the potential *bias* and ensuring that the performance (accuracy, precision, robustness and stability) of a *model* is adequate. Some studies have mentioned that users may seek explanations for *verification* purpose when there is a deviation or an inconsistency between what is expected and what has occurred [31], [32]. Miller [9] and Samket et al. [29] both propose that people want explanation to facilitate learning by identifying the hidden laws of nature and "extracting the distilled knowledge" in order to predict and control future phenomena. Some also need explanations to improve the efficiency of *their decisions making process* when using an AI product. *Compliance* with regulation is also a key reason for seeking to improve the explainability of AI systems. Indeed, the European Union set a "The right to explain" regulation whereby individuals subject to a decision made by an AI system should be provided with an explanation of why a specific decision has been made. Further on, end-users may also wish to understand what could be changed to obtain a

[1] https://eli5.readthedocs.io/en/latest/
[2] https://aix360.mybluemix.net/
[3] https://ethicalml.github.io/xai/index.html
[4] https://xai-aniti.github.io/ethik/

better result [27]. Overall, this research stream highlights that the need of an explanation can be motivated by a variety of reasons, from a diverse set of users [33].

Visual analytics for XAI. In parallel, there has been burgeoning efforts around developing explainable user interfaces, specifically with interactive visual analytics capability to support model explanation, interpretation, debugging and improvement [33]–[36]**.** Some tools are specifically designed for technical profiles to inspect and analyze the model they build or use. For example, *Prospector* [37], *Gamut* [38] and *What-If Tool* [39] are similar tools designed to help data scientists understand the impact of input features on the model predictions, investigate the outcome of the model globally and locally. Other solutions like *Shapash or Waston Openscale* have given more consideration to non-technical users. *Shapash* [40] provides an visual support with both global and local explanations for data scientist but only local explanations for end users. *Waston Openscale* [5] only provides information to explain the model output of a given instance.

Theoretical constructs of XAI from social science. A few studies have also looked at the challenge of explaining AI models from a social science perspective, leveraging knowledge into how human usually reason and explain from philosophy and psychology. For instance, Wang et al. [4] proposed a user centric XAI framework that links concepts in human reasoning process with explainable AI techniques. We leveraged their work when we developed our interview guide. Liao et al [5] developed an algorithm-informed XAI question bank where user needs for explainability are represented in terms of prototypical questions. They first identified a list of XAI methods and then mapped them to the questions that users might ask about AI models or AI products. They designed the questions leveraging the taxonomy of Lim&Dey and by interviewing 20 UX and design practitioners who work in the AI domain, they refined and enriched the user questions representing explainability needs.

Our work shares the similar goal of understanding from people their explainability needs. However, Liao et al's [5] work result of XAI question bank is only based on the point of view of design practitioners and their perception of others. In our work, we sought to get direct reactions from a variety of stakeholders, technical and non-technical, involved the AI project lifecycle. Our analysis thus contributes practical knowledge about what an explain means to real users of AI.

3 User study

Our user study aims at understanding from a variety of real-world users involved in the development, deployment and use of AI techniques, what their needs and expectations are in terms of explaining AI solutions. The context of this specific study is a large financial institution where AI models are routinely being deployed for a variety of use

[5] https://www.ibm.com/cloud/watson-openscale

cases. We will call it “test institution” in the remainder of the paper. We conducted the user study in two phases: (i) individual user interviews where we identified the standard user profiles and common themes related to users concerns about the explanations and (ii) group workshops where we refined those findings.

3.1 Individual user interview

The user interviews aimed at discovering: (i) users’ knowledge of AI and level of trust in AI models (ii) users’ needs with respect to understanding and explaining AI models, along with their current approach for doing so (iii) users’ pain points during the process.

Participants. We recruited participants (on a voluntary basis) seeking to obtain a variety of profiles from different business lines and functions. Specifically, we listed the user profiles that we would like to interview and dispatched broadly recruitment emails throughout the test institution asking for volunteers to share their experience around the XAI topic. In the end, our participant pool was composed of 33 persons with different responsibilities and skills: data scientists, model developers, inspectors, risk analysts, business analysts, project managers, relationship managers among others. We note however that we had a limited number of end users (relationship managers, clients) that have direct usage of model predictions or are affected by the model predictions.

Questionnaire guide. We designed an interview guide to ensure that we cover the same topics consistently across all the interviews. We organized it around four main categories:

- Professional activities: Background questions about current professional activity.
- Knowledge of AI models: Questions around participants’ knowledge of AI models and their interaction with AI models (modeling, model management or use of model predictions).
- The need of understanding the AI model: Questions leading participants to describe situations where they usually need to understand AI models (entire AI model behavior or model predictions), how they proceed and what their main concerns are.
- The need of explaining the AI models to others: Questions leading participants to describe situations where they usually need to explain AI models to another person (entire AI model behavior or model predictions) and what the questions from those third parties are.

Procedure. Before the formal interview, we conducted several pilot tests to verify the appropriateness of the questions and clarify aspects that were ambiguous. Each individual user interview lasted approximately 45 minutes. We conducted them via skype meeting due to the constraints from the sanitary situation. During each session, at least one UX designer animated the session, and one data scientist or business analyst observed. The moderator started by giving a brief introduction of the project and asked the participants for consent to record the interview for the analysis purposes. Then, the

moderator let the participant talk about their background experience. After that introductory discussion, the moderator asked the questions following the interview guide. The moderator encouraged the participants to provide as many details as possible. Participants could also use screen sharing or send screenshots to illustrate their answers whenever relevant.

Analysis. In addition to our notes, we transcribed all the recordings manually so as to ensure that we did not discard any useful information. This generated a large amount of qualitative data for analysis (around 1400 minutes of interviews were recorded and transcribed). For the analysis procedure, we decided to rely on thematic analysis. Thematic analysis is a systematic method of breaking down and organizing rich data from qualitative research by tagging individual observations and quotations with appropriate codes, to facilitate the discovery of significant theme [40]. A theme is a description of a belief, practice, need or another phenomenon that is discovered from data. Overall, our analysis process followed mainly five steps:

- *Transcription*: write down the participants' verbalization. This was done by hand which was tedious but enabled at the same time to help us increase our familiarity with the data.
- *Coding*: assign preliminary codes to the transcripts based on its content. A code could be a word or a phrase that acts as a label or a segment of text.
- *Categorization*: group the codes with similar meanings.
- *Search*: look for patterns or themes in the code groups that represented the information related to the explanation.
- *Review*: return to the transcription and compare the themes to make sure that our themes are useful and accurate.

Specifically, we printed out all the transcripts and posted them on a white board based on the categories that we had defined a priori: understanding and explaining. Then we highlighted with different colors the segments of texts associated with different themes.

In the end, the in-depth analysis of the transcripts from those interviews led to the identification of eight themes, which we call explanation focus, and which correspond to specific explanation concerns of users. Those eight explanation focuses were: justification of a local result, model performance, feature importance, feature analysis, model mechanism, business expertise, business impact simulation and recommendation.

Another important output of this first phase of our user study was the articulation of the diverse needs throughout the whole AI project lifecycle. We defined three phases: modeling, decision-making (model management) and the use of model, with three associated user profile namely model developer, model owner (the person that requested the development of the model and that will take responsibility for it) and finally the model users.

The purpose of the interview was to better understand and characterize user needs for explanations and preference in terms of information provided for the sake of explanation.

3.2 Group workshop

To fine-tune our findings, we conducted several group workshops with members from one specific data science team within the test institution. The goal of those workshops was to provide feedback on the eight explanation focuses and also on the three standard user profiles. To anchor the discussions we considered specific use cases.

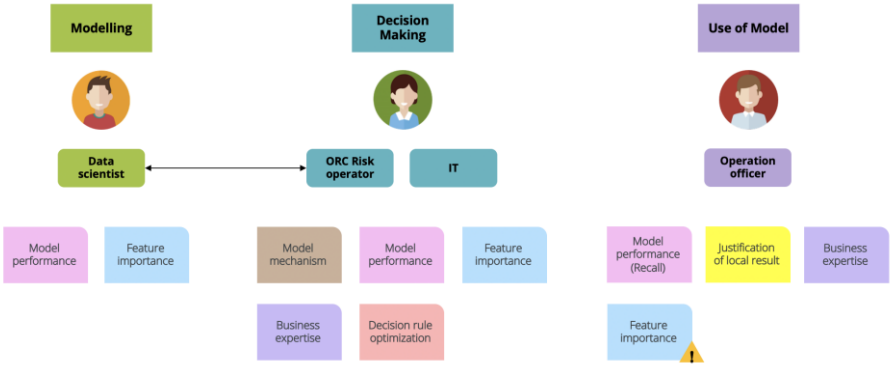

Fig. 1. Fraud detection use case example. Note: participants put *feature importance* in the use of model phase, which means the operational officer may be interested in knowing the feature contributions in predicting a fraud. However, there is a concern about the level of details of features exposed to the end user due to the risk associated with the ability to reverse-engineer such a system.

Participants. We invited all members of the data science team to participate in brainstorming sessions. In the end, we recruited six data scientists, one data engineer and one business analyst. We organized four workshops, three with data scientists, and one with the data analytics engineer manager and business analyst manager as they both served similar project management function in the development of AI projects.

Procedure. For each session, we proceeded as follows: A UX designer served as moderator and was supported by one technical profile for technical questions and a second person to take notes. First, we briefly introduced the eight explanation focuses, providing definitions. Second, we asked participants to comment on each definition and help refine them. Once we had a shared understanding of the explanation focuses, we discussed the four standard use cases one by one. We asked them to brainstorm on the stakeholders involved in each of the three phases of the development process and identify which explanation focuses were relevant based on their professional

experience. An example of output from those discussions based on a fraud-detection use case is presented in **Fig. 1.**

4 Discussion

4.1 Explanation focus

An explanation focus constitutes an "atomic" piece of information that contributes to a specific need for model explanation. After discussion in the group workshops and further discussions among the research team, we decided that the initial eight explanation focuses could be reduced to five.

Table 1. Definition of Explanation focus

Explanation focus	**Definition**
Model mechanism How does the algorithm work? What kind of algorithm was used?	Description of logic of the chosen class of algorithms.
Model performance Is the model performance good enough (precise, accurate, robust, and reliable)?	Information about model performance.
Main contributors What are the main contributors to the model predictions globally and locally?	Information about the contribution of the set of input features to the model predictions globally and locally.
Input feature behavior What is the relationship between each input feature and the model predictions?	Information about (i) the nature of the relationship between each input feature and the model predictions (Linear, monotonous or more complex) and (ii) the effect of changing the value of a given input feature on the prediction.
Example-based justification Why this prediction given this instance? How should this instance change to get a different prediction?	Information justifying the specific prediction for a given instance including both why and why not another output.

Specifically, considering the type of information that each explanation focus provides, two of them could be merged into one broader definition. Specifically, the *recommendation* explanation focus, which involved explaining what needs to change to obtain a different result from the model was merged with the *justification of a local result*. We also modify the label of that broader group to *example-based justification*, which focuses on providing information justifying the specific prediction for a given instance including both why and why not another output. Regarding the *business*

expertise explanation focus, what we meant initially was to verify if the overall relationship between a given input feature and model predictions made sense from a business perspective. This is more related to the user's goal instead of the focus of the explanation. We merged it with *sensitivity analysis* that we renamed to *input feature behavio*r with larger definition on the relationship between each input feature and the model predictions. Regarding the *business impact simulation* explanation focus, we observed from the user interviews that some users needed to be able to determine the optimal parameters of the model (i.e. probability thresholds). We recognized that this was also about the user's goal to ensure the capability of the model predictions in real world use. Hence we merged it with *model performance*.

In the end, we propose the five explanation focuses with their definitions and related explanation needs in **Table 1**.

4.2 User profiles

From the study, we also identified three standard user profiles who are the target recipients of explanations throughout the AI project lifecycle. This finding is also consistent with Ribera & Laperdriza's work [41]. They categorized the *explainees* in three main groups based on their background, goals and relationship with the AI product: developers & AI researchers, domain experts whose expertise is used for the system to make decisions, and lay users who are final recipient of the model decisions. We added the notion of the AI project lifecycle into the explanation needs because we think it is significant to better understand the context in which the users need different explanations. **Table 2** articulates the motivations that we identified behind each explanation focus for each user profile. Note that we have two technical users associated with model development: the model developer and the model validator. They have similar motivation though the model developer typically has the ability to directly modify the model while the model validator can only make recommendations.

Model developers / validators. They are both technical profiles, typically data scientists. They need to understand what drives the model to make sure the behavior of the model is coherent with the use case and the data. They need information to check potential problems such as bias, generalization, information leakage, stability. The model developers also need to explain their findings to two kinds of people: (i) the model validator and (ii) the model owner. To the first person the explanation is geared towards explaining the inner workings of the model so that they are assured that the model development followed the appropriate discipline. To the model owner, the content of the explanation is intended to make sure that the model answers the business needs, and that the performance and limitations of the model are well understood. The model validators need to be satisfied that the model works as intended and that it does not display problematic behaviors. They possibly need to report those findings to the model owner.

Table 2. Explanation focuses to user profiles taking into account their different needs

Name	**Model developer/ validator**	**Model owner**	**Model user**
Model mechanism	Ensure that model (logic) is coherent with the use case and the data.	Ensure that model (logic) is coherent with the use case and the data. Curiosity: Learn and apprehend a new concept.	Curiosity: Learn and apprehend a new concept.
Model performance	Explain the strengths and limitations of the model (error, false positive, bias). Ensure that model (performance) is coherent with the use case and the data. Investigate the sensitivity and stability of the model performance to modeling changes.	Ensure that model (performance) is coherent with the use case and the data. Ensure that the model robustness and stability is sufficient for the use case. Provide support to choose among multiple models or to choose optimal parameters of the model (e.g. Thresholds).	Be convinced that the model (performance) is coherent with the use case and the data. Curiosity: Understand what performance means for an AI model.
Main contributors	Ensure that the input features of the model are coherent with the use case and the data by checking globally and locally possibly leading to the modeling changes. Investigate model errors (false positives/negatives/no discriminative feature/no bias).	Ensure that the input features of the model are coherent with the use case and the data, that no discriminative feature were used, by checking globally and locally possibly leading to the modeling changes. Develop or confirm insights into the contribution of the input features to the model predictions or for a given instance.	Understand the contribution of the input features for a given instance possibly explaining to a third party. Develop or confirm insights into the contribution of the input features to the model predictions or for a given instance.
Input feature behavior	Ensure that the input features of the model are coherent with the use case and the data (no bias) possibly leading to the modeling changes.	Ensure that the input features of the model are coherent with the use case and the data (no bias) possibly leading to the modeling changes. Develop or confirm insights into influence of input data on the model predictions.	Develop or confirm insights into influence of input data on the model predictions.
Example-based justification	Ensure that the model behavior is coherent with the use case and relevant input data. Investigate model errors (false positives/negatives).	Ensure that the model behavior is coherent on a given instance.	Understand the model behavior for a given instance possibly explaining to a third party.

Model Owners. They need to understand enough the model to determine if it is coherent with the use case and the data, if performance and limitations are taken into account for deployment. Additionally, they may develop or confirm business insights into the model behavior. They need occasionally to explain to the model users how the model makes decision and seek to also instill trust in the model to be deployed. The model owners are the ones taking the responsibility of the model in the end and thus their need for explanation is driven by that responsibility.

Model Users. They need to be convinced that the model is coherent with the use case and the data in order to develop trust in the model. Moreover, they may be curious about the underlying decision rules of the model predictions and they should be able to understand specific prediction for a given instance. They may be required to explain a specific prediction to a third party (customer).

In general, from the user study, we noticed that people asked different questions in understanding and explaining AI model behavior or their predictions, however those questions can be answered by providing the same information. This is consistent with the Lim&Dey's findings that users use different strategies to check model behavior and thus ask different questions for the same explanation goals ([42], [43]). Nevertheless, the way to represent this explainable information varies from one user to another. Therefore, it will be insightful to match the explanation focus with the explainability techniques. Likewise, each explainability technique can be represented by different visualizations (e.g., tables, graphs, interactive interfaces).

4.3 Mapping

In our broader exploration of XAI we have carried out a thorough state-of-the-art study on some of the most popular explainability methods [1], [2], [21]–[26], [44]. We proceeded to map those to the explanation focuses that we have outlined [2], [21]–[23], [26], [44], as a way to connect our findings from the user study with the technical XAI landscape.

To approach the mapping, we organized a group discussion where data scientists and UX designers gathered to exchanges the findings from both sides (UX and AI). The discussion was organized around three main phases. First, the data scientists pointed out the principles of different explainability methods, the meanings of some common performance metrics, the type of information that each explainability technique brings and the suitable cases for each technique from a data scientist point of view. Second, the UX designers shared the five explanation focuses, and gave some details on the information that users required, and the potential needs associated with each explanation focus. Finally, the two teams worked together and tried to match the suitable explainability techniques to the explanation focus by going through one explanation focus at a time. The result of those interactions is summarized on **Fig. 2.** This visual summary aims at being a starting point for this exercise that should be enriched based on further addition of explainability techniques or refinement of existing

ones. At this point, some of the explanation focuses are associated with only one specific explainability method: specifically, the information that corresponds to 'Model performance' is solely given by *Performance Metrics*. By contrast, others are supported by multiple techniques; for example, we can use either Permutated feature importance or SHAP summary plot to give information about what are the main drivers of the model prediction, i.e., *Main contributors*.

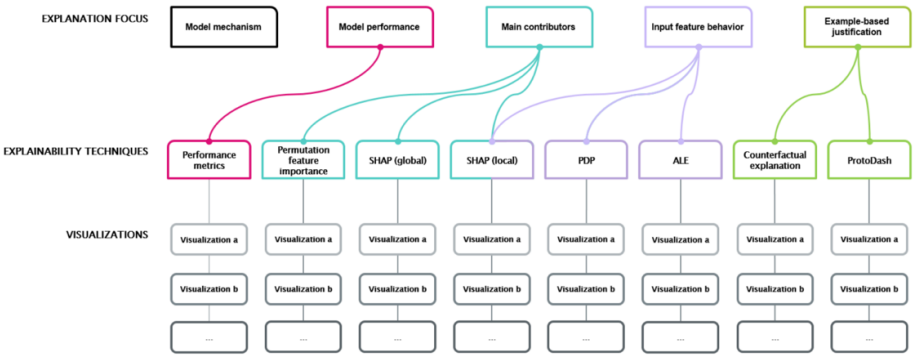

Fig. 2. Mapping between explanation focus and explanation techniques

Overall, the explanation focuses provide a novel level of granularity in the definition of explanation in the context of AI models. The initial mapping provides an overarching framework to articulate the link between the user needs and technical solutions, as well as to identify both technical and UX areas that may be underserved at this point. This relationship can serve as a guide for choosing explainability techniques and visualizations to compose coherent explanation solutions depending on explanation focus of users.

For instance, our study led us to realize that there were limited coordinated efforts dedicated to the explanations of model mechanism, each person resorting to their own inspirations for such tasks. Similarly, for specific explainability techniques, different versions of visualization could be designed for different users and contexts, which need to be defined and evaluated. Finally, from a technical perspective, it will be interesting to see which explainability methods are widely available, or only for a subset of models, i.e. where there are further technical approaches to be defined and in the same perspective, which ones are used most often and in which context.

5 Conclusion

While there had been great progress in the research field of XAI, particularly development of novel XAI techniques, we observed that there was limited work on refining what explainability means to real users. This paper describes a user-centered analysis of what explanation means to different real-world practitioners of AI, when, what and why they need an explanation. We identified five explanation focuses and three standard user profiles within the AI project lifecycle. Based on these findings, we

propose a framework mapping explanation focuses and explainability techniques. We believe this could also guide the design of explainability visual supports and novel explainability techniques and coordinate further efforts in this challenging multidisciplinary topic.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," 2016.

[2] S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions."

[3] B. Kim and U. T. Austin, "Examples are not Enough , Learn to Criticize ! Criticism for Interpretability," no. Nips, 2016.

[4] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing Theory-Driven User-Centric Explainable AI," no. May, 2019, doi: 10.1145/3290607.XXXXXXX.

[5] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," pp. 1–15, 2020, doi: 10.1145/3313831.3376590.

[6] D. Gunning, "Explainable Artificial Intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*, 2017. .

[7] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," *ACM Int. Conf. Proceeding Ser.*, pp. 195–204, 2009, doi: 10.1145/1620545.1620576.

[8] F. Doshi-Velez and B. Kim, "A Roadmap for a Rigorous Science of Interpretability," no. Ml, pp. 1–13, 2017.

[9] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2018, doi: 10.1016/j.artint.2018.07.007.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *arXiv*, pp. 1–45, 2018.

[11] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, and D. L. McGuinness, "Explanation Ontology: A Model of Explanations for User-Centered AI," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12507 LNCS, no. Ml, pp. 228–243, 2020, doi: 10.1007/978-3-030-62466-8_15.

[12] A. Barredo Arrieta *et al.*, "Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[13] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," *Proc. - 2018 IEEE 5th Int. Conf. Data Sci. Adv. Anal. DSAA 2018*, pp. 80–89, 2019, doi: 10.1109/DSAA.2018.00018.

[14] V. Arya *et al.*, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," 2019, [Online]. Available:

http://arxiv.org/abs/1909.03012.

[15] J. Parekh, P. Mozharovskyi, and F. d'Alche-Buc, "A Framework to Learn with Interpretation," 2020.

[16] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," *FAT* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar.*, pp. 56–67, 2020, doi: 10.1145/3351095.3372870.

[17] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *arXiv*, 2020.

[18] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges," no. 01, pp. 417–431, 2020, doi: 10.1007/978-3-030-65965-3_28.

[19] G. Vilone and L. Longo, "Explainable Artificial Intelligence: a Systematic Review," *arXiv*, no. Dl, 2020.

[20] A. A. Freitas, "Comprehensible classification models," *ACM SIGKDD Explor. Newsl.*, vol. 15, no. 1, pp. 1–10, 2014, doi: 10.1145/2594473.2594475.

[21] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *arXiv*, 2018.

[22] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 82, no. 4, pp. 1059–1086, 2020, doi: 10.1111/rssb.12377.

[23] Q. Zhao and T. Hastie, "CAUSAL INTERPRETATIONS OF BLACK-BOX MODELS QINGYUAN ZHAO AND TREVOR HASTIE Department of Statistics, Stanford University," 2016.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 1527–1535, 2018.

[25] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," *J. Comput. Graph. Stat.*, vol. 24, no. 1, pp. 44–65, 2015, doi: 10.1080/10618600.2014.907095.

[26] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2019-November, pp. 260–269, 2019, doi: 10.1109/ICDM.2019.00036.

[27] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electron. J.*, pp. 1–52, 2017, doi: 10.2139/ssrn.3063289.

[28] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12269 LNCS, no. 01, pp. 448–469, 2020, doi: 10.1007/978-3-030-58112-1_31.

[29] W. Samek, T. Wiegand, and K. R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv*,

2017.

[30] I. Nunes and D. Jannach, “A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems.”

[31] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, “Designing Theory-Driven User-Centric Explainable AI,” no. May, 2019, doi: 10.1145/3290607.XXXXXXX.

[32] D. J. Hilton and B. R. Slugoski, “Knowledge-Based Causal Attribution. The Abnormal Conditions Focus Model,” *Psychol. Rev.*, vol. 93, no. 1, pp. 75–88, 1986, doi: 10.1037/0033-295X.93.1.75.

[33] B. Y. Lim and A. K. Dey, “Investigating intelligibility for uncertain context-aware applications,” in *UbiComp’11 - Proceedings of the 2011 ACM Conference on Ubiquitous Computing*, 2011, pp. 415–424, doi: 10.1145/2030112.2030168.

[34] B. Y. Lim and A. K. Dey, “Evaluating intelligibility usage and usefulness in a context-aware application,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8008 LNCS, no. PART 5, pp. 92–101, doi: 10.1007/978-3-642-39342-6_11.

[35] J. Krause, A. Perer, and K. Ng, “Interacting with predictions: Visual inspection of black-box machine learning models,” *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 5686–5697, 2016, doi: 10.1145/2858036.2858529.

[36] S. Coppers *et al.*, “Intellingo: An intelligible translation environment,” *Conf. Hum. Factors Comput. Syst. - Proc.*, vol. 2018-April, 2018, doi: 10.1145/3173574.3174098.

[37] B. Y. Lim and A. K. Dey, “Toolkit to support intelligibility in context-aware applications,” p. 13, 2010, doi: 10.1145/1864349.1864353.

[38] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, “Bringing Transparency Design into Practice,” pp. 211–223, 2018, doi: 10.1145/3172944.3172961.

[39] C. Kulesza and S. Principles, “Principles of Explanatory Debugging to Personalize Interactive Machine Learning,” 2015, doi: 10.1145/2678025.2701399.

[40] Rosala Maria, “How to Analyze Qualitative Data from UX Research : Thematic Analysis,” *Nielsen Norman Group Publication*, 2019. .

[41] M. Ribera and A. Lapedriza, “Can we do better explanations? A proposal of user-centered explainable AI,” *CEUR Workshop Proc.*, vol. 2327, 2019.

[42] B. Y. Lim and A. K. Dey, “Evaluating Intelligibility Usage and Usefulness in a Context-Aware Application.”

[43] B. Y. Lim and A. K. Dey, “Investigating Intelligibility for Uncertain Context-Aware Applications,” 2011.

[44] A. Dhurandhar *et al.*, “Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives,” *arXiv*, no. NeurIPS, 2018.