Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users

Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, Marcin Detyniecki

Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users

Clara Bove
clara.bove@lip6.fr
clara.bove@axa.com
Sorbonne Université, CNRS, LIP6
F-75005 Paris, France
AXA
Paris, France

Jonathan Aigrain
jonathan.aigrain@axa.com
AXA
Paris, France

Marie-Jeanne Lesot
marie-jeanne.lesot@lip6.fr
Sorbonne Université, CNRS, LIP6
F-75005 Paris, France

Charles Tijus
tijus@lutin-userlab.fr
Laboratoire CHArt-Lutin, University
Paris 08
Paris, France

Marcin Detyniecki
marcin.detyniecki@axa.com
AXA
Paris, France
Polish Academy of Science
Warsaw, Poland

ABSTRACT

The increasing usage of complex Machine Learning models for decision-making has raised interest in explainable artificial intelligence (XAI). In this work, we focus on the effects of providing accessible and useful explanations to non-expert users. More specifically, we propose generic XAI design principles for contextualizing and allowing the exploration of explanations based on local feature importance. To evaluate the effectiveness of these principles for improving users' objective understanding and satisfaction, we conduct a controlled user study with 80 participants using 4 different versions of our XAI system, in the context of an insurance scenario. Our results show that the contextualization principles we propose significantly improve user's satisfaction and is close to have a significant impact on user's objective understanding. They also show that the exploration principles we propose improve user's satisfaction. On the other hand, the interaction of these principles does not appear to bring improvement on both dimensions of users' understanding.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **HCI theory, concepts and models**.

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03…$15.00
https://doi.org/10.1145/3490099.3511139

KEYWORDS

explainable AI, human-centered AI methods, user studies, interface design

ACM Reference Format:
Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3490099.3511139

1 INTRODUCTION

As Machine Learning (ML) models become increasingly accurate and accessible, they are now used in a wide variety of domains to support humans to make important decisions, such as making medical diagnosis [35], assessing recidivism risks for prisoners [38], or filtering applicants for a managing position [33]. Yet, a more widespread adoption of ML models is for now limited by their difficulty to provide explanations about the rationale behind their predictions [5, 36]. Also, unintended behaviors of deployed models, such as biased predictions or the existence of adversarial examples, fuel the call for more interpretability [7, 26, 34]. These limitations led to an increased interest in *eXplainable Artificial Intelligence* (XAI).

In this research field, users are generally classified into categories depending on their expertise level/domain, that can have different needs and goals [25]: (i) AI practitioners, who are (at least) knowledgeable about ML, (ii) agents, who are (at least) knowledgeable about the involved ML application domain, and (iii) non-expert users, who are neither knowledgeable about AI nor the application domain. In this work, we focus on improving the understanding of ML explanations for the latter non-expert users. It has been established [18] that such users are in general more interested in understanding the rationale behind a specific prediction, i.e. local explanations, rather than the overall rationale of a model (i.e global

explanations). Therefore, local feature importance explanations, such as those provided by SHAP [21], appear to be appropriate for such users. However, even though they match their expectations, it has recently been shown that these local feature importance explanations tend to be misleading for non-expert users [7]. They may indeed contribute to the creation of false or incomplete beliefs about the ML model. It has also been observed that the extent to which non-expert users understand these explanations is sensitive to whether they are presented as charts or as textual explanations [32]. Thus, we argue there is still some more work needed to enhance how local explanations are presented to non-expert users.

We consider two such enhancement methods: we study how effective contextualizing and allowing the exploration of local feature importance explanations are to improve the quality of explanation for non-expert users, considering two components for this quality, namely subjective satisfaction and objective understanding. Contextualization is considered an essential part of the explanation process by the social science field [22, 24]. In the XAI community, several works showed the positive impact of contextualizing explanations by adding domain knowledge [17, 27, 29] or by displaying relevant training examples [12, 23, 35]. However, most of these works target expert users, even though non-expert users are likely to also benefit from additional contextual information. Allowing the exploration of the explanations is also an important requirement to help non-expert users build better mental models [7, 16]. It is considered a promising way to enhance local explanations by the XAI community, as we describe in more details in Section 2.

In this paper, we present generic design principles to improve understanding and satisfaction of non-expert users through contextualization and exploration. Regarding contextualization, we propose to enhance local feature importance explanations by adding complementary information at three levels: on the ML system, on the ML application domain and on external factors. Regarding exploration, we propose to provide several explanation displays in order to match different user needs, as well as allowing the interaction with example-based explanations. We also provide an implementation of these principles into a user interface for an insurance scenario. We evaluate the effectiveness of our propositions through a moderated study with 80 participants. The experiment we conduct in a moderated lab setting shows that the contextualization principles we propose significantly improve user's satisfaction and is close to have a significant impact on user's objective understanding. They also show that the exploration principles we propose improve user's satisfaction. On the other hand, the interaction of these principles does not appear to bring improvement on both dimensions of users' understanding.

2 RELATED WORKS AND RESEARCH QUESTIONS

A currently very active domain of XAI investigates how to design explainable interfaces for different types of explanations. In particular, local explanations such as Shapley values [21] are widely used in practice [9]. Yet, it has been shown that they can be misleading or deceptive [3] because they do not justify the influence of a feature's on the prediction [1]; they can also lead to form false or incomplete mental models about how the explained ML prediction model works [7]. In this section we review recent works in XAI that aims at improving the understanding of such local explanations, considering in turn two types of enrichments: first we analyze methods and tools that contextualize the explanations, then we analyze works that allow users to explore the explanations.

2.1 Contextualizing explanations

It has been shown that the context of the prediction and basic domain knowledge may be important for non-expert users to better interpret ML explanations [15, 17]. For instance, Bellotti *et al.* [2] argue that automated systems need to share sufficient information about the context (what the system does and how it does it), so users are able to understand their behavior. However, for now there is no consensus on what it means in practice to provide contextual information for ML explanations. In the literature, the main approach consists in extracting automatically the context from the ML system or the data [2, 12, 19, 23, 27, 27, 29, 35, 37].

Indeed, context can first be directly extracted from the ML system, e.g. through Knowledge graphs (KG), in order to add domain knowledge on top of local explanations. For example, KGs can encode better data representations [29], structure a prediction model in a more interpretable way [27] or adapt semantic similarity for local explanations [17]. Context can also be extracted from the dataset in the form of examples like counterfactuals [21]. For a medical diagnosis tool, it has been shown that some doctors seek to verify the correctness of explanations provided for a medical diagnosis through counterfactual examples and nearest neighbors present in the training set [35]. These propositions have been designed to help AI-experts and domain experts to better understand the ML explanations. However, they can be relevant for non-expert users as well. Martin *et al* [23] show that these users are more likely to ask for concrete examples of output from the training set. Gomez *et al.* [12] propose to contextualize local explanations by adding a visual representation of the dataset. For one instance, the explainable interface graphically represents for each feature where its values lie within the density distribution of the training set. However, the lack of evaluation in this work makes it difficult to deduce whether this proposition is useful to them.

To conclude, contextualization is an interesting approach to improve the intelligibility of ML explanations. However, the current methods to contextualize them have mainly been designed for AI-experts and domain experts [10]. They can still be relevant for helping non-expert users as it has been demonstrated by Gomez *et al.* [12]. Yet, it remains important to compensate their literacy gap [4], in particular for their domain knowledge.

2.2 Exploration in explanations

As compared to contextualization, exploratory methods have been more extensively studied to improve local explanations intelligibility. However, in the context of explainable interfaces, the term "exploration" actually refers to two distinct principles. On one hand, it can refer to allowing users to interact directly with the ML model to observe the impact on the explanations. On the other hand, it also refers to allowing users to navigate between several kinds of explanations. Several works actually propose a combination of these two principles.

Interacting with the ML model. Several works in the literature allow users to easily change the input values of the ML model to observe the impact on the prediction and on the explanations. Krause *et al.* [16] describe in a case study how interactive dependence plots can help data scientists assess the relevant of ML models. *Gamut* [14] is another interactive visual analytic system designed for data scientists to help them understand generalized additive models (GAMs). Cheng *et al.* [6] show that having an interactive interface improves non-users understanding and self-reported understanding, at the cost of an increased time spent on the interface.

Navigating between different explanations. Another aspect of exploration resides in allowing users to navigate between different kinds of explanations. For instance, Collaris *et al.* [8] state that five fraud agents responded positively to being able to navigate between a local feature dashboard (with feature importance, partial dependence plots and distribution in training dataset) and a global rule dashboard (representation of locally extracted decision rules). Wang *et al.* [35] report that medical doctors request to have access to neighbor and counterfactual instances to better interpret the prediction of a diagnosis tool. Gomez *et al.* [12] also combine local feature importance explanations with counterfactual examples. The *Gamut* tool [14] combines local feature importance explanations with data density estimations.

Overall, exploration is considered as an important feature to include in explainable interfaces by the XAI community. This is why we include this notion in the design principles we propose and present in Section 3.

2.3 Research questions

This paper aims at studying the explanations provided in the form of enriched local feature importance for non-expert users, considering the two types of enhancements discussed in the previous sections, namely contextualization and exploration. More precisely, the aim is to examine how effective they are, individually and combined, to improve the explanation quality for users with no expertise, neither in the ML nor in the involved application domains. As discussed in more details in Section 5.1, we consider two components for this explanation quality, distinguishing between objective understanding, which assesses the extent to which users actually understand the explanation, and subjective satisfaction, which assesses the extent to which users appreciate the interface. More precisely, the study is driven by the following research questions and hypotheses:

- **RQ1** : How effective are contextualizing and allowing exploration for improving non-experts users' understanding of Local Feature Importance explanations?
  - H.1.1 : Contextualizing these explanations improves non-expert user understanding
  - H.1.2 : Allowing exploration in these explanations improves non-expert user understanding
  - H.1.3 : Contextualizing and allowing exploration in these explanations improve even more non-expert user understanding
- **RQ2** : How effective are contextualizing and allowing exploration for improving non-experts users' satisfaction of Local Feature Importance explanations?
  - H.2.1 : Contextualizing these explanations improves non-expert user satisfaction
  - H.2.2 : Allowing exploration in these explanations improves non-expert user satisfaction
  - H.2.3 : Contextualizing and allowing exploration in these explanations improve even more non-expert user satisfaction

In order to answer these questions and reject null hypotheses, we develop an interface offering enhanced local feature importance explanations: we investigate new modalities for designing contextualization and exploration for the case of non-expert users, as described in the next sections.

3 DESIGN PRINCIPLES

This section presents the generic XAI principles we propose for contextualizing and allowing exploration in local feature importance explanations: their respective description, purpose and level are discussed in turn in the following sections and summarized in Table 1. We propose an implementation of these principles in an insurance usage scenario described in Section 4.

3.1 Contextualizing explanations adding transparency

We propose XAI principles for contextualizing local feature importance explanations by providing more information at three levels: about the ML System, about the ML application domain and about external factors influencing indirectly the prediction. This additional information makes the ML system more transparent about its purpose.

3.1.1 ML Transparency. As non-expert users can find it difficult to get a global view on the ML system that generates the prediction, we propose a ML transparency principle, that aims at providing guidance about how to interpret the explanations users get for a prediction.

Non-expert users do not know how the model has been trained and which attributes it uses to make a personalized prediction. Moreover, they most likely never interacted before with local feature importance as explanations and do not know how to interpret them. Thus, it is important to be more transparent about the overall ML system so users understand its purpose and basic operations, as argued by Bellotti *et al.* [2]. For example, local explanations can be misunderstood as global ones if the users are not told that the displayed effect is only true for this specific prediction.

We propose that this ML transparency is accessible at the explanation level, meaning users have it prior to interacting with the explanations, so as to better interpret them.

3.1.2 Domain transparency. As previously presented, explanations should provide some brief justification [1]. Thus, we propose to associate local feature importance explanations with additional global information related to the ML application domain. The domain transparency principle provides domain knowledge and aims at helping users understand why a feature is used by the ML system, and how it might impact the prediction, regardless of its effect.

The type of explanations we consider in this research cannot make clear why a feature has a specific influence on the prediction.

Also, non-expert users lack knowledge on the domain of applied ML. Hence, they might not understand why some non-intuitive features are needed in this context for calculating the prediction. Justifying the feature importance with respect to the applied domain is needed by these non-expert users to better understand the prediction.

This domain transparency is generic, i.e. applicable to all instances, and should be paired with each local feature importance explanation to improve understanding of explanations' operations. We also believe this information has to be provided by domain experts, rather than extracting this knowledge from the system [17, 27, 29], or showing examples from the training dataset [12, 35].

3.1.3 External transparency. Local feature importance provides explanations about a given prediction model but external factors can also influence the outcome. We propose an external transparency principle that makes more explicit the impact of such factors. We call external information the type of knowledge which is not domain specific and differs from the information considered in the previous paragraph.

Indeed, some external events can affect the prediction because of real-life context (e.g. external events such as the COVID crisis that indirectly influences the prediction through the dataset) and algorithmic processes (e.g. data that are collected but not used). For instance, some information a user is requested to give can be excluded from the ML model by design (e.g. personal information like name, gender, or phone number can be asked to communicate with users, but it is not used by the ML prediction model), but users may believe it is taken into account for the prediction they get. Thus, it is important to be transparent with the users about which factors impact or not the prediction, even though this information might be external to the model.

This additional external information should be displayed at the explanation level so users have all the elements needed to better interpret the explanations.

3.2 Allowing exploration of explanations with interactive features

We also propose design principles for allowing users to explore the explanations at two levels: setting an interactive display of the explanations at an overall explanation level, and showing example-based explanations at a feature level.

3.2.1 Interactive Display. Local feature importance explanations usually display features in decreasing order of the absolute feature importance values. Users see the major positive influence shown at the top together with the major negative influence. This is a faithful representation of the ML model behavior. To have a more user-centered approach, we propose an interactive display principle to allow users to adapt the display of the feature according to their own needs and goals.

Indeed, users may want to test different hypotheses when interacting with a ML system to help them make a decision regarding the prediction [6, 8, 14]. For non-expert users, it is important to provide exploratory paths as they could be confused about what values can possibly be modified.

Thus, this interactive display should be accessible at the top of the explanations level so users can choose their display preferences to get the explanations in the most relevant way possible for them.

3.2.2 Example-based explanations. Local feature importance explanations reveal feature effects on a given prediction for each attribute. Because the explanations are local, the feature effects are specific to each prediction. We propose to make explicit that the local feature importance explanations are only true for one instance, by showing examples of prediction variations when changing one feature value.

Like counterfactuals [12, 16, 35], this example-based explanations principle should emphasize the impact of potential other values of each feature on the prediction. Indeed, it is not intuitive for non-expert users that the score of one attribute is specific to each instance. They may believe it is the same score for everyone (e.g. for a car insurance pricing service, a specific car model would always has the same impact on the premium) or always the same independently from the value of this attribute (e.g. any car model would have the same impact on the premium). Thus, it is important for non-experts users to clarify that the explanations are only valid for their own instance, so they do not build a wrong mental model of the ML System for future interactions.

The example-based explanations should appear at a feature level as a second layer of information for users to test their hypothesis on the potential effect of other feature values on the prediction.

4 APPLICATION: IMPLEMENTATION OF THE DESIGN PRINCIPLES IN AN INSURANCE SCENARIO

This section presents the application of the XAI principles we propose, as described in Section 3, into an insurance-related interface. We describe the usage scenario in Section 4.1 and the design process for implementing the principles we propose in the user interface in Section 4.2.

4.1 Usage scenario

We apply the principles we propose in a car insurance pricing interface. In this scenario, users provide several pieces of information regarding their insurance settings and background (coverages and options for the vehicle, personal bonus/malus, insurance history), the vehicle to insure (car's details, its usage and parking) as well as personal information (name, age and license information for each driver, address). This information is usually required by insurers to estimate a price according to each individual risk to have accidents and/or damages. The aim of the XAI interface is that prospective clients using this service to calculate a personalized price for a new car insurance can understand how their information impact the price they get.

4.2 XAI Interface

The implementation of the principles we propose, as presented in Section 3, is illustrated in Figures 1 and 2. We describe in the following paragraphs the design of these explanations with the implemented principles.

Type of principle	Principle	Description	Purpose	Level
Contextualization Provide users with missing contextual information	**ML Transparency**	Give transparency on the ML system's scope and basic operations. It should provide guidance regarding how to interpret the explanations.	Understand the ML system and its basic operations for explanation interpretation	At explanation level
	Domain Transparency	Pair each local feature importance explanation with global information provided by a domain expert. It should provide some brief justification about how a feature might impact the prediction regardless of its value.	Compensate for lack in domain knowledge	At feature level
	External Transparency	Complete the explanation with any other relevant information that could justify the prediction. It should provide more transparency on real-life context or the algorithmic process.	Add elements of contextualization which are not directly related to the ML system	At explanation level
Allow exploration Provide users with interactive features to test their hypotheses	**Interactive Display**	Allow users to adapt the display of the explanations according to their needs and goals. It should provide relevant options of display.	Ease access to most relevant explanations according to the users' goal.	At explanation level
	Example-based explanations	Provide an interactive example-based explanation for each feature. It should help users understand the impact on the prediction of different values per feature.	Understand potential feature importance effects on the prediction	At feature level

Table 1: Proposed XAI principles to improve understanding of local feature importance explanations for non-expert users. We describe and define the purpose of each principle we propose for contextualizing and allowing exploration. We also define the level of the ML explanations where the described principle is more valid: "explanation level" refers to principles that apply to the overall ML explanations for one prediction; "feature level" refers to the principles that apply to each feature explanation.

4.2.1 Card-based design. We apply a card-based design for the display of the explanations, as illustrated in Figure 1. Compared to classic local feature importance presentation [21], this design choice allows us to associate more content and interactions with the initial explanations we generate from ML interpretability solutions. Thus, we consider features individually and adapt the length of the card to the amount of content to display. A card contains two parts with different pieces of information related to the feature.

The top part displays the feature importance explanation: it contains the feature's label, its value and its effect on the prediction. We believe it is important for labels to be user-friendly so we propose to name them with non-technical labels. Also, we propose to design visually the effects on the prediction so users can identify quickly what the effect is on the prediction: e.g. for a price prediction, the effect is displayed in green if it decreases the price, in red if it increases it. Finally, we provide a more user-friendly visual representation of the feature with an illustrative icon.

The bottom part of the card is dedicated to contextualize the feature importance explanation with the domain transparency, and to allow exploration with example-based explanations, as discussed in the next subsections.

4.2.2 Implementing contextualization principles.

ML Transparency. We design an onboarding text above the cards (see Figure 1) of local feature importance explanations: it explains how the price for a car insurance is estimated by the ML system and makes explicit which users' personal information is used to give the personalized price. In addition, it provides information about how to read and interact with the feature-associated cards.

Domain Transparency. Each feature-associated card contains two complementary pieces of information: we display local feature importance as the basic explanation at the top of the card, and we pair it with more generic information about how the feature can impact on the price in the context of car insurance on the bottom (see Figure 1). This information is provided by an expert, i.e. an actuary in the considered scenario.

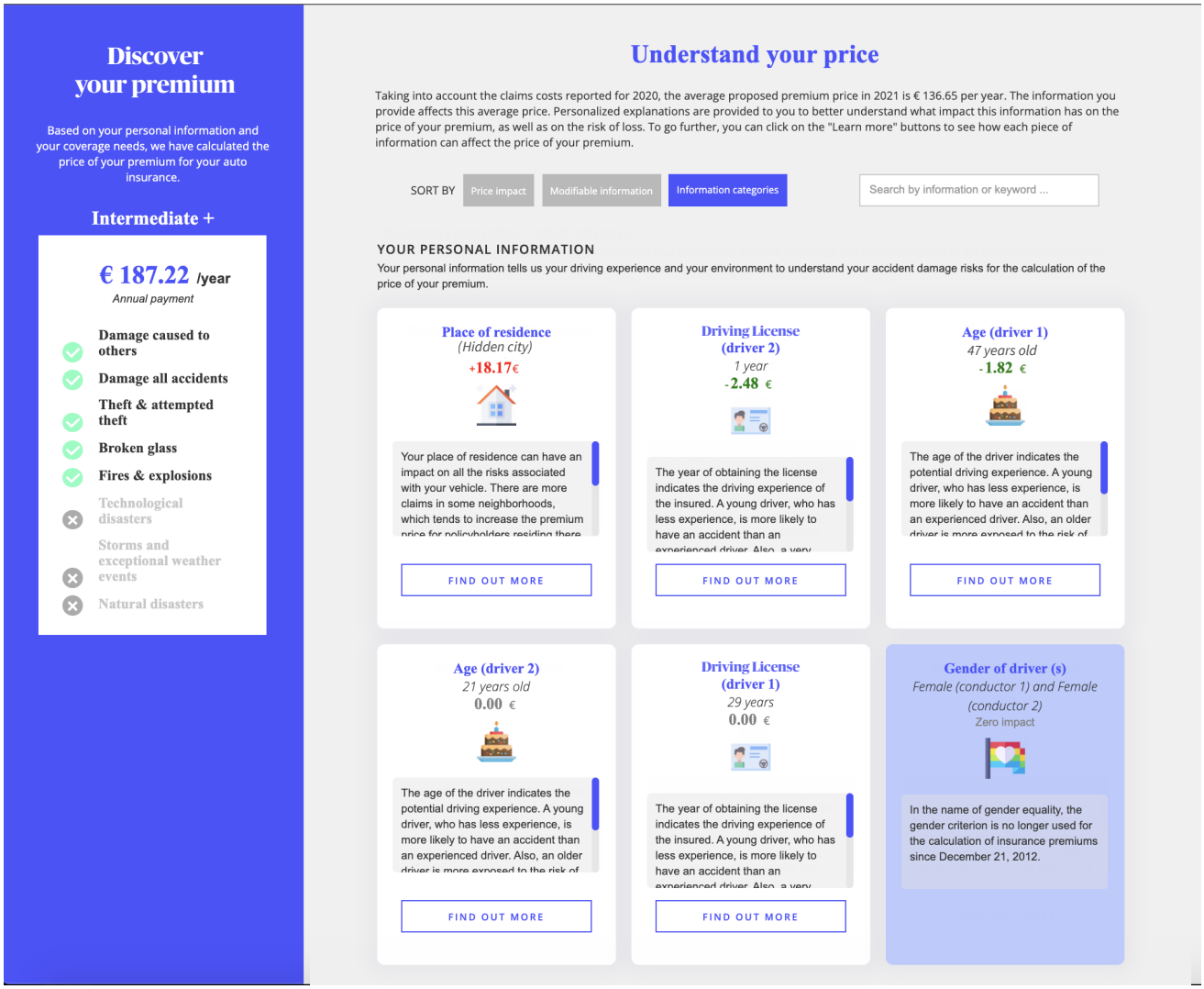


Figure 1: Application of contextualization and exploration principles in a fictive insurance-related scenario. This interface presents a personalized premium price for a prospective client on the left, and explanations on the right. *Note: The interface has been translated from the original language used for the evaluation.*

External Transparency. We introduce an external factor card into the list of feature-associated cards. It has a similar design to feature-associated ones, except that it has a different background color to be visually differentiated by users (see Figure 1). It displays a feature that is external to the ML model but has contextual importance for users. In the context of an insurance-related service, this generic principle applies to gender, information requested from the users so that the system knows how to address them but not used by the prediction model. Users may be suspicious about how their gender can be used to affect the price they get for a car insurance. Therefore, we explicitly display that this piece of information is not used by the model.

4.2.3 *Implementing exploration principles.*

Interactive display. We design filter buttons above the list of the feature-associated cards (see Figure 1), allowing users to change the ordering of the cards according to their goals. We propose three sorting options to match users' needs. First, cards can be displayed in decreasing order of the absolute values of the feature importance, so users can see which features influence most the price they get. Second, cards can be sorted so as to display first the cards that correspond to actionable features, i.e. features that can be realistically edited by users (e.g. users can switch the type of coverage they want, but they cannot change the date when they obtained their driving license), so that they can try to optimize the price they get [31]. Third, cards can be sorted according to the categories of information they contain, so as to follow the logic of the input stage (i.e. when users fill in their information). Thus,

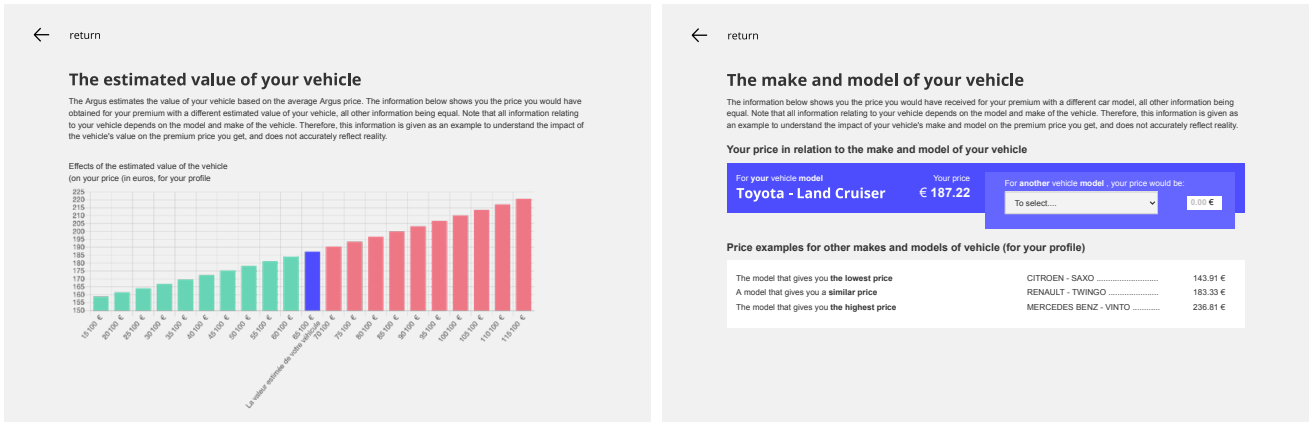


Figure 2: Application of the example-based principles for allowing exploration. For features with continuous values, we use bar graph to display example-based explanations as illustrated on the left. For features with categorial values, we display a drop-down list of the most frequent values of the feature with associated prediction, as well as three relevant examples of feature values, as illustrated on the right. *Note: The interface has been translated from the original language used for the evaluation*

users can find a logical path between the input stage and the output stage (i.e. display of the predicted price with explanations).

Example-based explanations. We place on each feature-associated card a button to access a second page displaying details in the form of example-based explanations, as illustrated in Figure 2.

In the considered pricing scenario, most features take numerical, continuous values. For these features, we propose to display a bar graph with up to twenty potential values and the associated predicted prices. Bars have difference colors to allow users to identify easily the different effects on the predicted price: blue identifies the user's value or values with the same predicted price; red (resp. green) is used for values increasing (resp. decreasing) the predicted price.

For categorical features, we propose to display a drop-down list of the most frequent values of the feature with the associated predicted prices. In addition to this list, we also display three more examples of feature values: one for the highest and lowest predicted prices, as well as one for a similar predicted price but with a different feature value than the one from the user. This allows users to know where their information fit in the overall data distribution.

4.2.4 Combining contextualization and exploration principles. We believe that the ML transparency principle can be beneficial to both exploration principles in the explainable interface. As described in Section 4.2.3, the principles we propose introduce new modalities of interaction and new explanations. In a hybrid approach, the ML transparency can make explicit how to use these new exploratory features.

For the interactive display, we implement an introduction text at the beginning of each category of feature-associated card for each filter option we designed (see Figure 1). The purpose of this introduction text is to help users understand what the categories of features are and how to interact with them.

For the example-based explanation, we implement information about the purpose of these second layers of explanations in addition to the local feature importance one, and guide users on how to interpret them towards the prediction they get (see Figure 2).

5 EXPERIMENT

To answer the studied research questions and to evaluate the effectiveness of the XAI principles we propose, we describe in turn below the material and the method we use to conduct the monitored study at the INSEAD-Sorbonne University Behavioural Lab.

5.1 Material

In this section, we present the interactive prototype we develop as the basis for the evaluation. We use a ML model to predict a personalized price for a prospective car insurance customer and extract explanations for this price with the SHAP method [21], as described in Section 5.1.1. We use this prototype to test our hypothesis towards the effectiveness of the XAI principles we propose on two dimensions of user's understanding, as described in Section 5.1.2

5.1.1 Interactive prototype. We develop an interactive prototype for a car insurance pricing interface, as described in Section 4.1. We describe in turn below the model we use in order to present SHAP and partial dependence explanations to users, the dataset and the explanation extraction.

Pricing model. We develop, with the help of an actuarial expert, and use a combination of two ML models to compute a personalized price for each user. The first model is a Gamma model which estimates the average price of a sinister for a specific person; the second one is a Poisson model which estimates the frequency of a sinister for a specific person. The final individualized price is obtained as the product of the two estimations.

Dataset. These models are trained using pg17trainpol and pg17trainclaim [11], two training datasets used for the 2017 pricing game of the French Institute of Actuaries. Pg17trainpol contains 100,000 policies for

private motor insurance and pg17trainclaim contains 14,243 claims for third-party liability risks of these 100,00 policies.

Explanation extraction. We use SHAP [21] to generate local feature importance explanations for the price estimation. It is one of the standard local feature importance methods. It provides the contribution of each feature value to the prediction as compared to the average prediction. To generate example-based explanations, we compute partial dependence plots for each feature. For a given feature, we compute the price obtained when this feature value changes while keeping all other feature values unchanged. Then, we adapt the display depending on whether the feature is continuous or categorical, as explained in Section 4.2.3.

5.1.2 Hypothesis testing. We use the interactive prototype to answer the studied research questions on the effectiveness of the principles we propose towards the objective understanding and satisfaction of non-expert users. We expect that the principles we propose increase both the objective understanding and user's satisfaction, as presented in Section 2.3. We also expect that the interaction of these principles improves even more both dimension of user's understanding. More formally, we consider null hypotheses of the form "the considered factor provides no significant improvement of the considered score" for each of the two factors (contextualization and exploration) and their interaction, and for each of the two scores (objective understanding and satisfaction).

5.2 Method

We describe in turn the experiment setup, the evaluation questionnaires, the study procedure and the method to analyze the collected results. The method has been approved by an Institutional Review Board (IRB). We pre-tested it with 4 participants at the INSEAD-Sorbonne University Behavioural Lab to validate the understanding XAI interfaces and questionnaires presented in this section, and to adjust the vocabulary used in the questions.

5.2.1 Experiment setup. We recruited non-expert participants from a large open network of volunteers of the INSEAD-Sorbonne University Behavioural Lab. Participants were randomly assigned to one of the four versions of the interface, allowing us to compare the impact of both contextualization and exploration factors on scores of objective understanding and satisfaction. We discuss in turn below the participant recruitment and interfaces they were assigned to.

Participant recruitment. We recruited 91 participants from a large open network of volunteers at the INSEAD-Sorbonne University Behavioural Lab, filtered to meet the requirements of our experiments. Participants were aged from 18 to 35 (average: 24.5 ± 3.8), had various demographics (e.g. gender, job position, level of study, driving experience). To ensure the participants were non-experts in both AI and insurance, we asked them to self-report their literacy for both topics on a 6-point Likert scale. We excluded the data of 2 participants who reported literacy scores between 4 to 5 at the end of the experiment, despite the initial filtering. After checking the screen recordings, we also excluded 9 participants who answered the questions without ever interacting with the interface. The results analyzed in the next sections thus rely on the evaluation collected from 80 participants, evenly distributed across the four versions of the interfaces we proposed. All participants were financially compensated at the end of the experiment.

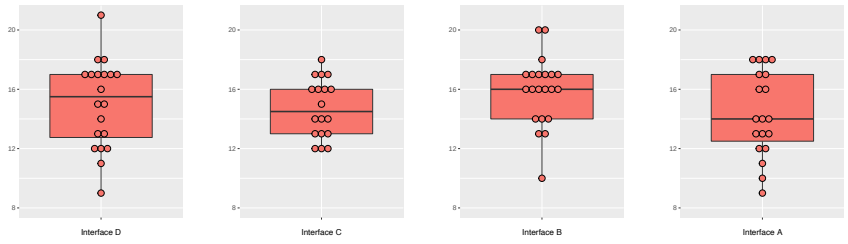
Tested interfaces. In this monitored experiment, we use 4 versions of our interface to evaluate all four conditions required for the hypothesis testing. One corresponds to the interface described in Section 4, and the others are partial variants which implement none or only one category of principles (contextualization vs exploration). We do so in order to be able to evaluate the impact of each factor, as well as their possible interaction when they are associated. More precisely, the different versions are designed as follows:

- Interface A is the baseline interface without any factor. It simply displays the local feature importance explanations with the card-based design described in Section 4.2.1. None of our design principles are applied in this version.
- Interface B is the contextualization factor interface. It adds to interface A the three principles we propose for contextualization: ML transparency, domain transparency and external transparency (see Section 4.2.2).
- Interface C is the exploration factor interface. It adds to interface A the two principles we propose for allowing exploration: the interactive display and the example-based explanations (see Section 4.2.3).
- Interface D is the interaction interface. It combines all the principles of contextualization and exploration (see Section 4.2.4). Figures 1 and 2 present screenshots of this version.

5.2.2 Evaluation questionnaires. Evaluating the effectiveness of explanations remains a challenging task [6, 7, 20, 24, 25] for which numerous methods and quality criteria have been proposed (e.g. see the survey proposed by Hoffman *et al.* in [13]). A consensus has recently been reached, according to which this assessment needs to take into account two distinct components, evaluating both objective understanding and subjective satisfaction [6, 7]. In this work, we follow this approach and describe in the following paragraphs the two questionnaires we use in the experiment, as well as an additional demographic questionnaire.

Objective understanding. We propose a questionnaire approach, similar to Cheng *et al.* [6]. Each item in the questionnaire is a statement, for which users can either answer "true", "false" or "I don't know". We design three types of questions to capture different components of user understanding:

- *(i) Explanations' scope* questions measure the extent to which users understand what information the ML system is using to give a prediction. *e.g. Feature X impacts the prediction Marianne gets.*
- *(ii) Explanations' effects* questions measure the ability of users to understand the type of effect a feature importance has on the prediction they get. *e.g. Feature X has a positive effect on the prediction Marianne gets.*
- *(iii) Explanations' locality* questions measure the users' understanding of the difference between the influence of their attributes and global explanations. *e.g. Feature X would probably have a different impact on the prediction for another person.*



#1 Contextualization and Exploration	#2 Exploration only	#3 Contextualization only	#4 No factor
15.05 (± 2.96)	14.32 (± 2.60)	15.90 (± 2.32)	14.37 (± 2.87)

Table 2: Objective understanding scores for all four conditions of the 2x2 factorial design. An overview of the descriptive statistics is displayed on top with boxplots figures.

For each question, an expected answer is predefined. We consider a participant provides a correct answer if his/her answer is identical to the expected one.

Self-reported satisfaction. We adapt the eight item self-reporting questions from the Explanation Satisfaction Scale [13], in order to assess users' satisfaction. Participants are required to answer on a 6-point Likert scale, from "Strongly disagree" (1) to "Strongly agree" (6), as it has been shown that 6-point response scales are a reasonable format for psychological studies [30].

Demographics. In addition to the previous items which are related to our research questions, a demographic questionnaire includes two questions regarding the participant literacy in artificial intelligence/machine learning and insurance, again using 6-point Likert scales, from "Not familiar at all" to "Strongly familiar", to ensure that participants are indeed non-expert users. It also asks participants their familiarity with driving, car insurance, driving frequency and claim experiences. Finally, we collect basic demographic information such as age, gender and education level. Participants can also share their insights and comments on the study in open response questions.

5.2.3 Study procedure. We conduct the user study in a lab setting at INSEAD-Sorbonne University Behavioural Lab, as it has been shown that the presence of a moderator increases participants' focus [7]. It also allows them to ask questions throughout the evaluation to make sure they understand the instructions.

After giving written consent and prior to the experiment, participants are introduced to the following experimental scenario: "Marianne, a 43 year-old woman, is looking for a new insurance for the car that she and her 21 year-old daughter drive. She decided to use our XAI interface to understand the impact of her information on her insurance price, and has now some questions about the explanations she receives". The role of the participants is to advise her about these explanations. This scenario allows us to present the same information and explanations to all participants, which makes the comparison and the statistical analysis significantly easier than if participants inputted their own information into the ML system.

Then, each participant is randomly assigned to one version of the interface for the evaluation. They take the objective understanding questionnaire (see Section 5.2.2) while interacting with the interface, and then answer the subjective satisfaction questionnaire (see Section 5.2.2). At the end of the experiment, participants complete the demographic survey.

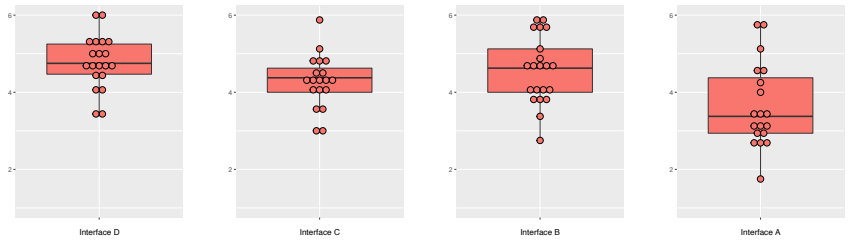
5.2.4 Data analysis. We remove one extreme outlier (below 'Q1 - 3xIQR' for the exploration factor regarding the satisfaction rate). As the collected data are normally distributed, we use 2x2 factorial ANOVA to analyze the effects of the two factors, contextualization and exploration, to test our hypotheses as presented in the previous Section. Table 4 displays the results for the scores obtained in the experiment. The objective understanding is rated from 0 to 22 corresponding to the number of correct answers for the 22 questions of the objective understanding questionnaire. The user's satisfaction is reported from 1 to 6 corresponding to the average score over the eight satisfaction's dimensions. The significance level is defined as $\alpha = .05$. We do not use the Bonferroni correction since we compare conditions that are orthogonally manipulated. Tables 2 and 3 also show comparative boxplots for the objective understanding and satisfaction scores obtained for all four conditions (contextualization principles effects, exploration principles effects, the interaction of both, and the absence of principles), with one datapoint for each participant.

6 RESULTS

We use the results presented in Tables 2, 3 and 4 to answer the two research questions we consider regarding objective understanding in Section 6.1 and user's satisfaction in Section 6.2.

6.1 RQ1: How effective are contextualizing and allowing exploration for improving non-experts users' objective understanding of Local Feature Importance explanations?

We analyze the significant effects of both contextualization and exploration factors on user's objective understanding score. As neither exploration factor nor the interaction of the two factors show significant impacts, we select the best model to analyze these statistical differences and use one-way ANOVA to measure the effect of contextualization on the objective understanding score. The



#1 Contextualization and Exploration	#2 Exploration only	#3 Contextualization only	#4 No factor
4.79 (± .69)	4.27 (±.69)	4.57 (± .86)	3.65 (± 1.10)

Table 3: Satisfaction scores for all four conditions of the 2x2 factorial design. An overview of the descriptive statistics is displayed on top with boxplots figures.

analysis of Table 4 leads to three main observations commented in turn below. First, contextualization leads to the biggest improvement in objective understanding, and is close to reach the level of statistical significance. In contrast, both exploration and the interaction of contextualization and exploration do not improve objective understanding overall.

Contextualization improves objective understanding. On the boxplots presented above Table 2, we can see the interface including contextualization principle only (interface B) shows the highest improvement in objective understanding with an average score of 15.90 correct answers out of 22, i.e. .85 point more than when these principles are paired with exploration ones (interface D) or 1.53 point more than when no principles are applied (interface A). When comparing the average means of conditions with contextualization principles applied (interfaces B and D) in Table 4, we observe that the contextualization factor increases by +1.15 points the objective understanding score. This difference is not statistically significant at 5% level however it is close (t=1.90 p=.06).

Although **we fail to reject the null hypothesis, these observations lead us to believe that contextualizing local feature importance is a promising tool to improve non-expert users objective understanding (H.1.1).**

Exploration does not have a significant impact. Table 2 shows that participants with the interface including exploration principles (interface C) obtain the lowest average score of objective understanding of all four conditions. When comparing the impact of exploration factor, we observe a similar trend as the average score for all conditions including exploration principles is .48 point lower than when not applied. Yet, we do not observe a significant impact of exploration principles on objective understanding.

Thus, we **fail to reject the null hypothesis** and are not able to demonstrate the positive effect of exploration on the objective understanding of local feature importance in our context (H.1.2).

The interaction of contextualization and exploration does not have a significant impact neither. Previous observations suggest a promising positive effectiveness of contextualization but reject exploration one regarding the objective understanding. When analyzing the interaction effect in a two-way ANOVA, we see no statistically significant impact.

Thus, we **fail to reject the null hypothesis** and are not able to demonstrate that the interaction of contextualization and exploration principles improves even more objective understanding of non-expert users (H1.3).

6.2 RQ2: How effective are contextualizing and allowing exploration for improving non-experts users' satisfaction of Local Feature Importance explanations?

Similarly to the previous analysis for objective understanding, we analyze the significant effects of both contextualization and exploration factors on user's satisfaction score. As the interaction of the two factors shows no significant impact again, we select the best model to analyze these statistical differences and use two-way ANOVA to measure the effect of contextualization and exploration factors on the satisfaction score. Table 3 shows that all three conditions with the principles we propose have an average satisfaction score higher than when no principle is applied (+.62 point for exploration principles, +.93 point for contextualization principles, +1.14 point for the combination of both principles). These differences are significant for both contextualization and exploration factors, which leads us to conclude that both factors significantly improve users' satisfaction. In contrast, the interaction of the two factors does not show a significant impact on users satisfaction. These conclusions are discussed in turn below.

Contextualization significantly improves users satisfaction. Table 3 shows that contextualization principles (interface B) obtain a higher satisfaction rate as they increase by .93 point the average satisfaction score compared to the interface without these principles (interface A), and by .30 point as compared to the interface with exploration principles (interface C). The positive effect of contextualization principles on users satisfaction can also be observed by the datapoint distribution for each participant in the boxplots displayed above Table 3. This difference is also observed in the two-way ANOVA analysis in Table 4 as the average mean

Objective understanding				
	With factor	**Without factor**	**One-way ANOVA**	
	means (sd)	**means(sd)**	***t*-value**	***p*-value**
Contextualization	15.49 (± 2.64)	14.34 (± 2.73)	1.90	.06°
Exploration	14.68 (± 2.78)	15.13 (± 2.59)		

Satisfaction				
	With factor	**Without factor**	**Two-way ANOVA** (intercept mean = 3.76)	
	means (sd)	**means(sd)**	***t*-value**	***p*-value**
Contextualization	4.68 (± .77)	3.96 (± .89)	3.80	.0003***
Exploration	4.53 (± .69)	4.11 (± .98)	2.15	.03*

Significance code: *** p<.001 ; ** p<.01 ; * p<.05 ; ° p<.1

Table 4: Comparing improvement of objective understanding between two factors: contextualization and exploration. The results of a one-way ANOVA regarding the significant effect of contextualization factor on user's objective understanding are displayed on the top. The results of a two-way ANOVA regarding the significant effect of contextualization and exploration factors on user's satisfaction are displayed on the bottom.

for contextualization factor is +.72 point significantly higher than the average mean for interfaces without (t=3.80 p=.0003).

Thus, we **reject the null hypothesis** as contextualization parameter is greater than the claimed value and conclude that **contextualization significantly improves non-expert users' satisfaction (H2.1)**.

Exploration also significantly improves users satisfaction. Similarly to interfaces including contextualization principles, Table 3 shows that the interface including the exploration ones (interface C) increases users satisfaction by .62 point as compared to the interface without any principle applied (interface A). The positive effect of the exploration principles can also be observed by the datapoints distribution for each participant in the boxplots displayed above Table 3. When analyzing the impact of the exploration factor in Table 4, the results of the two-way ANOVA analysis shows that the average mean for the exploration factor is +.42 significantly higher than the average mean for interfaces without it (t=1.90 p=.03).

Thus, we **reject the null hypothesis** as the exploration parameter is greater than the claimed value and conclude that **exploration significantly improves non-expert users' satisfaction (H2.2)**.

The interaction of both principles does not have a statistically significant impact. First, Table 3 shows that the combination of both principles (interface D) has the highest improvement as it increases by +1.14 points participants' satisfaction rates as compared to the interface without any principles (interface A), by +.22 point as compared to the interface with only contextualization principles (interface B) and by +.52 as point compared to the interface with only exploration principles (interface C). On the boxplots figures displayed above Table 3, we observe that the 1st quartile for the combination condition is 4.47, which +.10 point higher than the 3rd quartile for the condition without any principle applied, meaning that 75% of participants interacting with the contextualization and exploration gave higher satisfaction rates than 75% of participants using interfaces without any principle applied.

Yet, the interaction of the two factors has no statistical significant impact. Thus, **we fail to reject the null hypothesis** and are not able to demonstrate the positive effect of both principles interaction on users satisfaction (H.2.3).

7 CONCLUSION

In this paper, we propose generic design principles for contextualization and exploration of local feature importance explanations for non-expert users. We also propose an implementation of these principles into a user interface for a car insurance pricing scenario. The experiment we conduct in a moderated lab setting shows that the contextualization principles we propose significantly improve user's satisfaction and are close to significantly improve user's objective understanding. Also, the results show that the exploration principle we propose significantly improves user's satisfaction. On the other hand, the interaction of these principles does not appear to bring significant improvement on both dimensions of users' understanding.

It is noteworthy that the results we obtain differ from the ones presented in the close work of Cheng *et al.* in [6]. In their experiments, allowing users to interact with the ML model improves their objective understanding, but does not increase their satisfaction in the system, whereas we observe the opposite trend. One possible explanation for these diverging results could be the difference in the considered application domain: insurance is often perceived as an opaque industry [28], as confirmed by several participants of the preliminary user workshops we conducted. It is possible that participants in our experiments have low expectations when it comes to the transparency of insurance solutions, which could lead them to consider any insurance solutions that are willing to expose their ML model as more trustworthy. The same observation can be made about the contextualization principles we propose: it is possible that part of the observed improvement in satisfaction ratings is due to the perceived opaqueness of the insurance industry.

Future works will aim at investigating this hypothesis, in particular extending the conducted study to other application domains, in order to have a more comprehensive view of the impact of the principles we propose. Other directions for refining the conducted

study will focus on other possible effects of interest. The latter e.g. include a possible correlation between objective understanding and subjective satisfaction, or a possible effect of a notion of user engagement in the explanation interaction that could be derived from the collected information about their having a driving license. Another direction is to increase the number of participants: the current number is e.g. not high enough to allow a comparison of the three contextualization principles we propose (ML, domain or external transparency, as well as their combined effect). Future works will aim at performing a wider study making it possible to investigate their potential differences. Conducting more detailed analyses to evaluate the effect of the collected demographic information will also make it possible to obtain more detailed insights about the effectiveness of explanations provided to non-expert users, tackling one of the major current challenges of the XAI community.

ACKNOWLEDGMENTS

We would like to thank our research colleagues Vincent Grari who provided us with the ML system and the expert knowledge on car insurance; Thibault Laugel who helped us extract explanations from SHAP method; Anne Sheehy who provided us support in the statistical analysis; and anonymous reviewers for their valuable comments. We also warmly thank Hoai Huong Ngo, Germain Dépetasse and Sébastien Robin, members of the INSEAD-Sorbonne University Behavioural lab, who supported us with the user study.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human-Computer Interaction* 16, 2-4 (2001), 193–212.

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'18*. Association for Computing Machinery, New York, NY, USA, 1–14.

[4] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[5] Carrie Cai, Martin Stumpe, Michael Terry, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, and Greg Corrado. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19*. Association for Computing Machinery, New York, NY, USA, 1–14.

[6] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19*. Association for Computing Machinery, New York, NY, USA, 1–12.

[7] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proc. of the 26th Int. Conf. on Intelligent User Interfaces, IUI'21*. Association for Computing Machinery, New York, NY, USA, 307–317.

[8] Dennis Collaris, Leo M Vink, and Jarke J van Wijk. 2018. Instance-level explanations for fraud detection: A case study. arXiv:1806.07129

[9] Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proc. of the 25th Int. Conf. on Intelligent User Interfaces, IUI'20*. Association for Computing Machinery, New York, NY, USA, 510–518.

[10] Michael A DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The algorithm and the user: How can hci use lay understandings of algorithmic systems?. In *Extended Abstracts of the Int. Conf. on Human Factors in Computing Systems, CHI'18*. Association for Computing Machinery, New York, NY, USA, 1–6.

[11] Christophe Dutang and Arthur Charpentier. 2020. Package 'CASdatasets'.

[12] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: visual counterfactual explanations for machine learning models. In *Proc. of the 25th Int. Conf. on Intelligent User Interfaces, IUI'20*. Association for Computing Machinery, New York, NY, USA, 531–535.

[13] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608

[14] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19*. Association for Computing Machinery, New York, NY, USA, 1–13.

[15] Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. 2018. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In *Proc. of the Int. Cross-Domain Conf. for Machine Learning and Knowledge Extraction, CD-MAKE'18*. Springer International Publishing, Cham, 1–8.

[16] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'16*. Association for Computing Machinery, New York, NY, USA, 5686–5697.

[17] Freddy Lecue. 2020. On the role of knowledge graphs in explainable AI. *Semantic Web* 11, 1 (2020), 41–51.

[18] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proc. of the Int. Conference on Human Factors in Computing Systems, CHI'20*. Association for Computing Machinery, New York, NY, USA, 1–15.

[19] Brian Y Lim. 2012. *Improving understanding and trust with intelligibility in context-aware applications*. Ph. D. Dissertation. Carnegie Mellon University.

[20] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. In *Proc. of the Int. Conf. on Machine Learning, ICML'16 - Workshop on Human Interpretability in Machine Learning*. Association for Computing Machinery, New York, NY, USA, 36–43.

[21] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proc. of the Int. Conf of Advances in Neural Information Processing Systems, NeurIPS'17*. Curran Associates Inc., Red Hook, NY, USA, 4765–4774.

[22] Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press, Cambridge, MA, USA.

[23] Kyle Martin, Anne Liret, Nirmalie Wiratunga, Gilbert Owusu, and Mathias Kern. 2019. Developing a catalogue of explainability methods to support expert and non-expert users. In *Proc. of the In. Conf. on Innovative Techniques and Applications of Artificial Intelligence, IAAI'19*. Springer-Verlag, Berlin, Heidelberg, 309–324.

[24] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[25] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. arXiv:1811.11839

[26] Boris Ruf, Chaouki Boutharouite, and Marcin Detyniecki. 2020. Getting Fairness Right: Towards a Toolbox for Practitioners.

[27] Md Kamruzzaman Sarker, Joshua Schwartz, Pascal Hitzler, Lu Zhou, Srikanth Nadella, Brandon Minnery, Ion Juvina, Michael L Raymer, and William R Aue. 2020. Wikipedia knowledge graph for explainable AI. In *Proc. of the Iberoamerican Conf. of Knowledge Graphs and Semantic Web, KGSWC'20*. Springer International Publishing, Cham, 72–87.

[28] Daniel Schwarcz. 2014. Transparently Opaque: Understanding the Lack of Transparency in Insurance Consumer Protection, 61UCLA L. *Rev* 394 (2014), 400.

[29] Ramprasaath R Selvaraju, Prithvijit Chattopadhyay, Mohamed Elhoseiny, Tilak Sharma, Dhruv Batra, Devi Parikh, and Stefan Lee. 2018. Choose your neuron: Incorporating domain knowledge through neuron-importance. In *Proc. of the European Conf. on Computer Vision, ECCV'18*. Springer, Cham, 540–556.

[30] Leonard J Simms, Kerry Zelazny, Trevor F Williams, and Lee Bernstein. 2019. Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological assessment* 31, 4 (2019), 557.

[31] Ronal Singh, Paul Dourish, Piers Howe, Tim Miller, Liz Sonenberg, Eduardo Velloso, and Frank Vetere. 2021. Directive explanations for actionable explainability in machine learning applications. arXiv:2102.02671

[32] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proc. of the 26th Int. Conf. on Intelligent User Interfaces, IUI'21*. Association for Computing Machinery, New York, NY, USA, 109–119.

[33] Patrick Van Esch, J Stewart Black, and Joseph Ferolie. 2019. Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior* 90 (2019), 215–222.

[34] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.

[35] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19*. Association for Computing Machinery, New York, NY, USA, 1–15.

[36] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proc. of the 26th Int. Conf. on Intelligent User Interfaces, IUI'21*. Association for Computing Machinery, New York, NY, USA, 318–328.

[37] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proc. of the Int. Conf. on Intelligent User Interfaces, IUI'20*. Association for Computing Machinery, New York, NY, USA, 189–201.

[38] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.