Sorbonne Université

École doctorale Informatique, Télécommunications et Électronique (Paris)

Équipe LFI, LIP6

Conception et évaluation d'interfaces utilisateur explicatives pour systèmes complexes en apprentissage automatique

Par Clara Bove

Thèse de doctorat d'Informatique

Présentée et soutenue publiquement le ** "mois" 2023

Devant le jury composé de :

Nicolas Labroche	LIFAT, Université de Tours	Rapporteur
Katrien Verbert	AugmentHCI, KU Leven	Rapportrice
Jean-Gabriel Ganascia	LIP6, Sorbonne Université	Examinateur
Theodoros Evgeniou	Decision Sciences and Technology Management, INSEAD	Examinateur
Marie-Jeanne Lesot	LIP6, Sorbonne Université	Directrice de thèse
Charles Tijus	Lutin-CHart, Université Paris 08	Directeur de thèse
Marcin Detyniecki	AXA, Paris	Directeur de thèse

Designing and evaluating explanation user interfaces for complex Machine Learning systems

Acknowledgements

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

Résumé

Cette thèse se place dans le domaine de l'IA eXplicable (XAI) centrée sur l'humain, et plus particulièrement sur l'intelligibilité des explications pour les utilisateurs non-experts. Le contexte technique est le suivant : d'un côté, un classificateur ou un régresseur opaque fournit une prédiction, et une approche XAI *post-hoc* génère des informations qui agissent comme des explications ; de l'autre côté, l' utilisateur reçoit à la fois la prédiction et ces explications. Dans ce contexte, plusieurs problèmes peuvent limiter la qualité des explications. Ceux sur lesquels nous nous concentrons sont : le manque d'informations contextuelles dans les explications, le manque d'orientation pour la conception de fonctionnalités pour permettre à l'utilisateur d'explorer et la confusion potentielle qui peut être générée par la quantité d'informations.

Nous développons une procédure expérimental pour concevoir des interfaces utilisateur explictives et évaluer leur intelligibilité pour les utilisateurs non-experts. Nous étudions des opportunités d'amélioration XAI sur deux types types d'explications locales : l'importance des variables et les exemples contrefactuels. Aussi, nous proposons des principes XAI génériques pour contextualiser et permettre l'exploration sur l'importance des variables; ainsi que pour guider les utilisateurs dans l'analyse comparative des explications contrefactuelles avec plusieurs exemples. Nous proposons une application de ces principes proposés dans deux interfaces utilisateur explicatives distinctes, respectivement pour un scénario d'assurance et un scénario financier. Enfin, nous utilisons ces interfaces améliorées pour mener des études utilisateurs en laboratoire et nous mesurons deux dimensions de l'intelligibilité, à savoir la compréhension objective et la satisfaction subjective. Pour l'importance des variables locales, nous montrons que la contextualisation et l'exploration améliorent l'intelligibilité de ces explications. De même, pour les exemples contrefactuels, nous montrons qu'avoir plusieurs exemples plutôt qu'un améliore également l'intelligibilité, et que l'analyse comparative est un outil prometteur pour la satisfaction des utilisateurs.

À un niveau fondamental, nous considérons la question théorique des incohérences éventuelles de ces explications. Dans le contexte considéré dans cette thèse, la qualité d'une explication repose à la fois sur la capacité du système d'apprentissage automatique à générer une explication cohérente et sur la capacité de l'utilisateur final à interpréter correctement ces explications. Cependant, il peut y avoir plusieurs limitations: d'un côté, la littérature a rapporté plusieurs limitations techniques de ces systèmes, rendant les explications potentiellement incohérentes ; de l'autre, des études utilisateurs ont montré que les interprétations des utilisateurs ne sont pas toujours exactes, même si des explications cohérentes leur ont été présentées. Nous étudions donc ces incohérences et proposons une ontologie pour structurer les incohérences les plus courantes

de la littérature. Cette ontologie constitue un outil pour comprendre les limites actuelles en XAI pour éviter les pièges des explications.

Abstract

This thesis focuses on human-centered eXplainable AI (XAI) and more specifically on the intelligibility of Machine Learning (ML) explanations for non-expert users. The technical context is as follows: on one side, either an opaque classifier or regressor provides a prediction, with an XAI post-hoc approach that generates pieces of information as explanations; on the other side, the user receives both the prediction and the explanations. Within this XAI technical context, several issues might lessen the quality of explanations. The ones we focus on are: the lack of contextual information in ML explanations, the unguided design of functionalities or the user's exploration, as well as confusion that could be caused when delivering too much information.

For solving these issues, we develop an experimental procedure to design XAI functional interfaces and evaluate the intelligibility of ML explanations by non-expert users. Doing so, we investigate the XAI enhancements provided by two types of local explanation components: feature importance and counterfactual examples. Thus, we propose generic XAI principles for contextualizing and allowing exploration on feature importance; and for guiding users in their comparative analysis of counterfactual explanations with plural examples. We propose an implementation of such principles into two distinct explanation-based user interfaces, respectively for an insurance and a financial scenarios. Finally, we use the enhanced interfaces to conduct users studies in lab settings and to measure two dimensions of intelligibility, namely objective understanding and subjective satisfaction. For local feature importance, we demonstrate that contextualization and exploration improve the intelligibility of such explanations. Similarly for counterfactual examples, we demonstrate that the plural condition improve the intelligibility as well, and that comparative analysis appears to be a promising tool for users' satisfaction.

At a fundamental level, we consider the issue of inconsistency within ML explanations from a theoretical point of view. In the explanation process considered for this thesis, the quality of an explanation relies both on the ability of the Machine Learning system to generate a coherent explanation and on the ability of the end user to make a correct interpretation of these explanations. Thus, there can be limitations: on one side, as reported in the literature, technical limitations of ML systems might produce potentially inconsistent explanations; on the other side, human inferences can be inaccurate, even if users are presented with consistent explanations. Investigating such inconsistencies, we propose an ontology to structure the most common ones from the literature. We advocate that such an ontology can be useful to understand current XAI limitations for avoiding explanations pitfalls.

Publications

The work conducted during the Ph.D program has led to the following publications:

Mentioned in this thesis

- Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualising local explanations for non-expert users: an XAI pricing interface for insurance. In *IUI Workshop on Transparent Explanations in Smart Systems (TExSS)*. CEUR, 2021

- Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proc. of the 27th Int. Conf. on Intelligent User Interfaces*, IUI '22, 2022.

- Clara Bove, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In *Proc. of the 28th Int. Conf. on Intelligent User Interfaces*, IUI '23, 2023

Submitted work

- Clara Bove, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. An ontology of inconsistencies in ML explanations. In *ArXiv*, 2023

Other work (collaboration)

The following work is not mentioned in this manuscript, but have been conducted in parallel as collaborations on related topics:

- Clara Bove, Jonathan Aigrain, and Marcin Detyniecki. Building trust in artificial conversational agents. In *IUI Workshop on Conversational User Interface (CUI)*. CEUR, 2021

Contents

1

Introduction

Artificial Intelligence (AI) is increasingly more powerful and has reached significant scientific and technological advances over years. In particular, Machine Learning (ML) models are achieving unprecedented levels of performance when learning to solve increasingly complex computational tasks (e.g., the development of Deep Learning models, in particular for Natural Language Processing tasks through Large Language Models). As a result, the applications of Machine Learning are diverse and widespread. They concern multiple industries, among which for instance healthcare (e.g., skin cancer detection with radiography pictures), insurance (e.g., fraud detection in claims), human resources (e.g., filtering applicants for a job position), law (e.g., predicting chances of recidivism for prisoners), cybersecurity (e.g., spam detection in emails) to name a few.

However, such progress leads to an increase in the complexity and sophistication of Machine Learning models. For example, tree based methods like XGB are considered extremely fast, stable, fast to tune and robust to randomness. Because these types of very accurate models are highly opaque, they are often referred to as *black-boxes*: despite these models being able to produce useful predictions, it is not possible to get any information about their internal workings. Such models can be perceived as tools whose behaviors are not clear. This opacity can create many issues when the Machine Learning models are misused, with intent or ignorance, in situations where the decisions have high impacts. They can even be dangerous or have disastrous consequences, as shown by several use cases. For example, in 2018, a pedestrian woman died after a self-driving vehicle failed to analyze the situation [1] (a pedestrian crossing the street outside of the pedestrian crossing area) and prevent the crash. The opacity of the Machine Learning model used in such application had made it impossible for

[1] https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

the ML researchers who designed the model to assess the error and debug it, leading to a deadly consequence. In another disastrous example, the COMPAS software used by several jurisdictions in the US to predict the recidivism risk of convicts was found to be racially biased in 2016 [2]. The opacity of the Machine Learning model used in this software was not only making this bias difficult to detect, but also did not allow non-AI practitioners (e.g., expert users such as judges or non-expert users such as jurors) to be able to perceive nor understand such risks, leading to juridical and ethical issues.

These limitations fuel an increasing demand for transparency from various stakeholders in AI (Preece et al., 2018; Sokol and Flach, 2020; Goodwin et al., 2022). When decisions derived from such systems can have a high impact on people's life or society, there is a need for understanding how such decisions are provided by Machine Learning systems (Lipton, 2016; Wachter et al., 2017; Goodman and Flaxman, 2017). Public and private institutions have started to address this issue. For example, the European Union enforces the "right to an explanation" for citizens in the *General Data Protection Regulation* (GDPR). Hence, the organization responsible for the development of an algorithm is compelled to explain its decisions to the concerned citizen. For instance, when a citizen applies for a loan through an automated system, the insurance company is required to ensure that the latter can provide him/her with information to explain the decision. It is likely that the use of AI will be regulated and governed to ensure that it does not have a negative impact on people or society.

Explainable AI

As discussed in more details in Chapter 2, the field of eXplainable AI (XAI) has emerged in the recent years as the scientific answer to these societal issues, and generally aims at addressing the problem of the opacity of AI systems. It is among the hottest topics in AI research, as shown by the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years (Barredo Arrieta et al., 2020).

The term XAI has initially been popularized by the DARPA (Defense Advanced Research Projects Agency) in a call for research proposals on AI explainability (Gunning, 2017). It can be defined as *"the ability to explain or to provide the meaning in understandable terms to a human"* (Doshi-Velez and Kim, 2017). Generally, XAI works refer to all the initiatives aiming at making ML models understandable to human (Adadi

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

and Berrada, 2018). Various related terms are being used by the research community to describe these works such as interpretability, explainability, transparency, intelligibility, comprehensibility, accountability to list a few. There are often disagreements on their respective scopes and the extent to which they are redundant, complementary or distinct. Several works propose to distinguish the subtle differences between these terms and their use (see Barredo Arrieta et al. (2020); Liao and Varshney (2021); Bellucci et al. (2022)). This thesis does not go into that direction and uses the terms of explainability, interpretability and intelligibility interchangeably.

Numerous approaches have been proposed to make Machine Learning models understandable from a technical point of view (see Barredo Arrieta et al. (2020); Verma et al. (2022); Guidotti et al. (2018) for recent surveys). Such models can be trained for various ML tasks such as classification (see Umadevi and Marseline (2017); Kowsari et al. (2019); Chen et al. (2021a) for some surveys), regression (see Fernández-Delgado et al. (2019); Emmert-Streib and Dehmer (2019) for some surveys), clustering (see Patibandla and Veeranjaneyulu (2018); Ahmad and Khan (2019) for some surveys), recommendations (see Koren et al. (2009); Zhang and Chen (2020) for some surveys) to name a few. This thesis focuses on the case of supervised learning tasks, classification and regression.

To make these models more transparent, two kinds of predictive models and explanation strategies are distinguished (see Biran and Cotton (2017); Guidotti et al. (2018); Barredo Arrieta et al. (2020); Zhou et al. (2021) for some surveys). First, some models have a simple and small structure and thus are considered to be interpretable by nature. This may for instance be the case of decision trees with low depth. Other predictive models, called black-box models, have complex structures and cannot be considered as interpretable. Neural networks for instance belong to that category. They are then paired with an additional system dedicated to the generation of explanations, called explainers. For instance post-hoc explainers are built on top of the predictive model to enrich their prediction with additional information (see e.g., Adadi and Berrada (2018); Linardatos et al. (2020); Barredo Arrieta et al. (2020) for some surveys). Different types of explanations can be generated from these approaches: **feature importance** techniques assign a score to input features based on how useful they are at predicting an output (e.g., LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017)), **rules** as a knowledge base that collectively make up the prediction model (e.g., RuleMatrix (Ming et al., 2018) and Anchors (Ribeiro et al., 2018) and **counterfactual** instances, that exemplify the minimal modifications that would lead to a different prediction (e.g., Growing spheres (Laugel et al., 2018a) FACE (Poyiadzi et al., 2020) and DiCE (Mothilal et al., 2020)). The generated explanations can be lo-

cal, i.e. associated to a specific prediction, or global, i.e. describing the whole model's behavior.

Human-Centered eXplainable AI

Despite having multiple XAI approaches produced by AI researchers, there are very few successful examples of XAI in real-world AI applications (Liao and Varshney, 2021). Explainability should be inherently human-centered and developing XAI applications requires to center the technical development on people's explainability needs. XAI has become an increasingly multidisciplinary research field, with relative growth in papers belonging to diverse non-computer scientific fields. The Human-Centered Explainable AI research community has emerged very recently on the scientific landscape (Wang et al., 2019; Ehsan et al., 2021), and refers to interdisciplinary works in XAI, including Human-Computer Interaction researchers and design practitioners, aiming to address the needs of the end-users and adapt accordingly the presentation of Machine Learning explanations.

Another active area of research in XAI concerns the evaluation of Machine Learning explanations. There is a need to ensure that the XAI approach used to generate explanations is adapted to the needs of the end-users, and that these explanations can be interpreted by the latter. Yet, current research in XAI is generally proposed from a computational point of view, and lack empirical research in understanding users' needs of ML explanations in their usage (Keane et al., 2021; Verma et al., 2022; Shang et al., 2022). Also, very few of them have been user tested or evaluated: only 21% of counterfactual approaches surveyed by Keane et al. (2021) have been user tested. There is often no or little empirical evidence to prove the relevance of one approach as compared to another. Thus, evaluating the intelligibility remains a challenging task. The research community has been lacking guidance over the method to apply and the measures to use when evaluating such approaches (Nunes and Jannach, 2017; Adadi and Berrada, 2018; Guidotti et al., 2018; Chromik and Schuessler, 2020; Arora et al., 2022). Although there have been some contributions on this topic (see Chromik and Schuessler (2020); Rong et al. (2022b); Arora et al. (2022) for recent surveys on XAI evaluation), the research community seems to be lacking guidance on the method and measures to use in order to assess the quality of an explanation with users (Nunes and Jannach, 2017; Adadi and Berrada, 2018; Guidotti et al., 2018).

Motivations

Research in XAI is aiming at presenting the users with intelligible explanations. We investigate in particular three research directions in this thesis.

First, the notion of an explanation and its purpose has been studied in various fields of research (e.g., philosophy, cognitive sciences, human-computer interaction (Miller, 2019)) and in different contexts (e.g., in social interactions, between human and various kinds of machine (Sokol and Flach, 2020). Hence, it requires to investigate outside of the scope of research in computer science and incorporate key insights in social sciences in order to have human-centered approach in XAI. We describe such insights in Section 2.1.

Moreover, multidisciplinary contributions in XAI are investigating the design of user interfaces for different types of explanations and users needs. These interfaces, called explanation user interfaces (XUIs) (Chromik et al., 2021) are defined as the sum of outputs of an XAI system that the user can directly interact with. Such interfaces are designed for various types of explanations (e.g., counterfactual explanations in visual interface (Gomez et al., 2020) or interactive interface for rules (Ming et al., 2018)), as well as for the users' various needs and levels of expertise (e.g., a loan applicant with little to no knowledge in AI or in finance).

Finally, evaluating and measuring the quality of an explanation is a challenging task. Different methods from social sciences (e.g., application-grounded and human-subjects evaluations (Doshi-Velez and Kim, 2017; Nauta et al., 2022)) and metrics in XAI (e.g., objective measures such as understanding, usability; subjective measures such as trust, fairness (Chromik and Schuessler, 2020; Rong et al., 2022b)) are used in recent contributions on human-centered XAI.

Contributions

In this thesis, we consider the following context: on one side, a classifier or a regression opaque model provides a prediction, and a post-hoc XAI approach generates pieces of information that act as explanations; on the other side, a user receives both the prediction and the explanations.

We aim at investigating the process for designing and evaluating Machine Learning explanations so that the end-users understand and find them useful. We focus on the needs of non-expert users whose questions are oriented towards the understanding of the predicted outcome. Hence, we use two types of local explanations to build the process: local feature importance and plural counterfactual examples.

We analyze local approaches for feature importance explanations (e.g., LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017)) and highlight the lack of both contextual information and interactive functionalities needed to compensate for the knowledge gaps of the non-expert users and allow them to test their hypotheses. Thus, we investigate two design enhancements for explanation user interfaces, namely contextualizing and allowing exploration. We also investigate the method and the metrics that can be used for the evaluation of the intelligibility of such explanations. We propose design principles and an implementation into an explanation user interface for an insurance scenario. This interface is then used to conduct a user study in lab settings and we measure two dimensions of the intelligibility, namely objective understanding and subjective satisfaction. We demonstrate that these design enhancements improve the intelligibility of such explanations.

Regarding counterfactual explanations, we highlight that most of these approaches have not been user tested. Yet, when having plural counterfactual explanations (i.e., for one instance and a given prediction, having multiple counterexamples instead of a single one), too many examples can create confusions. In this work, we investigate the intelligibility of counterfactual explanations, and compare two cases: having a single example and having plural examples. We also investigate design enhancements for comparative analysis functionalities in an explanation user interface. This work allows us to refine the proposed evaluation method: we propose enriched questionnaires for measuring intelligibility and we also integrate qualitative measures to collect insights from the user study.

At a fundamental level, we consider the theoretical question of possible inconsistencies in Machine Learning explanations. In the explanation process considered for this thesis, the quality of an explanation relies both on the ability of the Machine Learning system to generate a coherent explanation, as well as on the ability of the end user to make a correct interpretation of the explanations. Yet, there can be limitations on both sides: the literature has reported several technical limitations of such systems, making the explanations potentially inconsistent; also, user studies have demonstrated that users' inferences are not always accurate, even if they have been presented with consistent explanations. Thus, we investigate such inconsistencies and propose an ontology to structure most common ones from the literature. This ontology can be useful to understand the current limitations in XAI, and avoid explanations pitfalls.

Document structure

The thesis is structured as follows. After a brief overview of the very vast domain of eXplainable AI, Chapter 2 presents the recent research direction in XAI with a focus on the human aspect of Machine Learning explanations.

Chapters 3 to 5 discuss in turn the three contributions summarized above: Chapter 3 is dedicated to the study of design enhancements for local feature importance explanations in an XUI for non-expert users; similarly, Chapter 4 is dedicated to design enhancements for counterfactual explanations with plural examples; and Chapter 5 is devoted to discussing and structuring the notion of inconsistencies in ML explanations.

Finally, the manuscript ends by summarizing the contributions of this thesis and discussing the perspectives it opens.

2

Background and related works

So called black-box machine learning models achieve very high accuracy levels possibly at the expense of their intelligibility and transparency. For instance, they include deep learning models. Hence, the research community has been very active on the topic of eXplainable AI (XAI), and numerous approaches have been proposed to improve both this needed algorithm transparency and human understanding of such models (see Adadi and Berrada (2018); Mohseni et al. (2018); Guidotti et al. (2018); Barredo Arrieta et al. (2020) for some surveys). Since 2020, contributions in XAI have shifted from a technical focus to more human-centered approaches that emphasize on the users and how to presentation them with ML explanations (Liao and Varshney, 2021; Chromik and Butz, 2021; Ehsan et al., 2022; Szymanski et al., 2022b). This human-centered research field has become a prominent interdisciplinary domain in the past years, including machine learning, data science and visualization, human-computer interaction, design, psychology or law. Research in XAI takes insights from social sciences fields, to be able to present qualitative explanations to the end-users, and to evaluate them. There is also an increasing number of contributions from the Human-computer interaction (HCI) research community on user interfaces for the display of ML explanations.

In this chapter, we first discuss the key notions of an explanation from research in social sciences in Section 2.1. We then present the challenges and current research directions of explainability in Section 2.2. In Section 2.3, we review some HCI contributions for explanations user interfaces (XUIs). We then discuss another XAI challenge regarding the evaluation of the explanations in Section 2.4 and conclude in Section 2.5.

2.1 | Explanation in social sciences

The purpose of research in XAI is to provide users with explanations regarding a model's behavior and the prediction. A first important step in this process is to reflect on what is an explanation generally speaking, and to take perspective outside the computing perimeter. The notion of an explanation has been widely studied in various fields of research. In particular, insights from social sciences works provide a better understanding on the needs for an explanation, how do people build and communicate the latter. De Graaf and Malle (2017) argue that because people assign human-like traits to artificial agents, they will expect explanations to have similar conceptual framework as the one used to explain human behaviors. Hence, we review the process of explaining from the point of view of social sciences. Here, our intention is not to make an exhaustive list of all the discussions around this subject, we summarize the few elements that reach a consensus and that we believe are relevant for the continuation of our work. We discuss in turn below the context of an explanation and its key dimensions.

2.1.1 | Context of an explanation

An explanation can be defined as "an answer to a why-question" (Lewis, 1986; Dennett, 1989; Overton, 2011; Miller, 2019) and should provide a reason that justifies what happens (Lipton, 1990; Dennett, 2017). For example, a student can ask a teacher why he/she received a given grade. In such a case, the need for an explanation may stem from a lack of understanding or a feeling of injustice about the received grade. In the literature, there are many reasons to justify such need, and how the latter is formulated. There are different points of view that overlap but do not completely align. Here, the objective is not to identify their complementarities but just to point out their diversity.

Following the discussion of Keil (2006), people are said to be driven to acquire explanations to apprehend the world (e.g., few months after their first words, children start asking "why" to understand their environment) and to provide them in an attempt to communicate an understanding (e.g., a friend explaining why he/she has failed to honor a commitment). Keil (2006) argues that people tend to ask for an explanation to improve their understanding of someone or something, so they can build a mental model that can be used to make an informed decision. Building a mental model refers to the process to constructing hypothethical knowledge (Carroll and Olson, 1988; Wickens et al., 2015) about a specific event, so as to better under-

stand it (Norman et al., 1983). More precisely, Malle (2006) argues that people tend to ask for explanations for two reasons. On one hand, explanations are requested to find meaning when there are contradictions or inconsistencies with prior knowledge. It is argued that they are most needed for events or observations that are perceived as being abnormal or unexpected (Hesslow, 1988; Hilton, 1996), or when the explanation may lie in a new field (Doshi-Velez and Kim, 2017). On the other hand, explanations are needed to create a shared meaning of something, adapt others' beliefs and impressions. When applied to Machine Learning, Adadi and Berrada (2018) define at least four sub-reasons for the need for explanations: to justify a predicted outcome, to control the model (e.g., to mitigate potential flaws or mistakes), to improve the model or to discover new facts.

Miller (2019) proposes to distinguish between four classes of *"why-questions"* that can be answered with explanations, called "whether-questions", based on the Ladder of Causation of Pearl and Mackenzie (2018). First, the **what-questions** (e.g., *"What event happened?"*) seek to determine, from some observed events, which unobserved events could have happened as well. In addition, there are **how-questions** (e.g., *"How did the event happen?"*) and **what if-questions** (e.g., *"What event would have happened if this was different?"*), that seek to determine the set of causes that would prevent an event from happening. Finally, **why-questions** (e.g., *"Why did the event happen?"*) seek for causes that can be used to simulate alternative causes to see if a factual event would still happen.

People tend to explain the cause of an event relative to some other event that did not occur (e.g., "Why P rather than Q?") (Hilton, 1990). P refers to the event that occurred, called the fact, and Q refers to the hypothetical case that was expected but did not occur, called the foil (Lipton, 1990). Hence, explanations are requested for a reason, and often a why-question implies an underlying hypothesis. Offering an explanation requires to identify this underlying whether-question it should answer (Bromberger, 1966). According to Van Bouwel and Weber (2002), there are three types of whether-questions: on properties within an object (e.g., "Why does object a have property P rather than property Q?"), between objects themselves (e.g., "Why does object a have property P while object b has property Q?"), and within an object over time (e.g., "Why does object a have property P at time t, but property Q at time t'?").

2.1.2 | Key dimensions of an explanation

In social sciences, there have been many rich discussions over the years around the concept of an explanation (see Lewis (1973); Lipton (1990); Malle (2006) to name a

few), and its various definitions seem to be used interchangeably between authors, often demonstrating conflation of some related terms (e.g., an explanation can be defined as information about the cause of an event; the concept of causality has brought many discussions in the literature (Hume, 2000; Lewis, 1986; Hilton, 1990)). The objective of this section is not to offer an exhaustive overview of such discussions, but to present some key insights of the surveyed literature made by Miller (2019). In particular, there are four important dimensions to consider when defining explanations: they are **contrastive** as they are acting as an answer to a why-question in the form of "Why event P happened instead of event Q"; they are **selected** in a biased manner among series of possible causes; they are **social** as they are part of a conversation or interaction; and they tend to be **causal**. Miller (2019) also argues that all these dimensions convey a **contextual** nature of the explanations. We describe in turn below these points.

Explanation are contrastive Explanations are said to be contrastive by nature, as people tend to explain the cause of an event relative to some other event that did not occur (e.g., “Why P rather than Q?”) (Hilton, 1990) which requires to identify the underlying whether-question it should answer. As presented in the previous section, most whether-questions require contrastive explanations, the foil being explicitly formulated (e.g., "Why did you open the door rather than the window?") or implicitly formulated (e.g., "Why did you open the door?") (Hilton and Slugoski, 1986; Hesslow, 1988; Lipton, 1990; McGill and Klein, 1993). In the case of an implicit whether-question, the explainer (i.e., the person who receives the question and provides the explanation) must understand the foil, as there can be many depending on the context the question is asked (Hesslow, 1988; Lipton, 1990): in the same question "Why did you open the door?", the foil could be either "rather than the windows" or "rather than turning on the air conditioning", and they do not seek for the same nature of answer in order to explain both the fact and the foil (Miller, 2019). Thus, an explanation necessarily depends on the context behind the foil, so that it can meet the needs of the one requiring the explanation (Hoffman et al., 2018).

Explanations are selected Explanations are usually requested when people do not have the necessary elements to understand an event. Hence, these people usually do not have access to the causes, and infer them from contextual observations and prior knowledge (Malle, 2006). People select some of these inferred causes as the explanation, based on the goal of the explanation (Mill and Robson, 1973). The explanations

are said to be selected and Miller (2019) argue that there are three cognitive processes in use to select the causes.

First, the causal connection process is used to identify the causes of an event. There are two sub-cognitive processes at play for the causal connection: abductive reasoning refers to hypotheses that people formulate and test to infer causes of an event; and simulation refers to counterfactual events that people simulate to draw a good explanation. Then, the explanation selection process is used to retain a small subset of identified causes as explanation. It is argued that people tend to select what they believe are the most relevant causes (Trabasso and Bartolone, 2003). Various criteria for selection are considered, such as abnormality (Hilton and Slugoski, 1986), intentionality (Hilton and John, 2007) or necessity (Woodward, 2006) to name a few. Finally, the explanation evaluation process is used to assess the quality of an explanation. Different criteria can be used to assess the explanations: the probability of the explanation of being true (Josephson and Josephson, 1996), simplicity (Read and Marcus-Newhall, 1993) and coherence with prior beliefs (Thagard, 1989) for example.

Explaining is a social process Once the explanation is selected and evaluated, people communicate it. Hilton (1990) says that explaining is a social process where "someone explains something to someone". Thus, explanations is an instantiation of a conversation with a dialog between two parts. There are two stages for communicating these explanations: first the diagnosis stage, in which an explanation is selected; and then the explanation stage, which is the social process of conveying this to someone. Hilton (1990) argues it is important for this explanation to be relevant to the question asked so that it is considered a good one (i.e., that it answers the appropriate whether-question). As such, the explanation should follow the rules of *Grice's maxims of conversation* (Grice, 1975), namely the quality (i.e., making sure that the piece of information is of high quality), the quantity (i.e., providing the right quantity of information), the relation (i.e., only providing information that is related to the conversation) and the manner (i.e., being intelligible: avoid obscurity of expression, ambiguity, be brief and orderly).

Explanations are causal In the discussion proposed by Miller (2019), it is finally argued that an explanation is less about probabilities and more about causes. Explanations should provide a reason that justifies what happens (Dennett, 1989, 2017; Biran and Cotton, 2017; Antaki and Leudar, 1992). In philosophy, the causal dimension is an important topic, and it is embedded in many definitions of an explanation (Kelley, 1987; Lewis, 1986; Josephson and Josephson, 1996; Hume, 2000). Lewis (1986)

says that "[explaining] an event is to provide some information about its causal history". Josephson and Josephson (1996) also say that "an explanation is an assignment of causal responsability". Moreover, Hume (2000) assimilates an explanation to a counterfactual: there is a causal relation between two types of events, if events of one type are always followed by events of the other type (Lewis, 1973; Hilton, 1988). Finally, Halpern and Pearl (2005) propose another formalization of causality: the world is assumed to be characterized by the values of two kinds of variables (exogenous variables, whose values are external to the model; and endogenous variables, whose values are determined the exogenous variables) and some variables may have causal influence on others.

2.2 | Explaining machine learning models

As presented in Chapter 1, explaining machine learning models is a challenging task. As models are increasingly more complex and sophisticated, it is more difficult, if not impossible, for people to interpret them. This has encouraged a surge for research on explainable AI (XAI) over the years (Lipton, 2016; Wachter et al., 2017). Research in XAI has received a significant amount of scientific attention in the last decade (see Barredo Arrieta et al. (2020) for recent survey). Generally, XAI works refer to all the initiatives aiming at making ML models understandable to human (Adadi and Berrada, 2018). Various terms have been used in the literature to describe contributions in XAI (such as interpretability, explainability, transparency, intelligibility to name a few), and researchers often disagree on their scopes and relations (Liao and Varshney, 2021). This thesis uses the terms of explainability, interpretability and intelligibility interchangeably. Many techniques have been proposed for that purpose (see Adadi and Berrada (2018); Mohseni et al. (2018); Guidotti et al. (2018); Barredo Arrieta et al. (2020) for some surveys) to meet the growing needs for more transparency over algorithmic decisions of the users and regulators.

In Section 2.2.1 we introduce the considered issues for explaining ML models, and present different approaches to make such models more transparent in Section 2.2.2 and different types of explanations that can be generated in Section 2.2.3. We then discuss the current research direction towards human-centered XAI in Section 2.2.4.

2.2.1 | Considered issues

The considered case for explainability in this thesis refers to ML models mainly for tabular data, as illustrated in Figure 2.1. The baseline system to explain is composed

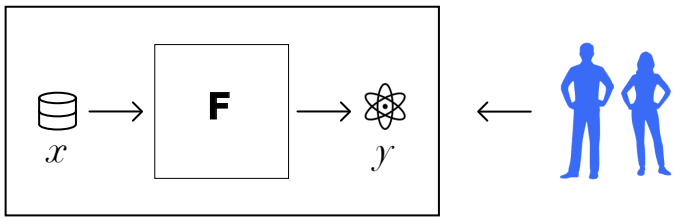


Figure 2.1: ML model for regression or classification tasks. For instance x, model F that has been trained beforehand, provides prediction y. Users interact with this system.

of a model F that is providing a prediction y for instance x.

ML models can be trained for various ML tasks such as classification (i.e., y is a class), regression (i.e., y is a numerical value), clustering (i.e., input data in x are not labeled and y labels this input data), recommendations (i.e., y is a rank/score) to name a few. This thesis focuses on the case of supervised learning tasks, classification and regression.

Also, there can be various types of data needed by the ML models to complete such tasks: tabular features (with categorical and numerical values), text and image, as well as their combinations. This thesis considers ML models trained with tabular data.

In particular, we are interested in the prediction phase and consider that the model has been trained before. To illustrate this baseline system, the use case of a loan application in insurance is considered. In such a system, a ML model F (depicted as a box in Figure 2.1) is trained for a binary classification task: y (depicted as an atom icon) refers to the class predicted for the loan (either accept or reject the loan applicaton); x (depicted as a database icon) refers to the values of the instance, in the form of features that can numerical (e.g., "Salary", "Age") and/or categorical (e.g., "Current position", "City of residence"). If the loan application x is rejected by model F, the user will likely be needing explanations to understand such decision and system.

Explaining Machine Learning models is a challenging task because they often are exponentially more complex at the expense of their transparency. Due to this increased complexity, users are not able to understand the algorithmic processes nor the predicted outcomes. Yet, as presented in Section 2.1, people need explanations in situation where they are dealing with novelty or for surprising outcomes. This need is reinforced in situation with high impacts (e.g., in health), where it is important to

ensure that the prediction is reliable for the users to be using it (e.g., no incorrect prediction (Dietvorst et al., 2015), no hidden biases from the dataset (Larson et al., 2016). The considered issues in this thesis are related to the explainability of such complex models, called *"black-boxes"*.

2.2.2 | Some characteristics of ML explanations

To make ML models more transparent, numerous approaches have been proposed (see e.g., Biran and Cotton (2017); Guidotti et al. (2018); Molnar (2020); Barredo Arrieta et al. (2020); Zhou et al. (2021) for some surveys). These technical approaches generally allow to (i) extract the relevant information to explain a prediction from black-box ML models and (ii) translate this into an explanation that users can understand. These approaches aim at gaining the users' trust, and providing them with the causes of some events. Thus, they can be considered as key features to add to a ML system, so that it is possible to generate relevant explanations to the users. Among possible distinctions between such approaches, we consider the following two:

- **Self explaining vs *post-hoc* approaches:** the complexity of the model to explain allows to distinguish between two types of explaining approaches. The first one consists in building a self-explaining model , that is intrinsically explainable (e.g., low depth decision trees). For complex ML models, called "black-boxes", post-hoc approaches can be used to extract and generate various kinds of information that act as explanations. These approaches can be model-agnostic (i.e., they are independent from the prediction model) or model-specific (i.e., they only work for the interpretation of specific model like neural networks). This allows for more complexity in the ML model and flexibility in the settings of the ML system (e.g., the model may be modified or retrained without modifying the explainer).

- **Global vs Local:** Another important distinction can be made between global and local explanations that can be generated. Global approaches aim at explaining the model's behavior, whereas local approaches focus on explaining a single prediction. These two approaches allow to answer different types of questions: global approaches answer questions on the ML system itself such as "What is the system's overall rationale"; local approaches answer questions on the prediction such as "How should this instance change to get a different prediction Q?" for example Liao et al. (2020).

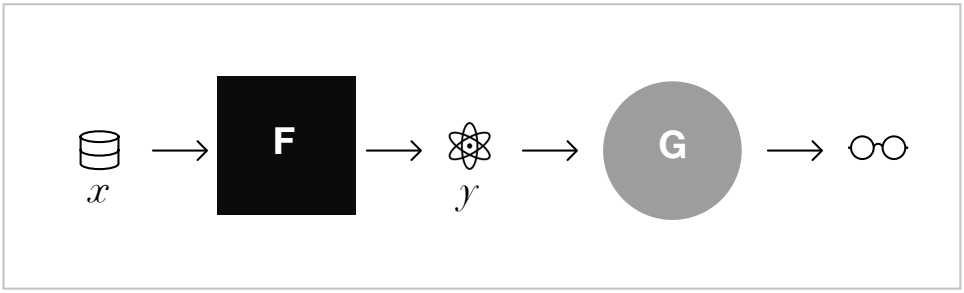


Figure 2.2: ML system to explain complex ML models: F is the pre-trained complex ML model; G is the post-hoc explainer; F and G combined represent the ML system. From left to right: for instance x, model F provides prediction y; explainer G generates local explanations (for y) or global explanation (for model F).

The considered models for this thesis are complex ones for regression and classification tasks. They are considered as opaque and not interpretable. We call ML system the combination of such a model and a post-hoc approach, called the explainer, as illustrated in Figure 2.2. The model F is giving a prediction y for instance x, and the explainer G (depicted as a disk) provides an explanation (depicted as glasses icon). This explanation can be local (i.e., explains the prediction) or global (i.e., explains the model's behavior).

2.2.3 | Some types of ML explanations

Different types of explanations can be generated for a ML system, and explainer G can be categorized into three types of outputs. First, **feature importance** refers to techniques to assign a score to input features based on how useful they are at predicting an output (e.g., LIME (Ribeiro et al., 2016) or SHAP method (Lundberg and Lee, 2017). Second, **rules** are defined as knowledge bases that collectively make up the prediction model (e.g., RuleMatrix (Ming et al., 2018) and Anchors (Ribeiro et al., 2018)). Finally, **counterfactual** instances, that exemplify the minimal modifications that would lead to a different prediction (e.g., Growing spheres (Laugel et al., 2018a), FACE (Poyiadzi et al., 2020) and DiCE (Mothilal et al., 2020)). The considered explanation types for this thesis are feature importance scores and counterfactual examples, and we describe these XAI approaches below.

Feature importance techniques assign scores to input features based on how useful they are at predicting an output. The higher the scores are, the more impact the

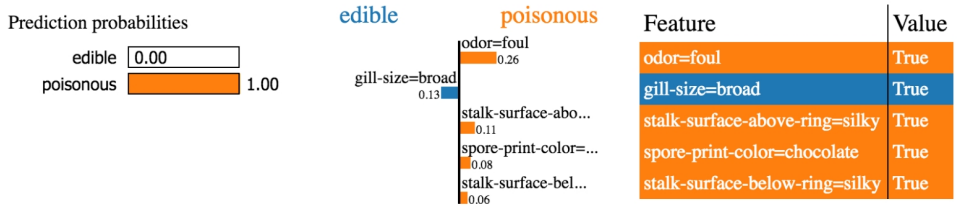


Figure 2.3: LIME (Local Interpretable Model-agnostic Explanations) explanations for a classification ML model for tabular data (figure reproduced from Ribeiro et al. (2016)).

corresponding features have on the output.

These explanations are displayed into graphs that show the weights of features. Based on the weights in the linear model (i.e., coefficients of the relationship between a given feature x and a target y, assuming that all the other features remain constant), the most important or influential features are placed at the top. Feature importance explanations can be local, i.e., they provide a weight for each feature describing its contribution to the final decision for a specific instance, or they can be global, i.e., they describe the weights of the features used by the model.

Several techniques have been proposed to generate this type of explanations, such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), randomization (Henelius et al., 2014) and pairwise interaction (Lou et al., 2013) to name a few. In particular, we detail below two of the most popular ones: LIME and SHAP.

- LIME stands for Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016). It is a local model that explains the prediction of an instance by analyzing its neighborhood in the dataset. It creates a transparent model that acts as a substitute model to explain the predictions of a black-box model. First, new samples are created by perturbing the instance for which an explanation is required. These new perturbed samples are run through the black-box model, and the changes in the prediction with respect to the original instance are measured, weighted by the proximity to the original instance. Based on this, a transparent model is trained to fit the predictions of the black-box model. The output of LIME for tabular data is a list of input features with importance scores, and features are displayed in a decreasing order of importance.

  As an illustration of how such an approach works, the example used by Ribeiro et al. (2016) is represented in Figure 2.3. In this example, a classifier predicts the editability of a mushroom, using only categorical features (here a binary clas-

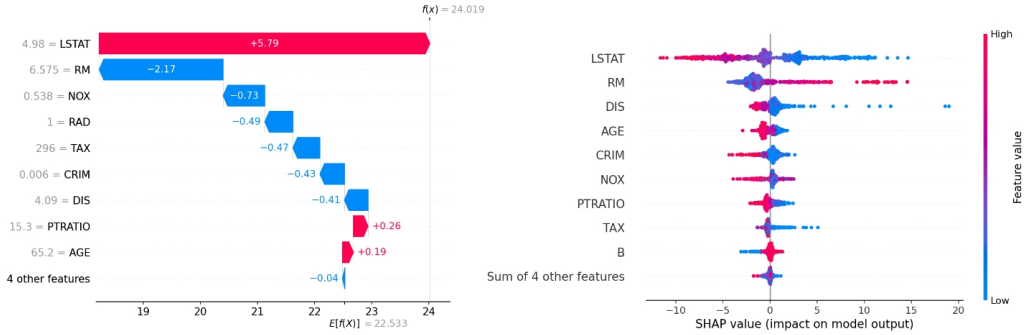


Figure 2.4: SHAP (SHapley Additive exPlanations) local (left) and global (right) explanations for a regression ML model for tabular data from (figure reproduced from Lundberg and Lee (2017)).

sification task for whether the mushroom is edible or poisonous). The left part of the illustration displays the prediction probabilities; the middle part displays a graph presenting the weight of each feature towards the predicted class in a decreasing order of importance; the right part display the initial value for each feature. For the mushroom considered here, the feature "odor", whose value "foul" is "true", has the most impact on the predicted class "poisonous".

- Another popular technique for generating feature importance is SHAP, standing for SHapley Additive exPlanations (Lundberg and Lee, 2017). Inspired by game theory, this method generates either local or global feature importance explanations by calculating Shapley values for inputs that were used in the ML model. These values define how much features contributed to the value of a given prediction, in comparison to the average prediction. They are obtained by computing the average difference in the value of predictions when including and omitting a certain feature value in increasingly large sets of other features.

  As an illustration of how such an approach works, the example used by Lundberg and Lee (2017) is reproduced in Figure 2.4. In this example, a regression model predicts house prices in Boston area, and the SHAP method provides explanations for these prices. The plot on the left shows features each contributing to pushing the model output from the base value. Features pushing the prediction higher are shown in red, those pushing the prediction lower are in blue. The plot on the right sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values to show the distribution of the impacts each feature has on the model output.

Counterfactual examples explain a predicted outcome y by identifying minimal changes that could be applied to the initial instance x so as to get another prediction. For instance for the loan application case, if the model F rejects the loan application, a counterfactual example provided by the explainer G explains what minimal changes on x would have made the model accept the loan instead.

Counterfactual examples are argued to be highly relevant forms of explanations based on arguments from cognitive sciences regarding their resemblance to human explanations (Wachter et al., 2017; Byrne and Tasso, 2019; Wang et al., 2019; Zhang and Lim, 2022). First, counterfactual examples possess the key property to be contrastive (Miller, 2019; Chromik et al., 2021; Zhang and Lim, 2022): they allow to answer to questions such as "Why Q rather than P?". It is argued that they are much more causally informative than factual explanations (e.g., feature importance approaches) (Byrne and Tasso, 2019; Warren et al., 2022), which is another important component of an explanation as described in Section 2.1.

There is a large variety of approaches to generate such explanations (see e.g., Guidotti et al. (2018); Verma et al. (2022) for recent surveys). In the survey proposed by Verma et al. (2022), 55 counterfactual approaches are compared for different ML system settings (i.e., these approaches can access to the model's internals or the model is a "black-box"; and they can be model-agnostic or specific). These approaches may differ on several other dimensions.

First, they can differ regarding the definition of counterfactual attribute such as sparsity, data manifold adherence, causality. For example, sparsity can be defined as minimizing the distance, for which various definitions can be considered, such as the Euclidean distance, the L1 norm (Wachter et al., 2017) or combinations with other norms so as to improve the change sparsity, in terms of the amount of modified features (Lash et al., 2017; Laugel et al., 2019; Le et al., 2020).

Moreover, counterfactual approaches may thus differ regarding the optimization problem, and how they can integrate constraints in the formulation of this cost function: for example, depending on how important one criterion may be for the relevance of the explanation to the users, algorithms should allow to generate only plausible (Looveren and Klaise, 2021) (e.g., with realistic suggested changes) or feasible (Poyiadzi et al., 2020) counterfactual examples (e.g., with actionable features when the users are looking to optimize the prediction), adding causal reasoning (Karimi et al., 2022), or users' preferences (Lash et al., 2017; Jeyasothy et al., 2022) (e.g., when expert users want to test their hypotheses).

Finally, recent works on counterfactual explanations propose approaches to generate plural counterfactual examples, and argue that having several examples can

Query instance (original outcome : 0)

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	22.0	Private	HS-grad	Single	Service	White	Female	45.0	0.01904

Diverse Counterfactual set (new outcome : 1)

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	70.0	-	Masters	-	White-Collar	-	-	51.0	0.534
1	-	Self-Employed	Doctorate	Married	-	-	-	-	0.861
2	47.0	-	-	Married	-	-	-	-	0.589
3	36.0	-	Prof-school	Married	-	-	-	62.0	0.937

Figure 2.5: DiCE (Diverse Counterfactual Examples) with 4 counterfactual examples (bottom) with sets of changes on diverse features for the model to predict "High Income" class instead of "Low Income" one for the instance query (top) (figure reproduced from Mothilal et al. (2020)).

help users to better interpret them (Ekstrand et al., 2014; Kunaver and Porl, 2017; Mothilal et al., 2020; Laugel et al., 2023). Indeed, a single counterfactual could be misleading as it could suggest changes that are not feasible or plausible for example (Dandl et al., 2020). Similarly to the methods described previously, these methods propose different approaches to generate a set of multiple counterfactual instances. For example, Mahajan et al. (2019) propose to integrate all the constraints defined above, so as to avoid requiring to choose between them. Other approaches emphasize the importance of generating a set of diverse counterfactual examples: diversity in terms of optimized metrics (Dandl et al., 2020) and diversity in the feature space (e.g., DICE by Mothilal et al. (2020) as illustrated in Figure 2.5).

2.2.4 | Towards Human-Centered eXplainable AI

The research community in Human-Centered XAI has recently emerged and considers more fundamental questions regarding ML explanations. In particular, the place of the user is considered to be central for the design of XAI approaches (Liao and Varshney, 2021; Chromik and Butz, 2021; Ehsan et al., 2022; Szymanski et al., 2022b). The different types of explanations generated by XAI approaches we presented in the previous Section 2.2.3 are not necessarily equally understood by the end-users, especially when they may lack literacy in AI. Making these explanations intelligible and useful for these users remains a challenging task (Dodge et al., 2019; Feng and Boyd-Graber, 2019; Schaffer et al., 2019). There is thus a call for more interdisciplinary

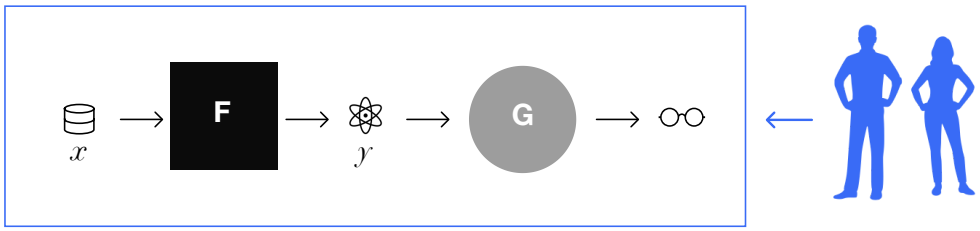


Figure 2.6: Human-Centered explainable AI: explaining the Machine Learning System (left) to the explainees (right).

works, in particular from social sciences and human-computer interaction fields, so as to have a user-centric approach when designing XAI approaches (Liao and Varshney, 2021; Ehsan et al., 2022). We present the current research direction to make intelligible explanations for the end-users in Section 2.2.4.1. We then discuss the notion of "explainees" as the users who receive the explanations in Section 2.2.4.2.

2.2.4.1 | Human factor in explainability

ML explanations should answer questions the users may have on such ML system. Thus, the human component is an inherently important factor for the quality of such explanations: knowing and understanding the users' needs is crucial to extract the relevant information from the model, and present it in an intelligible and useful manner for them.

We identify two parts in the explanation process which allows to refocus the user in this context :

- The *ML system* represents the set composed of the predictive model and the explainer (i.e., an XAI approach). Thus, the ML system is the provider of both the predicted outcome and the explanations.

- The *explainees* represent the users who interact with this system (i.e., the outputs of both the predictive model and the explainer). Thus, when provided to the end-users, the explanation should allow them to understand the prediction and/or the ML model. In this work, the users we consider are humans with various needs, expertise and prior knowledge, discussed below.

We illustrate this explanation process in Figure 2.6. We add the "user" component next to the ML system we consider in this thesis (see Figure 2.2). Although users are

placed at the end of this process, the ML system should be designed to answer their needs.

2.2.4.2 | The explainees

When interacting with a ML system, the explainees refer to the users that receive the explanations for such system. There can be an important diversity of motivations among these users in the real world (Chromik et al., 2021).

A classical approach consists in classifying the users depending on their expertise level/domain, with distinct needs and goals regarding the provided explanations (Mohseni et al., 2018; Bhatt et al., 2020):

- **AI practitioner** (e.g., ML engineers, data scientists, researchers): they are the designers of the ML system, and they are at least knowledgeable about ML. They can use explanations to understand how the trained model works. For example, data scientists can use the explanations to detect potential unwanted biases or to debug what the model has learned.

- **Domain agent** (e.g., regulators, domain experts): they are users of the ML systems and refer to the various stakeholders that at least knowledgeable about the involved ML application domain. They aim at understanding the model's behavior and the predictions, so that they can make an informed decision. For example, a fraud agent can use explanations to understand patterns of fraud in submitted claims to better detect fraudulent claims (Collaris et al., 2018).

- **Non-expert user:** They are also users of the ML system, but have little to no knowledge about ML nor the application domain. It has been established (Liao et al., 2020) that such users are in general more interested in understanding the rationale behind a specific prediction, rather than the overall rationale of a model. For example, an insurance customer whose loan application got rejected would be using explanations to understand the reasons for this reject or what to change in order to have it accepted.

As these users can have various motivations, it is important to understand also the range of underlying questions they may be asking when requesting for an explanation, as presented in Section 2.1. Liao et al. (2020) propose an XAI question bank based on a rich user study, that should help AI practitioners addressing the needs of their users. This bank is made of the 50 most common questions that various types of users (e.g., data scientists, customers, agents) have when interacting with an AI. There

are different categories of questions and we list some of them here: these questions can be oriented towards the design and performance of the ML system (e.g., "What kind of output does the system give" or "How accurate/precise/reliable are the predictions"?); on the model itself (e.g., "How does the system make predictions?"); on the predicted outcome (e.g., "Why/how is the instance given this prediction"); on contrastive outcomes (e.g., "What would the system predict if this instance changes to...?). Moreover, the answers could be different for the same question, depending on who is asking (Arya et al., 2019). For example, users can be differentiated depending on their prior knowledge (e.g., a doctor has expert knowledge that a patient does not have) and intention (e.g., data scientists need to monitor a model; a prospective client needs to optimize the output). Thus, the presentation of ML explanations should be adapted to the end-users and their needs, so that they are useful and help to make an informed decision.

In this thesis, we believe that providing the non-expert users with intelligible explanations is an important challenge, as they are the ones who might struggle the most to interpret them.

2.3 | From XAI to XUIs

In human-centered XAI, some contributions in HCI investigate how to design interfaces for different types of explanations in order to answer the needs of various users. These interfaces called eXplanation User Interfaces (XUIs) are defined as the sum of outputs of an XAI system that the user can directly interact with (Chromik and Butz, 2021), as illustrated in Figure 2.7. The XUI (depicted with blue box) is a part of the ML system, and act as the interactive (depicted with the blue double arrow) interface between the latter and the users. Several XUIs have recently been proposed for explaining to a user a specific prediction of a ML model: some interfaces display the information provided by a given explainer (e.g., the SHAP method by Lundberg and Lee (2017) is displaying local feature importance as explanations into plots); other interfaces allow to display interactive explanations of various types (e.g., ViCE tools by Gomez et al. (2020) for visual counterfactual explanations) and in various formats (e.g., explanation vocal interface proposed by Sokol and Flach (2018)). We present in turn below different directions for designing such XUIs: depending on the users' expertise level in Section 2.3.1 and the type of explanations interfaced in Section 2.3.2.

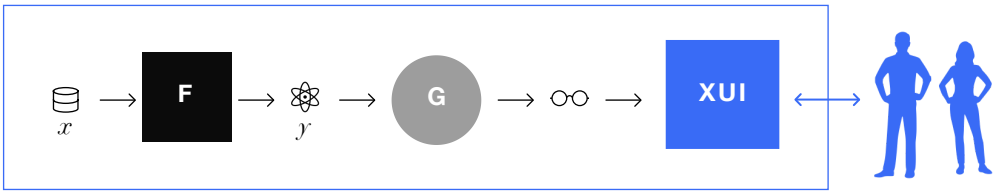


Figure 2.7: Explanation User Interface (XUI) in the explanation process

2.3.1 | XUIs for various expertise levels

As presented in Section 2.2.4.2, there are three main types of users that are usually considered regarding explanations: AI practitioners, domain expert and non-expert users (Mohseni et al., 2018).

The majority of current XUIs address the needs of AI practioners to better understand ML models. Many of them propose interactive visualization enhancements. For example, AutoAIViz (Weidele et al., 2020) is an experimental system for data scientists, that aims to visualize AutoAI's model generation process. Weidele et al. (2020) demonstrate that this interface helps users to complete the data science tasks, and increases their understanding in the AutoAI system. Another example is the What-if Tool proposed by Wexler et al. (2020). It allows AI practitioners to probe, visualize and analyze ML systems, with minimal coding. The What-If Tool lets practitioners test performance in hypothetical situations, analyze the importance of different data features and visualize the model behavior across multiple models and subsets of input data. It also lets practitioners measure systems according to multiple ML fairness metrics. iForest (Zhao et al., 2018) is a visual analytic system aiming at interpreting random forest models and predictions. It provides the AI practitioners with a summary of the decision paths of a random forest model, which reflects the working mechanism of the model and reduces the users' mental burden of interpretation. ViCE (Gomez et al., 2020) is another interactive visual analytics tool that generates counterfactual explanations to contextualize and evaluate model decisions.

Other XUIs are dedicated to users with advanced knowledge in the application domain. For instance in the medical domain, Wang et al. (2019) co-design with doctors an AI-driven medical diagnosis XUI with multiple explanations: feature value for time series, class attribution of predicted disease risk, feature attribution by vitals, and counterfactual rules indicating key rules for each prediction. Another example is RuleMatrix (Ming et al., 2018), an interactive visualization XUI to help domain experts with little expertise in machine learning to understand, explore and validate

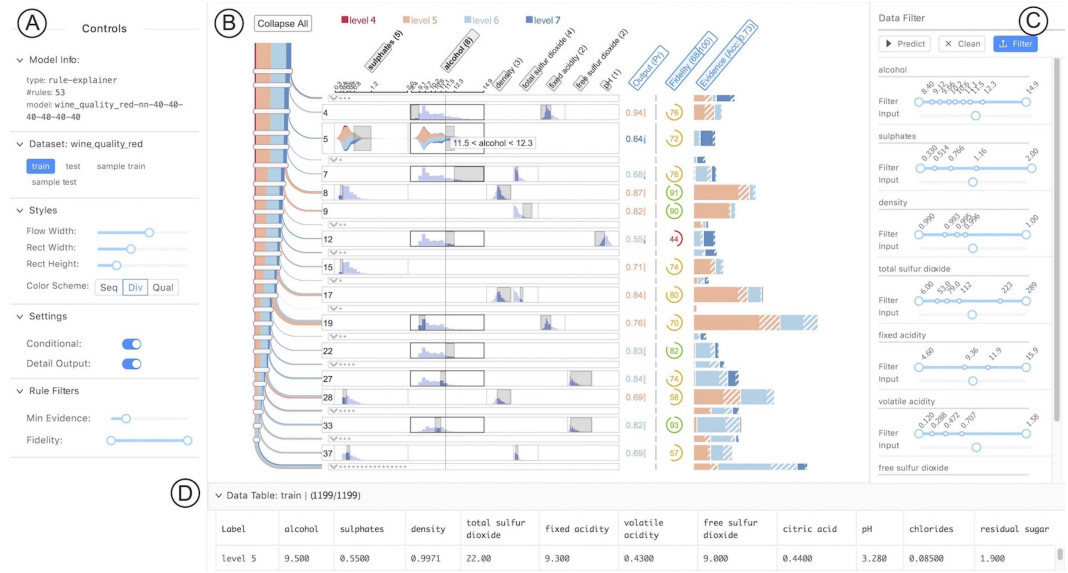


Figure 2.8: RuleMatrix (Ming et al., 2018): explanatory visual interface for non-AI expert users to understand the behavior of a trained neural network. (A) Control panel: users can specify the detail information to visualize (e.g., level of detail, rule filters); (B) Matrix: rule-based explanatory representation (row = rule; column = feature); (C) Data filter: users can filter or customize the input; (D) Data Table: users can filter the dataset.

classification models using rule-based explanation (see Figure 2.8).

Finally, few XUIs are designed for the needs of non-expert users. Their lack of knowledge in both ML and the applied domain may make it difficult to interpret explanations generated with current XAI approaches. In recent contributions, it has been demonstrated that various XAI design principles can improve the quality of these explanations for the non-expert users (Cai et al., 2019; Cheng et al., 2019; Yang et al., 2020; Ooge et al., 2023). For exemple, Cheng et al. (2019) explore interactive principles on XUIs (see Figure 2.9) and demonstrate that allowing users to modify the input values can improve their understanding of the ML system. Similarly, Ooge et al. (2023) propose a control mechanism and a visualization of its impact in an XUI for an e-learning recommendation system, which has been found useful and usable by adolescents.

2.3.2 | XUIs for different types of explanations

As presented in Section 2.2.3, there are several types of explanations associated with various types of interfaces, and we discuss in particular two of these below: XUIs with feature importance explanations first, and with counterfactual explanations then.

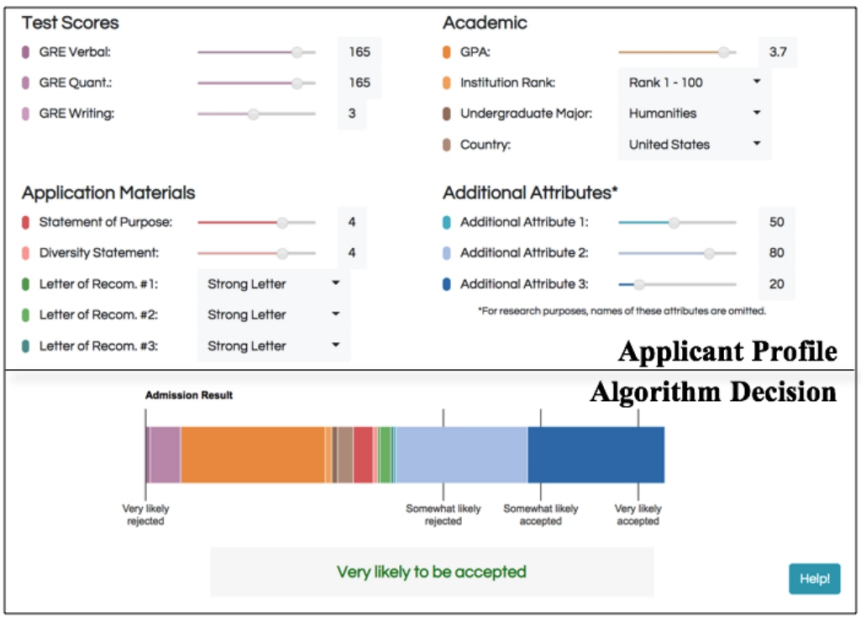


Figure 2.9: Interactive prototype proposed by Cheng et al. (2019) to communicate students how the university admission decision-making of the algorithm works. Top: Applicant profile with interactive sliders to modify the values of attributes. Bottom: colored bars to indicate the contribution of each feature and how they add up to reach the final decision.

Feature importance Among other, commonly used XAI approaches are SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016), as discussed in Section 2.2.3. Various XUIs have been proposed to improve the intelligibility and usefulness of such explanations. For example, in a dashboard for fraud detection (see Figure 2.10 by Collaris et al. (2018)), feature importance explanations are enhanced by adding other information (such as data distribution and other model explanations) needed in this context to enable fraud detection experts to effectively identify more potential fraud cases. Cheng et al. (2019) explore interactive principles on XUIs with feature importance explanations (see Figure 2.9). They demonstrate that allowing users to play with the input values while displaying feature importance weights can improve their understanding of the ML system.

This type of explanations is argued to be relevant for non-expert users. Szymanski et al. (2022a) compare six different versions of a conversational XUI for a pain-related health recommendation system: two different forms of text-based explanations, tags, word cloud, and two forms of feature importance explanations. The results of the qualitative user study demonstrate that the non-expert users ranked higher both versions of the XUI with feature importance explanations.

Figure 2.10: Feature dashboard (Collaris et al., 2018) for fraud agents. (A) bar chart for feature contribution to the target class. (B) partial dependence plots, showing the impact of changing the feature value (indicated with a vertical line) on the final prediction. (C) training data distribution as compared to current value (indicated with a vertical line).

Counterfactual explanations Very few XUIs propose to implement such explanations. For image data, it has been shown that having both normative explanations (i.e., examples that are similar from the prediction) and comparative explanations (i.e., examples that are different from the prediction) leads to better understanding of the underlying ML models (Cai et al., 2019). Similar works demonstrate that example-based explanations in general have a positive effect on users' trust in ML, regardless of their familiarity with it (Yang et al., 2020). In the case of structured data, using counterfactual examples as explanations has been explored in the ViCE tool (Gomez et al., 2020) but this proposition has not been evaluated experimentally.

Several forms of XUI have been proposed to present counterfactual explanations:

- Textual interface: for example in a user study conducted by Wang and Yin (2021), it is shown that textual counterfactual explanation increases users' objective understanding and satisfaction.
- Visual interface: for model validation, visual counterfactual explanations (ViCE) by Gomez et al. (2020) and aggregated visual counterfactual explanations (Ad-ViCE) by Gomez et al. (2021); for the understanding of ML model's performance across a wide range of inputs, the What-if tool by Wexler et al. (2020).
- Vocal interface: for example, Glass-Box Sokol and Flach (2018) is a voice-enabled device that provides class-contrastive counterfactual explanations when questioned by users for the understanding of automated decisions.

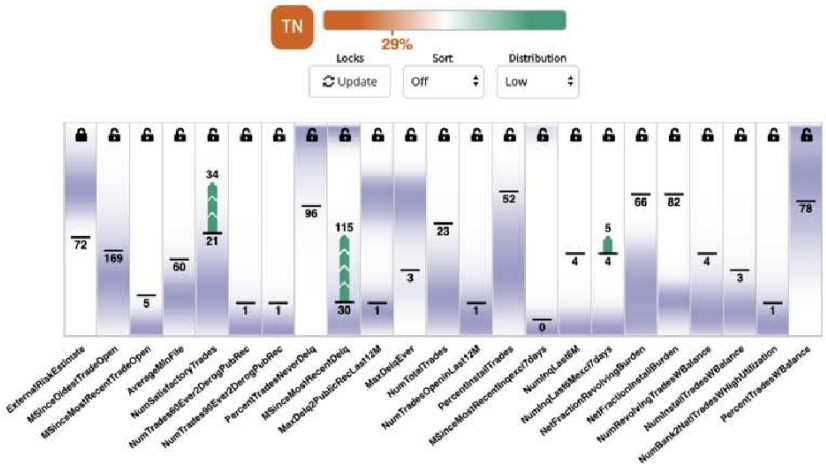


Figure 2.11: ViCE (Gomez et al., 2020): interactive and visual analytics tool that generates counterfactual explanations for Home Equity Loans predictions. Top: predicted class probability and control panel for the display of the features; Bottom: visual counterfactual explanations with numerical features, data distribution per feature, instance value, suggested changes for different classes in orange arrow and locker buttons to allow users to prevent changes on specific features.

2.4 | Evaluation in XAI

Most XAI approaches are proposed from a computational point of view, but lack empirical research in understanding users' needs of ML explanations in their usage. For example, the survey conducted by Keane et al. (2021) states that only 21% of counterfactual approaches have been user tested and evaluating them through user studies remains an important challenge to tackle (Verma et al., 2022; Shang et al., 2022). These concerns are also amplified by the challenge of defining the criteria that can be used to to measure the quality of an explanation (see Nauta et al. (2022); Rong et al. (2022b) for recent surveys on the variety of evaluation methods and criteria used in XAI). Despite the recent surge for contributions in XUIs as presented in Section 2.3, there is a lack of guidance over the method to apply and the measures to use when evaluating such explanations (Nunes and Jannach, 2017; Adadi and Berrada, 2018; Guidotti et al., 2018; Chromik and Schuessler, 2020).

We discuss first common methods to evaluate XAI approaches in Section 2.4.1. We then discuss the challenge of measuring the quality of an explanation in XAI and some current metrics that have been proposed in Section 2.4.2.

2.4.1 | Common methods to evaluate XAI approaches

To evaluate XAI approaches, several methods can be used depending on whether human subjects are involved or not. Most XAI approaches have not been user tested, meaning that the quality of an explanation is measured using computational proxy measures (Nauta et al., 2022). Other methods favor human involvement when evaluating such quality, in particular for XUIs. To conduct these kinds of evaluation, it is necessary that the participants recruited correspond to those who will use the interface, in a situation that is close to the real one. We discuss in turn below these different evaluation methods in Section 2.4.1.1, the participants' recruitement in Section 2.4.1.2 and the study design used for such evaluations in Section 2.4.1.3.

2.4.1.1 | Evaluation methods

A first distinction can be made between two types of evaluation methods (Doshi-Velez and Kim, 2017; Nauta et al., 2022): the functionally-grounded ones that do not involve human subjects, and the human subject ones that do involve them and that can be further decomposed into application-grounded and human-grounded evaluations. We describe these two strategies below:

Functionally-grounded evaluation is a strategy that does not need human experiments, but instead uses computational proxy measures (e.g., for explainability, model input can be perturbed in order to validate feature importance explanations (Nauta et al., 2022)). Some researchers argue that the functionally-grounded evaluation is time and cost saving (Markus et al., 2021), can be more easily scaled (Wang and Vasconcelos, 2020), and is more adapted for unethical or immature studies that would imply too much risks if there were users in the evaluation (Semuels et al., 2018; Hara et al., 2018). However, the difficult comparability between different automatic evaluation measures is a common problem (Tomsett et al., 2020; Rong et al., 2022a), and there is no guarantee that they truly reflect humans' preferences (Nguyen, 2018; Hase and Bansal, 2020).

Human subject evaluation defends the need to involve human subjects in the evaluation, in two ways: application-grounded evaluations involve domain experts within a real application (e.g., to verify if the model is aligned with human expert knowledge); human-grounded evaluations involve lay users on specific tasks to evaluate the quality of the explanation in general (Doshi-Velez and Kim, 2017). User studies in XAI can be crucial for the evaluation of the quality of ML explanations, especially when moving towards real-world products. We discuss current metrics for such evaluations in Section 2.4.2.

In this thesis, we adopt a user-centric approach in XAI and thus focus on human-grounded evaluation methods and metrics.

2.4.1.2 | Participant recruitment

In human subject evaluations, the recruiting method and number of participants are important criteria, and depend on the tasks that the latter have to complete. For example for application-grounded evaluation, domain expert participants are recruited from the field of study (e.g., doctors for a medical field), and it is considered to be more difficult because these users do not have much time and their compensations can be quite expensive for large scale studies (Rong et al., 2022b). For human-grounded evaluation, recruiting can be more flexible depending on the study design:

- The experiments can be run through a **crowd-sourcing** platform (e.g., MTurk [1]) where participants are directly recruited. Many experiments are performed on such platforms as they allow to recruit a large number of participants with minimum costs. Yet, the low financial compensations can affect the quality of the results of the experiment, as it may not motivate enough participants to complete tasks diligently or provide valuable feedbacks.

- The experiments can also be performed in an online setting (e.g., for questionnaires, using platform like Useberry or Qualtricx; for interviews, using Zoom or Teams and digital white board like Miro or Mural). It allows recruiting participants from professional and personal networks, with various demographics and geographies.

- Experiments can be performed in a lab setting: participants are recruited from the lab or academic networks. Although the financial contributions may be higher than for other experimental setups, it allows to monitor the study and to let participants asks questions when needed. It has been shown that the presence of a moderator increases participants' focus (Chromik et al., 2021). Studies in lab settings require more time to spend with participants, thus their number might be lower than with an online setting

In this thesis, we favor experiments performed in a lab settings to maximize the quality of the collected results.

[1] Amazon Mechanical Turk platform https://www.mturk.com/

2.4.1.3 | Study design

Another important dimension of human subject evaluation is the study design, which refers to two components: the approach used to evaluate an explanation (from an XAI approach with or without an XUI) and the experiment setup (i.e., what is evaluated/compared and how is it evaluated/compared).

First, there are three types of study approaches for the evaluation of an explanation in a study design, that depend on the data collected: qualitative evaluation (e.g., thematic analysis on results such as response to open-response questions), quantitative evaluation (e.g., statistics on scores in A/B testing setup) or a mixed approach, combining both quantitative and qualitative ones (Chromik and Schuessler, 2020; Vilone and Longo, 2021). Quantitative evaluations are particularly valuable when there is a large number of participants. Qualitative evaluations require more time for the analysis and are more suitable when the number of participants is relatively low. Recently, it has been demonstrated that performing both quantitative and qualitative analyses in XAI evaluation helps assessing user perceptions of the quality of the explanations (Meske and Bunde, 2022).

Another dimension of the study design is related to the experiment setup. Nunes and Jannach (2017) distinguish between four types of experimental combinations: single group (i.e., no alternative group), with and without explanations (i.e., no explanation is the alternative group), alternative explanations (i.e., varying information provided in explanations between groups with other aspects of user interface fixed) and alternative explanation interface (i.e., varying user interfaces between groups). The choice of the research question define the experimental condition and study approach. Moreover, there is a distinction between two experimental assignments (Chromik and Schuessler, 2020): between-subjects studies evaluate the difference between groups of participants; within-subjects studies evaluate differences within individual participants who are assigned to multiple treatments.

2.4.2 | The challenge of measuring the quality of an explanation

Although evaluation methods seem to be converging towards increasing user involvements, there is still much debate about the tasks for such evaluations and the measures used to assess the quality of the explanations (see Chromik and Schuessler (2020); Rong et al. (2022b) for recent surveys). We discuss these challenges regarding the evaluation task in Section 2.4.2.1 and the measures in Section 2.4.2.2.

2.4.2.1 | Overview of the current tasks for human subject evaluations

In the literature, there are discussions about additional elements involved in the development of study design for human subject evaluations. For example, Chromik and Schuessler (2020) propose a taxonomy of XAI evaluation involving humans, describing relevant dimensions of human subject evaluations in the context of an interaction with *"black-box"* models. In particular, there are three additional dimensions to consider and we discuss them in turn below: the level of task abstraction, the type of human involvement and the level of user tasks.

First, two levels of task abstraction can be distinguished (Chromik and Schuessler, 2020), which reflect the two types of human subject evaluation methods described in Section 2.4.1.1. For application-grounded evaluations, the task to complete usually requires a high level of participant expertise in real application (e.g., an AI-based diagnosis tool for doctors (Wang et al., 2019)). The XAI method can be evaluated by the intended users with respect to a particular task. In such evaluation methods, domain experts experiment with either the exact application task, or with a simpler/partial one, verify that the system succeeds in delivering the intended explanation. Another level refers to simpler task in human-grounded evaluations (e.g., to compare which type of explanations is better understood by users in a given context). These evaluations can be completed by non-expert users, and they are in general less restrictive and expensive than when having to recruit domain experts. The nature of the ML application to evaluate defines the level of task abstraction, helping to target the right users for the study.

Another distinction is made between two types of human involvement (Mohseni et al., 2018). First, participants can provide feedbacks on an actual explanation. These feedbacks are used to determine its quality. In another setup with no explanation provided to the participants, the latter can generate examples of reasonable explanations based on what they observe. In this feed-forward settings, these examples can be used as a benchmark for algorithmic prediction. The goal of the evaluation (whether it is for building the right explanations or evaluating if the explanations is right) is thus important to consider in order to define the right type of human involvement.

Finally, multiple user tasks have been proposed for XAI evaluation and different levels to elicit the quality of explanations based on the two taxonomies proposed by Chromik and Schuessler (2020) and Rong et al. (2022b):

- Verification task: participants rate their satisfaction with the explanations.
- Force choice task: participants choose from multiple competing explanations.

- Forward simulation task: participants are evaluated on their ability to predict the system's output based on the given explanations.
- Manipulation or counterfactual simulation task: participants predict what input changes are necessary to obtain an alternative output.
- Marginal effect queries: participants are evaluated on their ability to predict how changes in a given input feature will affect the prediction.
- Feature importance queries: participant are evaluated on their understanding of features weight on a given outcome.
- "Clever Hans" detection or failure prediction task: participants are evaluated on their ability to perceive where the model fails and debug these flaws.
- System usage task: participants are asked to use the system and its explanations for its primary purpose.
- Annotation task: participants provide a suitable explanation based on a given input and output.

2.4.2.2 | Overview of current objective and subjective metrics

When measuring the quality of an explanation, many metrics have been used in recent user studies (see Hoffman et al. (2018); Rong et al. (2022b) for recent surveys on XAI evaluation), depending on the explanation goal and method used to evaluate the explanations.

There are subjective metrics that rely on human judgment (e.g., the perceived understanding of an explanation (Cheng et al., 2019)). In quantitative analysis, these metrics are measured using a Likert scale, a widely used approach in research to scaling responses in a survey or questionnaire (e.g., satisfaction scale by Hoffman et al. (2018); questionnaire for usefulness of counterfactual explanations for recommendations by Shang et al. (2022)). The most common choices include between 4 and 7-point Likert scales. Simms et al. (2019) demonstrate that even numbers, i.e., 4-point and 6-point, are a reasonable format for psychological studies when there is a need to avoid having neutral scores. In qualitative evaluation, these metrics can be measured by performing thematic analysis on answers to open-response questions, interviews, observations (Clarke et al., 2015).

Because the metrics are subjective, they often report great variability in answers or observations.

Also, there are objective metrics that are impartial and quantifiable. They are thus independent of personal judgments and rely on factual data (e.g., score or time spent for task completion in quantitative study (Yang et al., 2020)).

We define below multiple metrics that have been used in recent user studies:

- **Trust** refers to the extent to which users know when to trust or distrust the model's recommendations (Zhang et al., 2020) and this notion is key for decision-making (Mohseni et al., 2018; Hoffman et al., 2018; Rong et al., 2022b). It can be measured through a self-reported answer (Ooge et al., 2023; Cheng et al., 2019), satisfaction questionnaire (e.g., trustworthiness is one of the eight dimensions of satisfaction measured by the Explanation Satisfaction Scale proposed by Hoffman et al. (2018) and detailed later), or it can be observed (e.g., in semi-structured interviews). When comparing both self-reported and observed trust, it allows to measure the extent to which a user might have over or under trusted the ML system (van der Waa et al., 2021), and the persuasive power of the explanations.

- **Understanding**: Quantifying users' understanding remains challenging in XAI (Lipton, 2016) even though it is an important goal of XAI approaches. The intelligibility of the explanations can be measured as objective and/or subjective understanding. For the measurement of the objective understanding, a proxy task is generally used to measure if the model's behaviour or prediction is reliably understood by the users. We present such tasks in Section 2.4.1. The subjective understanding is generally measured post-task through questionnaires with direct questions (with Likert scale) or open-response questions.

- **User Experience** refers to the notion of usefulness and satisfaction (Mohseni et al., 2018; Hoffman et al., 2018). Various metrics are used to measure user experience, such as helpfulness, workload, satisfaction, ease of use or performance in debugging. One generic approach have been proposed by Hoffman et al. (2018) where participants self-report their overall satisfaction with the explanations. It is composed of a questionnaire called the Explanation Satisfaction Scale: it gives the participant satisfaction statements in the form of *"The explanations provided by the interface are..."*, followed by one of the eight satisfaction dimensions, respectively "*understandable*", "*satisfying*", "*sufficiently detailed*", "*complete*", "*intuitive*", "*useful*", "*accurate*", "*trustworthy*". The participant provides a score on a Likert-scale for each statement. Another questionnaire has been proposed by Shang et al. (2022) to collect the users' needs for counterfactual ex-

planations in recommendations systems. It is composed of a set of 8 questions to be answered with self-reported scores using a Likert-scale on the following variables: decision utility, experience utility (negative and positive), user action, understand what, understand why, explanation need.

- **Human-AI performance**: applies to the specific case where a human is teaming with an AI agent to achieve a particular task. The goal of this collaboration is to improve the performance on the tasks when the AI is supporting the decision-making process (Mohseni et al., 2018). To measure this performance, various metrics can be used: the score of correctly predicted instances when the considered tasks is classification, the time spent to complete the task or the perceived performance.

- **Mental models** refer to the psychometric evaluation of the users' ability to build an accurate mental model of the XAI process (Hoffman et al., 2018). Kenny et al. (2021) argue that there are three sub-models to evaluate: the model of the domain, the AI and the explanations.

A consensus has recently been reached, according to which two distinct components need to be taken into account, evaluating both objective understanding and subjective satisfaction (Cheng et al., 2019; Wang and Yin, 2021; Chromik et al., 2021). For example to measure users' understanding, comparing objective and subjective measures can provide valuable insights on the actual understanding of the users (e.g., if they overtrust or undertrust the explanations). In addition, this comparison can be used to measure if users overrate the depth of their understanding (Chromik et al., 2021). Also, it has been demonstrated that performing both quantitative and qualitative analyses help assess users perceptions of the quality of the explanations (Meske and Bunde, 2022).

2.5 | Review

After reviewing different research directions, this chapter introduced a current active one towards human-centered explainable AI: XAI approaches are progressively more general, integrating users' needs in terms of explanations; moreover, the explanation generated by a given method constitutes the basic design material for the design of explanation user interface. Such interfaces are used to evaluate the quality of the explanations in user studies.

The work presented in this thesis falls under this human-centered direction. The context considered is the interaction between an ML system and non-expert users, as presented in Section 2.2.4: this ML system is composed of a black-box ML model such as a classifier or regressor, and an explainer that allows to generate local explanations such as feature importance or counterfactual examples. However, these generated explanations remain technical and are not sufficient to provide intelligible information for users who lack knowledge both in machine learning and in the application domain. This thesis proposes to tackle these challenges as follow:

- Due to their lack of knowledge in Machine Learning as well as in the application domain, non-expert users may find the local explanations generated by current XAI approaches too complex to interpret and incomplete. In Chapter 3, we propose generic XAI principles for contextualizing and allowing exploration on local feature importance explanations for non-expert users. We also propose an implementation of the principles into an XUI for an insurance scenario. We investigate the effectiveness of these two enhancements on the intelligibility of these explanations. We tackle the challenge of evaluating the quality of explanations and design a quantitative study to evaluate two dimensions of the intelligibility, namely objective understanding and satisfaction.
- For counterfactual approaches, as presented in Section 2.2.3, it is argued that having plural examples improves the quality of the explanations. Yet, this type of explanations may also be too complex for non-expert users, and they may lead to confusion due to the quantity of information. In Chapter 4, we investigate the effectiveness of having plural examples, as compared to a single one, on the intelligibility of counterfactual explanations. We study the effectiveness of comparative analysis functionalities to mitigate potential confusion of the users when having a large set of examples. Finally, we expand the study design proposed in Chapter 3 and conduct in addition a qualitative study.
- The overall intelligibility of ML explanations is further studied in Chapter 5. While analyzing the research directions and contributions presented in Section 2.2, various inconsistencies are identified at several levels. We propose an ontology to structure common inconsistencies in ML explanation that, at its first level, can be distinguished between limitations from the ML system itself and limitations in the ability for the end-users to interpret the explanations.

3

XUI for local feature importance explanations

In this chapter, we develop an experimental procedure for designing and evaluating an XUI with intelligible explanations in the form of feature importance for non-expert users. As discussed in Section 2.5, we consider the case of a ML system composed of a regression or classification model, paired with an XAI approach that generates local explanations. These approaches can be too complex to understand for users who lack literacy in AI. Interpreting the explanations in the context of the applied domain of ML can be difficult as well for users who also lack literacy in such a domain. Finally, these explanations do not allow to interact with them, which can increase the complexity for the users to build a mental model of the ML system.

There are three contributions for this work: we propose to enhance local feature importance explanations with contextualization and exploration design principles; we also propose an implementation of these principles into an XUI for an insurance scenario; finally, we use this enhanced XUI to conduct a user study with 80 participants. The results show the relevance of the proposed principles on both objective understanding and satisfaction.

This chapter is structured as follows: after reviewing the motivations in Section 3.1, we discuss the two considered enhancements for contextualization and exploration of ML explanations in Section 3.2. In Sections 3.3 and 3.4, we successively present the two design principles we propose for these enhancements. We propose an implementation of these principles into an XUI for an insurance pricing service in Section 3.5. In Section 3.6, we discuss the protocol we propose for evaluating these explanations user study and the results. We conclude in Section 3.8.

The work presented in this chapter has led to two papers:

- Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proc. of the 27th Int. Conf. on Intelligent User Interfaces*, IUI '22, 2022.
- Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualising local explanations for non-expert users: an XAI pricing interface for insurance. In *IUI Workshop on Transparent Explanations in Smart Systems (TExSS)*. CEUR, 2021

3.1 | Motivations

Current XAI approaches, as presented in the previous chapter, allow to extract various types of local explanations for users to gain insights or evidence on why a prediction has been made for one instance. Yet, we argue that there are some limitations for the intelligibility of such explanations, especially for users lacking knowledge in both AI and the applied domain. We discuss in turn below two limitations: the lack of contextual information and the lack of guidance for exploratory methods.

3.1.1 | Need for contextual information

As discussed in Section 2.1, various definitions of explanations imply that the latter have a contextual nature. Depending on the users asking for an explanation and their underlying questions, the answers (i.e., the explanations) should vary. We discuss here contextual information that allow to align the explanations with the explainees (i.e., users' levels of expertise and needs). In XAI, it has been argued that providing the users with such context on the prediction and with basic domain knowledge can be useful for them to better interpret ML explanations (Holzinger et al., 2018; Lecue, 2020). For instance, Bellotti and Edwards (2001) 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 the behavior of such systems.

However, for now there is no consensus on what it means in practice to provide contextual information for ML explanations. In the literature, most approaches propose to contextualize the explanations with information on the ML system or dataset (Sarker et al., 2020; Selvaraju et al., 2018; Gomez et al., 2020; Wang et al., 2019). In the first case for example, Knowledge Graphs (KG) can add domain knowledge on top of ML explanations, by encoding better data representations (Selvaraju

et al., 2018), structuring a prediction model in a more interpretable way (Sarker et al., 2020) or adapting semantic similarity for local explanations (Lecue, 2020). On the other hand, context can also be extracted from the dataset. For a medical diagnosis tool, Wang et al. (2019) demonstrate the usefulness of examples from the training set, so that doctors can verify their hypotheses regarding the prediction. Finally, context can be provided through a visual representation of the dataset. For example *ViCE* (Gomez et al., 2020) offers an XUI to represent visually each feature showing where its values lie within the density distribution of the training set. However, this approach has not been evaluated with end-users, which makes it difficult to assess whether contextualizing the explanation is useful to the users.

These current approaches to contextualize ML explanations have mainly been designed for AI-experts and domain experts (DeVito et al., 2018). Martin et al. (2019) argue that novice and non-AI knowledgeable domain experts are more likely to require local explanations contextualized by specific input-output examples. It remains important to compensate their literacy gap (Burrell, 2016), in particular for their domain knowledge.

3.1.2 | Need for guidance for exploratory methods

When interacting with an interface, users have to navigate through its space to be able to access relevant information. While exploring this space, users are able to develop a mental model of the interface and its basic operations (Chromik et al., 2021). As presented in Section 2.3, exploratory methods have been studied in various XAI approaches to improve the quality of ML explanation. Exploration is considered as an important feature to include in explainable interfaces by the XAI community. Allowing the exploration of the explanations can help non-expert users build better mental models (Krause et al., 2016; Chromik et al., 2021). In an XUI, the term "exploration" can refer to two types of design enhancements, detailed in turn below: allowing users (i) to interact directly with the ML model and (ii) to navigate between several kinds of explanations.

- **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. (2016) describe in a case study how interactive dependence plots can help data scientists assess the relevance of ML models. Hohman et al. (2019) propose another interactive visual ana-

lytic system designed for data scientists to help them understand generalized additive models (GAMs).

- **Combining different explanations**: another aspect of exploration resides in allowing users to navigate between different kinds of explanations. For instance, Collaris et al. (2018) propose an interface where users are able to navigate between a feature dashboard (with feature importance, partial dependence plots and distribution in training dataset) and rule dashboard (representation of locally extracted decision rules). The results of a user study conducted with fraud agents show the usefulness of such a combination of explanations for users' satisfaction. Wang et al. (2019) report that medical doctors request to have access to both feature importance and counterfactual instances to better interpret the prediction of a diagnosis tool. Gomez et al. (2020) also combine local feature importance explanations with counterfactual examples. The *Gamut* tool (Hohman et al., 2019) combines local feature importance explanations with data density estimations.

Again, the current exploratory approaches have mainly been designed for AI-experts and domain experts, and very few have been evaluated with non-expert users.

3.2 | Overview of the propositions for enhancing local explanations

We aim at studying local explanations for non-expert users, considering such enhancements, namely contextualization and exploration. More precisely, we examine how effective the enhancements we propose are, individually and combined, to improve the explanation quality for users with no expertise, neither in the ML nor in the involved application domains. We discuss in turn below these enhancements and the design approach for the implementation of the explanations in an XUI.

3.2.1 | Three levels of contextualization

Contextualizing ML explanations can remove some layers of opacity uncovered by the current XAI approaches, as discussed in Section 3.1.1. For local feature importance explanations, we argue that there are three remaining opaque areas for non-expert users, related to the context:

- **On the ML system**: due to their low literacy in ML, the mechanisms and processes of the ML system itself can not be understood by such users,
- **On the applied domain**: non-expert users lack knowledge in the applied domain of ML, which can make it hard to interpret provided information.
- **On external factors**: some piece of information might be missing from the explanations provided, although they are implicitly linked to the process of the ML system.

Providing such missing contextual information can therefore improve the intelligibility of the explanations. We study each of these areas individually in order to propose principles to address the opacity problem, as presented in Section 3.3.

3.2.2 | Two levels of exploration

In this work, we aim at enhancing the explanations provided by a local feature importance approach such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), as illustrated in Figures 2.3 and 2.4. As presented in Section 2.3, several works have demonstrated the value of exploratory explanations, yet with no clear guidance on how to design them in an XUI. We argue that there are two levels of exploration to consider:

- **Information display**: these explanations are usually displayed without considering which piece of information might be more useful to the users.
- **Alternative scenarios**: the information displayed is only valuable for the considered instance, and does not allow the users to see how it might change the prediction with different input values.

Thus, principles for designing exploratory explanations should provide the users with more relevant explanations according to their various needs, and allowing them to test multiple settings with the ML system so that they can progressively understand how to interpret the explanations.

3.2.3 | A card-based design approach in XUI

We first propose to apply a *card-based design* for the display of the explanations in the form of local feature importance. Compared to classic local feature importance

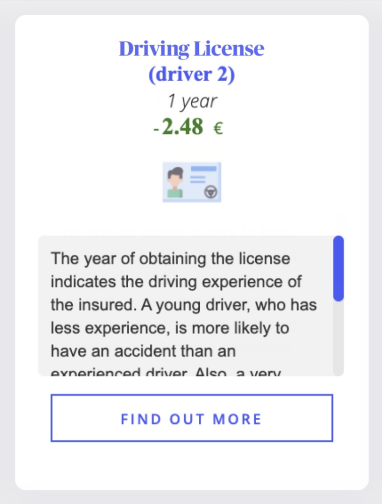


Figure 3.1: Proposed card-based design for the display of local feature importance explanations. Top part: feature label, feature value, importance score and a visual icon; Bottom: design enhancements as described in Sections 3.3 and 3.4. The insurance application framework is presented in Section 3.5.

presentation (see Section 2.2.3), this design choice allows us to associate more content and interactions with the initial explanations we generate from XAI approaches. Thus, we consider features individually and adapt the length of the card to the amount of content to display. A shown in Figure 3.1, commented below and detailed in the next sections, 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 the two design enhancements we propose: first, contextualizing the explanation with additional information; and second, exploratory functionalities to support the users' exploration, discussed in turn in Sections 3.3 and 3.4.

3.3 | Proposed XAI principles for contextualization

This section presents the generic XAI principles we propose for contextualizing local feature importance explanations. As introduced in Section 3.2, we propose to provide additional contextual information at three levels: about the ML system, about the ML application domain and about external factors influencing indirectly the prediction. Their respective description, purpose and level are discussed in turn in the following sections and summarized in Table 3.1. For each principle, we make explicit the corresponding resources to provide this additional information and to know at which level the latter should be implemented: at the feature level for the principles that give more precision to the explanation generated by the explanatory method; at the explanation level for the principles that are not specific to the instance or to the explanation, and that can therefore be applied to all the instances. These additional pieces of information should make the ML system more transparent about its purpose. We describe and illustrate the implementation of these principles in Section 3.5.

3.3.1 | ML transparency principle

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 users should interpret the explanations they get for a prediction.

Non-expert users do not know how the model has been trained and which features 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 (i.e., predictive model F and explainer G as illustrated in Figure 2.7) so users understand its purpose and basic operations, as argued by Bellotti and Edwards (2001). One of our particular concern is to introduce the difference between local and global explanations. Due to their lack of literacy in AI, users may not know the difference between these two. Thus, they might build an erroneous mental model of the overall ML system based on local explanations for a single instance, if they are not told that the displayed effects are only true for this specific prediction.

We propose that this ML transparency applies to the two components of the ML system, both over model F and explainer G. This piece of information is global and accessible at the explanation level, meaning users have access prior to interacting with the explanations, so as to better interpret them. We believe that this information has to

be provided by a ML expert to introduce this ML system, and translated into simple information so that it is comprehensible to people without technical knowledge.

3.3.2 | Domain transparency principle

We propose to associate local feature importance explanations with additional global information related to the considered 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.

Information provided by local feature importance 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 global, i.e., applicable to all instances, and should be paired with each local feature importance explanation to improve understanding of explanations' operations. We believe this information has to be provided by domain experts.

3.3.3 | External transparency principle

Local feature importance provides explanations about a given prediction model but external factors can also influence the outcome. 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 influence 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. This for example applies to personal information like name, gender, or phone number that can be asked to communicate with users, but not used by the ML prediction model. Yet, users may believe it is taken into account for the prediction they get. Thus, it is important to be transparent with users about which factors impact or not the prediction, even though this information might be external

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	Explanation
	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	Feature
	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	Explanation
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.	Ease access to most relevant explanations according to the users' goal.	Explanation
	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	Feature

Table 3.1: 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.

to the model. We propose an external transparency principle that makes more explicit the impact of such factors.

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

3.4 | Proposed XAI principles for exploration

This section presents the generic XAI principles we propose for allowing exploration on local feature importance explanations. They come into play at two different levels: offering an interactive display of the explanations at an overall explanation level, and showing example-based explanations at a feature level. Their respective description, purpose and level are discussed in turn in the following sections and summarized in Table 3.1.

3.4.1 | Interactive display principle

Local feature importance explanations usually display features in decreasing order of the absolute feature importance values (see Section 2.2.3). Users see the major positive influence 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 adapt the display of the feature according to their own needs and goals.

As presented in Sections 2.2.4.2 and 2.3.1, users may want to test different hypotheses when interacting with a ML system to help them make an informed decision regarding the prediction. For non-expert users, they may not know how to explore and we believe it is important to provide guidance regarding the possible exploratory paths (e.g., if the users are seeking to optimize the prediction, they could be confused about which values can possibly be modified).

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

3.4.2 | Example-based explanation principle

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.

Similarly to counterfactual explanations (see Section 2.2.3), this example-based explanations principle should allow to play with the values for each feature and emphasize the impacts of these changes on the prediction. Indeed, it is not intuitive for non-expert users that the score of one attribute is specific to one instance. They may believe it is the same score for everyone (e.g., for a motor insurance pricing service, a specific car model would always have 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-expert 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.

3.5 | Implementing XAI principles in a real life application

This section presents the application of the XAI principles we propose, as described in Sections 3.3 and 3.4, into an insurance-related interface, as illustrated in Figures 3.2, 3.3 and 3.4. We describe the usage scenario in Section 3.5.1 and the design process for implementing these principles in the user interface in Sections 3.5.2 to 3.5.4.

3.5.1 | Usage scenario: motor insurance pricing

We apply the principles we propose in a motor insurance pricing interface. In this scenario, users provide several pieces of information regarding their insurance settings and background (desired coverage and options for the car, personal bonus/malus, insurance history), the car 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.

Based on this information, a ML system predicts a price and provides local explanations, that are used as basis material for the design of the XUI. The aim of the XUI is that prospective clients using this service to compute a personalized price for a new motor insurance can understand how their information impact the price they get.

3.5.2 | Implementing contextualization principles

The implementation of the contextualization principles we propose, as presented in Section 3.3, is illustrated in Figures 3.2 and 3.3. We describe in the following paragraphs the design of these explanations with the implemented principles.

ML transparency We design an onboarding text above the cards (see A in Figure 3.3) of local feature importance explanations: it explains how the price for a motor 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. As stated in Section 3.3, this onboarding text is provided by a ML expert and translated into non-technical information by a designer.

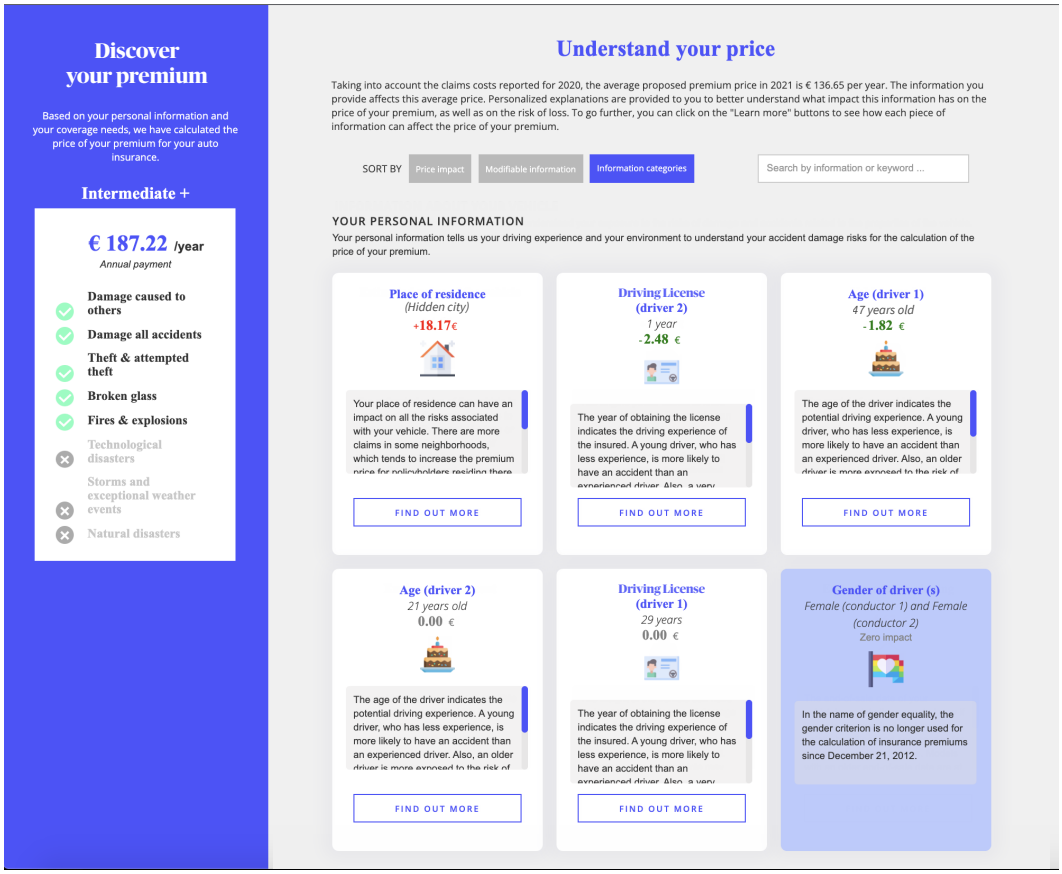


Figure 3.2: Application of contextualization and exploration principles in a fictive insurance-related scenario. This interface presents a personalized premium price for a prospective client on the left, and explanations on the right. See Figure 3.3 for zoomed views on the implementation of some contextualization principles and Figure 3.4 for some exploration principles *Note: The interface has been translated from the original language used for the evaluation.*

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 (see Section 3.2.3), and we pair it with more generic information about how the feature can impact the price in the context of motor insurance on the bottom (see B in Figure 3.3). As stated in Section 3.3, this information is provided by an insurance expert, i.e., an actuary in the considered scenario.

External transparency We introduce an external factor card into the list of feature-associated cards, as an additional one. It has a similar design to feature-associated ones, except that it has a different background color to be visually differentiated by users (see C in Figure 3.3). 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,

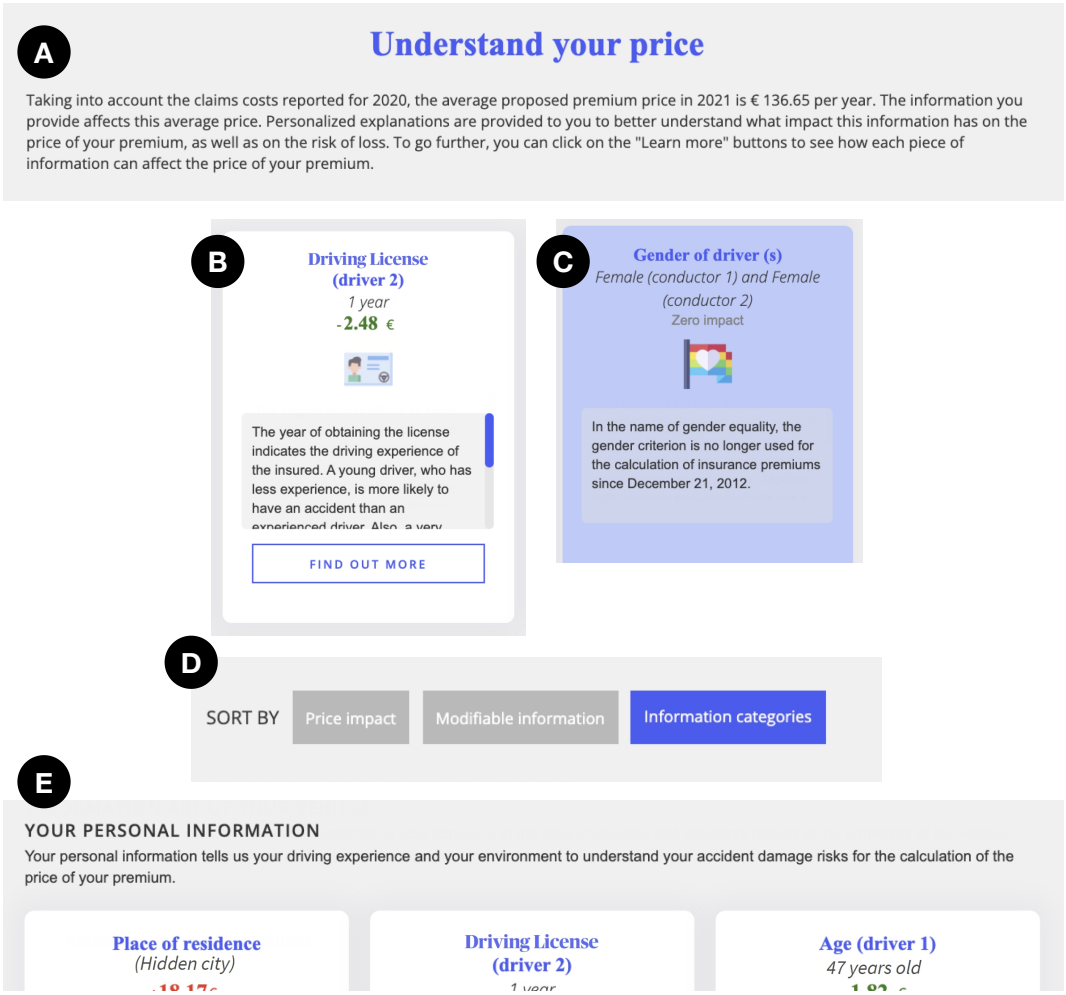


Figure 3.3: Implemented principles views: A. ML transparency B. Domain transparency and part of example-based explanations C. External transparency D and E Interactive Display *Note: The interface has been translated from the original language used for the evaluation*

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 motor insurance. Therefore, we explicitly display that this piece of information is not used by the model. Again, this information is provided by by the domain expert, i.e., an actuary in the considered scenario.

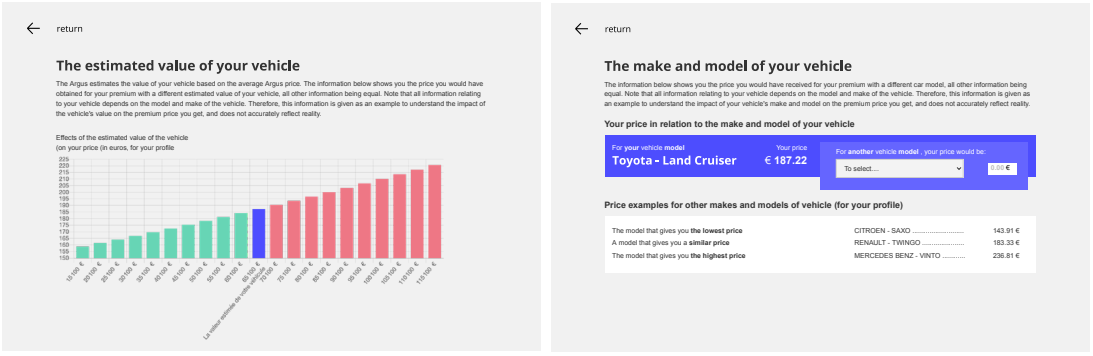


Figure 3.4: Application of the example-based principles for allowing exploration. Left: for features with continuous values, we use a bar graph to display example-based explanations. Right: for features with categorical values, we display a drop-down list of the most frequent values of the feature with associated prediction, as well as three relevant examples of feature values. *Note: The interface has been translated from the original language used for the evaluation*

3.5.3 | Implementing exploration principles

The implementation of the exploration principles we propose, as presented in Section 3.4, is illustrated in Figures 3.2 and 3.4. We describe in the following paragraphs the design of these explanations with the implemented principles.

Interactive display We design filter buttons above the list of the feature-associated cards (see D and E in Figure 3.3), 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. This is the usual display for feature importance explanations.

Second, cards can be sorted so as to display first the ones that correspond to actionable features, i.e., features that can be realistically edited by users so that they can try to optimize the price they get. For example, users can switch the type of coverage they want, but they cannot change the date when they obtained their driving license. This feature actionability is determined by an insurance expert, i.e., an actuary in the considered scenario.

Third, cards can be sorted according to the categories of information they contain (see Section 3.5.1), so as to follow the logic of the input stage (i.e., when users fill in their information). Thus, 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 (see B in Figure 3.3), to access a second page displaying details in the form of example-based explanations, as illustrated in Figure 3.4.

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. For example, on the estimated value of the car, the graph shows what would have been the predicted price if this value was higher or lower than the actual value. Bars have different 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. For example on the model and brand of the car, users can select another combination within the most frequent ones to see what would have been the predicted price. 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 feature value that differs from the one of the user. This allows users to know where their information fit in the overall data distribution and to better understand that the explanation is local.

3.5.4 | Combining contextualization and exploration principles

We believe that the ML transparency principles can be beneficial both to interactive display and example-based explanation principles in the XUI.

For the interactive display, we implement an introduction text at the beginning of each category of feature-associated card for each filter option we design (see E in Figure 3.3). 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 top part of interfaces shown in Figure 3.4).

3.6 | Experimental evaluation

To evaluate the effectiveness of the XAI principles we propose, we use the XUI presented in Section 3.5. We measure two dimensions of the intelligibility of the explanations, namely objective understanding and satisfaction. We conduct a first experimental evaluation online and use the results to finalize the evaluation method and material (see Appendix A). We then conduct a monitored users study with 80 participants at the INSEAD-Sorbonne University Behavioural Lab. The results show that the contextualization principles we propose significantly improve the users satisfaction and are close to have a significant impact on the users objective understanding. They also show that the exploration principles we propose improve the users satisfaction. On the other hand, the interaction of these principles does not appear to bring improvement on both dimensions of users' understanding. We describe in turn below the material in Section 3.6.1, the method we use to conduct the monitored study in Section 3.6.2 and detail the result of the study in Section 3.7.

3.6.1 | 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 motor insurance customer and extract explanations for this price with the SHAP method (Lundberg and Lee, 2017), as described in Section 3.6.1.1. We use this prototype to test our hypotheses towards the effectiveness of the XAI principles we propose on two dimensions of user's understanding, as described in Section 3.6.1.2.

3.6.1.1 | Interactive prototype

We develop an interactive prototype for a motor insurance pricing interface, as described in Section 3.5.1. We describe in turn below the model we use in order to provide the users with SHAP and example-based explanations, the dataset and the explanation extraction.

Pricing model We develop, with the help of an AXA 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 (Dutang and Charpentier, 2020), 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 100,000 policies.

Explanation extraction We use SHAP (Lundberg and Lee, 2017) to generate local feature importance explanations for the price estimation (see Section 2.2.3). To generate example-based explanations, we compute partial dependence plots for each feature (Greenwell et al., 2018). 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 3.5.3.

3.6.1.2 | Hypothesis testing

We aim at studying the explanations provided in the form of enriched local feature importance for non-expert users. More precisely, we examine the effectiveness of the two enhancements we propose, individually and combined, on the explanation quality for users with no expertise, neither in the ML nor in the involved application domains. As presented in Section 2.4 and discussed in more details in Section 3.6.2, we consider two components for this explanation quality, distinguishing between objective understanding, which assesses the extent to which users actually understand the explanation, and satisfaction, which assesses the extent to which users appreciate the XUI. 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 explanations improves non-expert user understanding
  - H.1.2: Allowing exploration in explanations improves non-expert user understanding
  - H.1.3: Contextualizing and allowing exploration in 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 explanations improves non-expert user satisfaction
- H.2.2: Allowing exploration in explanations improves non-expert user satisfaction
- H.2.3: Contextualizing and allowing exploration in explanations improve even more non-expert user satisfaction

We expect that the principles we propose increase both the objective understanding and users satisfaction. 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).

3.6.2 | 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 the INSEAD Institutional Review Board (IRB).

3.6.2.1 | 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 described below, allowing us to compare the impact of both contextualization and exploration factors on scores of objective understanding and satisfaction. We discuss in turn the participant recruitment and interfaces they were assigned to.

Participant recruitment We recruited 91 participants 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, they were asked 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. All participants were financially compensated at the end of the experiment.

Tested interfaces 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 3.5, 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 (see Figure 3.5). It simply displays the local feature importance explanations with the card-based design described in Section 3.2.3. None of our design principles are applied in this version.
- Interface B is the contextualization factor interface (see Figure 3.6). It adds to interface A the three principles we propose for contextualization: ML transparency, domain transparency and external transparency (see Section 3.5.2).
- Interface C is the exploration factor interface (see Figure 3.7). It adds to interface A the two principles we propose for allowing exploration: the interactive display and the example-based explanations (see Section 3.5.3).
- Interface D is the interaction interface. It combines all the principles of contextualization and exploration (see Section 3.5.4). Figures 3.2 and 3.4 present screenshots of this version.

3.6.2.2 | Evaluation questionnaires

In this work, we adopt the approach described in Section 2.4.2 and evaluate both the objective understanding and satisfaction. We 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 (see Appendix A), similar to Cheng et al. (2019). 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:

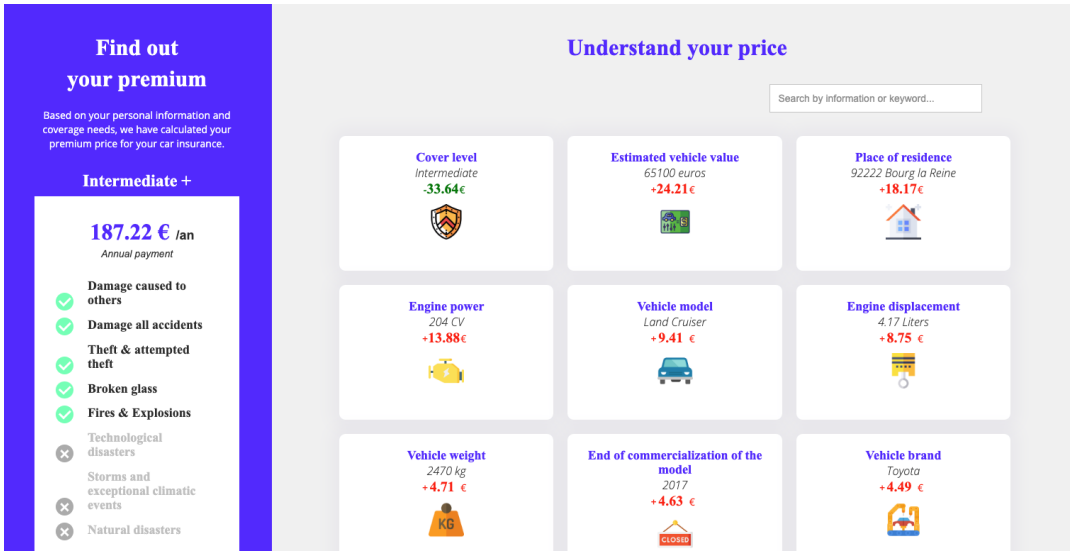


Figure 3.5: Interface A is the baseline version of feature importance explanations.

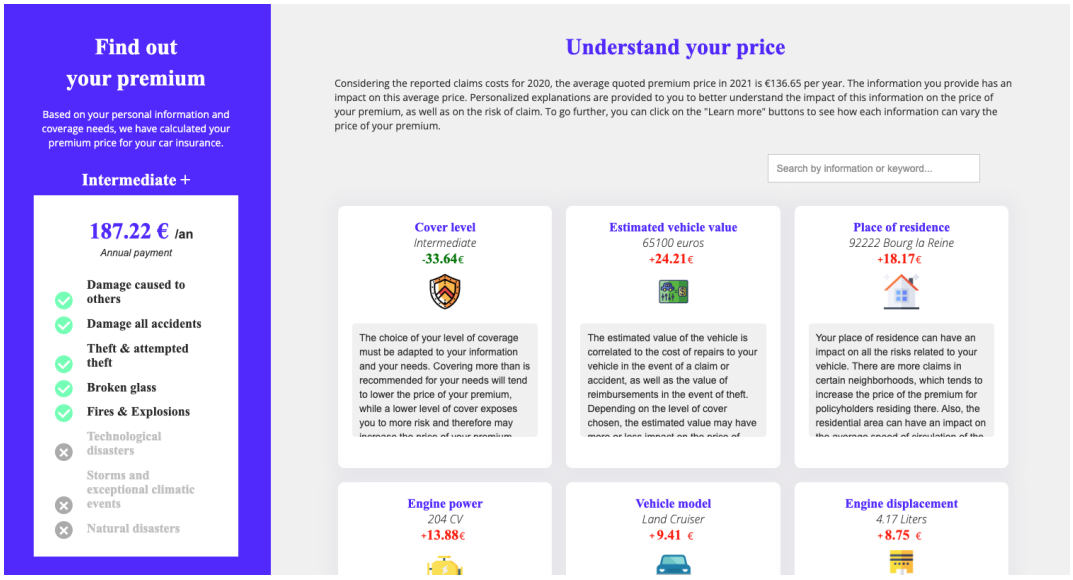


Figure 3.6: Interface B is the XUI we propose for implementing contextualization principles only.

- *(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.*
- *(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.*

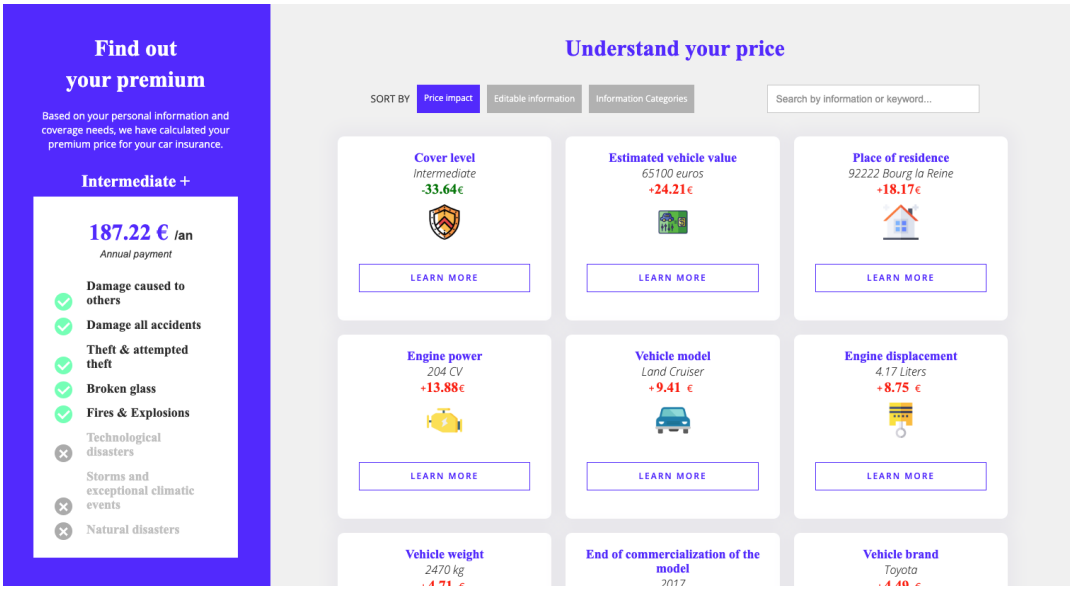


Figure 3.7: Interface C is the XUI we propose for implementing exploration principles only.

- *(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.*

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 Hoffman et al. (2018) in order to assess users' satisfaction (see Section 2.4.2). Participants are required to answer on a 6-point Likert scale, from "Strongly disagree" (1) to "Strongly agree" (6).

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, motor 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.

3.6.2.3 | Study procedure

The lab setting at INSEAD-Sorbonne University Behavioural Lab we apply allows participants 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 3.6.2.2) while interacting with the interface, and then answer the subjective satisfaction questionnaire (see Section 3.6.2.2). At the end of the experiment, participants complete the demographic survey.

3.6.2.4 | Data analysis

We remove one extreme outlier (below 'Q1 - 3*IQR' 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 3.2 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 users 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 3.3 and 3.4 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.

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 3.2: Comparing improvement of objective understanding between two factors: contextualization and exploration. Top: results of a one-way ANOVA regarding the significant effect of contextualization factor on users objective understanding. Bottom: results of a two-way ANOVA regarding the significant effect of contextualization and exploration factors on users satisfaction are displayed on the bottom.

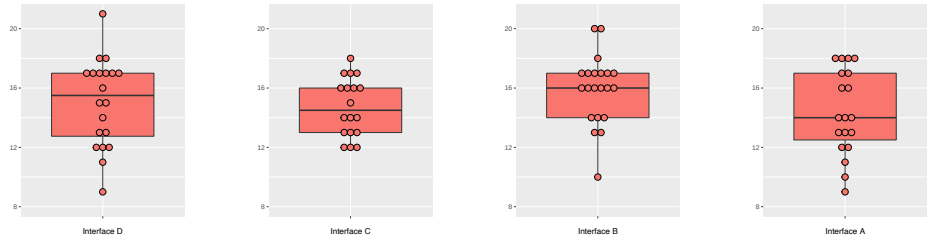
3.7 | Results

We use the results presented in Tables 3.2, 3.3 and 3.4 to answer the two research questions we consider regarding objective understanding in Section 3.7.1 and users satisfaction in Section 3.7.2.

3.7.1 | Objective understanding

We analyze the significant effects of both contextualization and exploration factors on users objective understanding score. As neither exploration factor nor the interaction of the two factors show significant impacts, we use one-way ANOVA to measure the effect of contextualization on the objective understanding score. The analysis of Table 3.2 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; on the other hand, 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 3.3, we can see the interface including contextualization principle only (interface B) shows the highest improvement in objective understanding with an av-



#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 3.3: 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.

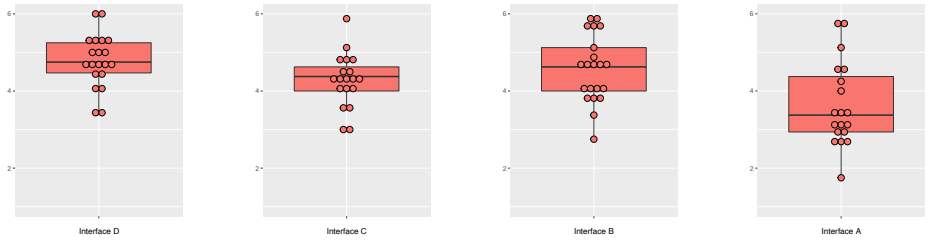
erage score of 15.90 correct answers out of 22, i.e., .85 point more than when these principles are paired with exploration ones (interface D) or 1.53 point more than when no principles are applied (interface A). When comparing the average means of conditions with contextualization principles applied (interfaces B and D) in Table 3.2, 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 3.3 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 either Previous observations suggest a promising positive effectiveness of contextualization but reject exploration one regarding the objective understanding.



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

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).

3.7.2 | Satisfaction

Similarly to the previous analysis for objective understanding, we analyze the significant effects of both contextualization and exploration factors on users satisfaction score. As the interaction of the two factors shows no significant impact again, we use two-way ANOVA to measure the effect of contextualization and exploration factors on the satisfaction score. Table 3.4 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. On the other hand, 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 3.4 shows that contextualization principles (interface B) obtain a higher satisfaction rate as they increase by .93 point the average satisfaction score as compared to the interface without these principles (interface A), and by .30 point as compared to the interface with ex-

ploration 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 3.4. This difference is also observed in the two-way ANOVA analysis in Table 3.2 as the average mean 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 3.4 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 3.4. When analyzing the impact of the exploration factor in Table 3.2, 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 3.4 shows that the interaction 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 3.4, we observe that the 1st quartile for the interaction 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).

3.8 | Conclusion

In this chapter, we present generic XAI principles we propose for contextualization and exploration of local feature importance explanations for non-expert users. We propose an implementation of these principles into an explanation user interface for a motor insurance pricing scenario. The experiment we conduct in a moderated lab setting shows that the contextualization principles we propose significantly improve users satisfaction and are close to significantly improve users objective understanding. Also, the results show that the exploration principle we propose significantly improves users 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. (2019). 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 (Schwarcz, 2014), as confirmed by several participants of the pilot study we conducted (see Appendix A). 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 and exploring 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, e.g., about their having a driving license. Another direction is to increase the number of participants: the current number is for instance not high enough to allow an individual comparison of the three contextualization principles we propose (ML, domain or external transparency, as well as their combined effect).

4

XUI with plural counterfactual explanations

In this chapter, we investigate the intelligibility of another type of local explanations, namely counterfactual examples, also considering of non-expert users. Similarly to the work presented in the previous chapter, we consider that a classifier or regressor is providing a prediction, and a post-hoc XAI approach is generating explanations. Recent works underline the issue that most counterfactual approaches have not been tested with real users: only 21% of the surveyed methods by Keane et al. (2021) have been tested. There is a lack of empirical research in understanding the users needs for counterfactual explanations in their usage (Keane et al., 2021; Verma et al., 2022; Shang et al., 2022). This also applies to the case of explanations in the form of plural counterfactual examples (see Section 2.2.3), although it is argued that they improve the quality of the explanations. Yet, it has been shown that too much information in the explanations may affect users trust (Kizilcec, 2016) and create confusions (Cai et al., 2019).

In this work, there are three contributions: we propose comparative analysis XAI principles to enhance the quality of plural counterfactual explanations; we also propose an implementation of such enhanced explanations in an XUI for a financial scenario; finally, we use this enhanced XUI to conduct a user study in a monitored lab setting with 112 participants. The results of the statistical analysis demonstrate the effectiveness of the plural condition, both on objective understanding and satisfaction scores, as compared to having a single counterfactual example. The qualitative analysis of the results shows that the proposed comparative analysis features are promising approaches to improve the intelligibility of such explanations. This work provides the first experiment, to the best of our knowledge, evaluating the intelligibility of plural

counterfactual examples for non-expert users.

The chapter is structured as follows. In Section 4.1, we discuss the process we propose for the design of an XUI for the non-expert users. Section 4.2 presents the proposed design enhancements for comparative analysis of plural counterfactual examples. Section 4.3 presents the illustration of such enhanced explanations in an XUI for a financial scenario. In Section 4.4, we discuss the protocol we propose for evaluating these explanations in a user study and present the results in Section 4.5. Finally, we discuss limitations and future works in Section 4.6.

The work presented in this chapter has led to the following paper:

- Clara Bove, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In *Proc. of the 28th Int. Conf. on Intelligent User Interfaces*, IUI '23, 2023

4.1 | Motivations

Various types of explanations can meet the needs of non-expert users, and some can be more relevant in a specific context. For instance, when a user wants to optimize the prediction he/she obtains, then a counterfactual example can be useful (see Section 2.2.3). However, as discussed in Chapter 3, it is not always easy for all users to accurately interpret the explanations provided. In fact, they can get lost in the amount of information (Kizilcec, 2016), and often lack guidance on where to look and how to interpret an explanation as a whole. We discuss below the process we propose for the design of a user-centered explanation: we first analyze the limitations of the considered explanation in Section 4.1.1; then we study the design opportunities to improve the quality of these explanations in Section 4.1.2; finally, we discuss the interface design approach in Section 4.1.3.

4.1.1 | Need for guidance

As discussed in Chapter 2, it has been argued in numerous work that counterfactual examples constitute a highly relevant form of explanation due to their resemblance with human explanations (Wachter et al., 2017; Wang et al., 2019; Miller, 2019; Byrne and Tasso, 2019; Zhang and Lim, 2022). Among others, they possess contrastive properties, i.e., they are formulated as answers to *Why not?* questions (see Section 2.1). Such explanations are particularly useful to users who are trying to understand why

they did not get a desired outcome, for instance using a canonical example, if a ML model predicts their loan application is denied (Ramon et al., 2021).

This work focuses on the case of plural counterfactual examples, i.e., when counterfactual explanations contain several examples (see Section 2.2.3). Indeed, it has been proposed to build approaches to generate so-called diverse counterfactual examples Kunaver and Porl (2017); Mothilal et al. (2020); Ekstrand et al. (2014), claimed to constitute more relevant and appropriate explanations. In this work, we only take into account the fact that they provide several examples instead of a single one and does not study the extent to which they differ one from another. Thus, we favor the word "plural" instead of "diverse".

Recent works underline the issue that these explainability methods have not been tested with real users, and that there is a lack of empirical research in understanding the users' needs for counterfactual explanations in their usage (Keane et al., 2021; Verma et al., 2022; Shang et al., 2022). This also applies to the case of explanations in the form of plural counterfactual examples (see Section 2.2.3), although it is argued that they improve the quality of the explanations. Yet, it has been shown that too much information in the explanations may affect users trust (Kizilcec, 2016) and create confusions (Cai et al., 2019). In the process of explanation assimilation, non-expert users may need to compare and analyze various information and we argue that they need guidance and complementary information due to their lack of knowledge in both machine learning and the applied domain.

4.1.2 | Design opportunity: two levels of guidance

In this work, we investigate (i) if plural counterfactual examples are indeed better than having a single one, and (ii) if we can mitigate the users' confusion with a comparative analysis enhancement when there is a high number of examples.

When there is a rich set of counterfactual examples, we argue that there are two levels of complexity for non-expert users:

- **On counterfactual example**: due to their low literacy in ML and in the applied domain, the nature of the explanations (such as suggested changes, combination of features, distance to initial values) may not be understood by non-expert users.
- **On the set of examples**: the users may not be able to assess and compare an example with another; also, this can create further confusion as to the nature of this type of explanation.

Figure 4.1: 3-column grid of cards. Each card represents a counterfactual example with suggested changes on one or multiple features values. The set of counterexamples is displayed on an interactive 3-column grid.

Guiding the users on the analysis and comparison of such examples can therefore improve the intelligibility of the explanations. We study these two levels in order to propose principles to address the confusion issue, as presented in Section 4.2.

4.1.3 | Interface design: a grid of cards

For counterfactual explanations with plural examples (see Section 2.2.3 and Figure 2.5), a rich set of counterfactual examples is generated, and each counterexample suggests changes on one or more values. To display such explanations, we apply a grid of cards design approach, as illustrated in Figure 4.1. The set of counterexamples is displayed on a grid and we consider counterexamples individually: each one is represented on its own card and adapt the length of the card to the amount of content to display. As compared to the design approach in Chapter 3, a card does not present only a single feature but multiples ones for each example.

Basically, each card displays labels of the data descriptive features whose values are modified in the considered counterexample, together with these new values that allow to reach the class opposite to the predicted one. Similarly to the work presented in Chapter 3, we propose to name these features with non-technical labels: we use the names known from the user (see Section 4.3.1). Also, we propose to group the features into contextual categories (see the interactive display principle we propose in previous work and present in Section 3.4), so users can identify quickly what category of information is impacted by the suggested change: e.g., for a loan application, the feature "age" is displayed under the category "Personal Information". This design choice allows us to associate more content and interactions for each counterexample provided by the ML system, as discussed in Section 4.2.1.

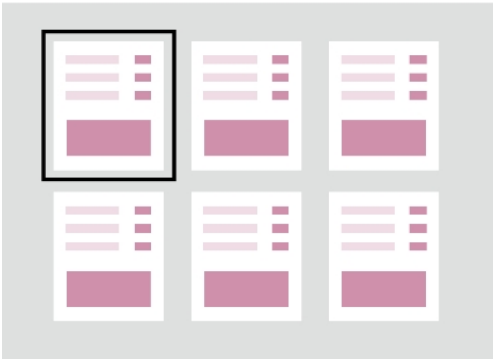


Figure 4.2: Highlight singularities design enhancement at the counterexample level. Top: highlight the non-zero differences with initial values. Bottom: highlight the value of the counterexample as compared to others with additional information derived from expert knowledge.

The set of cards is presented in a 3-column grid, as illustrated in Figure 4.1. When generating plural counterfactual examples, it can be difficult to present a rich set in one screen. Here, we use the grid so that users can scroll on the page to explore the different counterfactual examples. To navigate between the cards, we add interactive display tools to support this exploration: a search bar using key words to allow to automatically filter the set of cards (as illustrated in Figure 4.4). We also add a "sort by" button above the grid: users can sort the counterfactual examples by increasing or decreasing number of modified data feature. This design choice allows us to enhance this exploration with additional interactive tools, as discussed in Section 4.2.2.

4.2 | Proposed principles for comparative analysis

We propose comparative analysis principles that aim at making it easier for non-expert users to compare and analyze this rich set of counterexamples. We propose two XAI principles to do so: highlighting singularities of each example, presented in Section 4.2.1 and illustrated in Figure 4.2; and guiding the non-expert users to analyze and compare a rich set of counterfactual examples, presented in Section 4.2.2 and illustrated in Figure 4.3. They respectively apply at two levels we propose to distinguish: the first one corresponds to each counterfactual example represented in a card, individually; the second considers the grid with all counterfactual examples globally. Their description, purpose and level are discussed in turn below and summarized in Table 4.1. We propose an illustration of these principles in an XUI for a financial usage scenario in Section 4.3.

4.2.1 | Highlighting examples' singularities

The first principle aims at visualizing and assessing singularities in order to help the users differentiating one counterfactual example from another.

When interacting with plural counterfactual examples, users may not know how to interpret the proposed changes. In order to help users better interpret and assess each counterexample, it is important to be more precise regarding the meaning of the provided explanation. Also, the users may need extra information on the value of each counterfactual. Thus, we propose to highlight these two elements on the example card, as illustrated in Figure 4.2.

On the top of the card, we highlight the non-zero differences with initial values, that differs from the information retrieved by plural counterfactual methods such as *DiCE* (Mothilal et al., 2020) which displays the changed values as the counterfactual explanations.

At the bottom, we propose to pair the counterfactual example with new information derived from expert knowledge to highlight its value as compared to other example. For example, we add a feasibility score regarding the suggested changes on the example. For each data descriptive feature, three levels of feasibility are distinguished for the suggested variations: they can be either feasible, moderately feasible or hardly feasible, depending on the context of use. For counterfactual examples with more than one data descriptive feature variation, we adopt a pessimistic aggregation approach and display the lowest level of feasibility between all the involved modified data features. This level of feasibility aims at providing non-expert users with an additional element for the interpretation of the counterfactual. This information can be used to compare the counterexamples and select the most appropriate ones for the users.

We propose that these highlighted singularities features are accessible on each card of counterfactual examples, so that the non-expert users can better interpret and assess them.

4.2.2 | Guided comparison

As previously presented, we should also provide more guidance to the non-expert users on how to analyze a set of various example-based explanations. We propose a **guided comparison** XAI principle, that aims at underlying the differences between examples, and should match the users needs when comparing and analyzing them.

Non-expert users may get lost when exploring various examples, not knowing

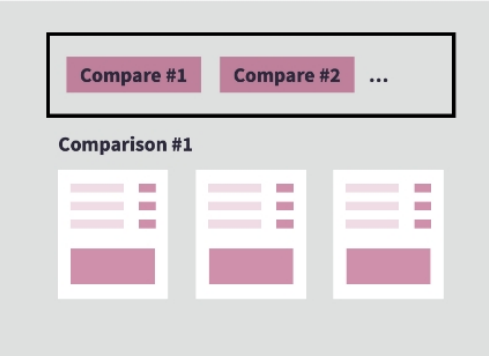


Figure 4.3: Guided comparison design enhancement at the grid level. Top of the grid: filtering buttons for the display of the plural counterfactual examples aiming at underlining the differences between examples.

where to start and how to navigate between them. Thus, they should have more guidance towards the directions they should follow when exploring and analyzing a set of counterfactual examples. We propose to offer users filtering options for the dynamic display of the plural counterfactual examples, as illustrated in Figure 4.3. These options allow users to change the ordering and/or the filtering of the cards to better compare them. The first option corresponds to the generic display of the cards as generated by the explainer. We propose two additional buttons with sorting/filtering options to guide comparison.

First, as the suggested plural examples are not necessarily diverse, we add an option that filters the cards to display only the most diverse ones. We apply a heuristic approach to select these most diverse examples from the generated set. First, we define ensembles of modified attributes; for each of these ensembles, we select the example with the closest proximity to input values; and then, we classify these examples by number of modified attributes. For instance, when there are plural cards that suggest changes on a similar feature, it will only display the one that is the closest to the instance value. This allows to present the users with a synthetic overview of all the closest and diverse counterfactual examples in real value.

Second, as features are grouped into contextual categories (see Section 4.1.3), we propose to add a filtering option that filters the cards regarding the frequency of changes by categories: there can be several counterfactual examples that suggest changes on the same data descriptive feature, which leads to define frequently modified features and further on to frequently modified feature categories. We add dynamically a sub-filtering option for each frequent category of change suggestions, in order to filter and only display the counterfactual examples offering such changes. We propose to display these sub-filtering options in a frequency decreasing order (i.e.,

Principle	Description	Purpose	Level
Highlight singularities	Enhance the counterfactual examples by highlighting two complementary information: the non-zero differences with initial values and the added value of the example as compared to others.	Help the users for an accurate interpretation of each counterfactual example	At the example level
Guided comparison	Offer the users pre-defined filtering options for the display of the plural counterfactual examples. These options should match the users needs when comparing and analyzing a set of examples.	Ease the analysis and comparison of plural counterfactual examples towards the predicted output	At the explanation level

Table 4.1: Design principles to improve the intelligibility of counterfactual explanations with plural examples for non-expert users. We describe and define the purpose of each principle we propose for adding comparative analysis features. We also define the level of the ML explanations where the described principle applies: "explanation level" refers to principles that apply to the overall ML explanations for one prediction; "example level" refers to the principles that apply to each counterfactual example.

from the category with most counterfactual examples to the one with the least), so as to analyze in which category of data descriptive features there are the most suggested variations.

We propose that these guided comparison features are accessible above the grid of cards, so that users can filter the cards to better navigate between them.

4.3 | Illustrating principles in a real life application

This section presents the application of the XAI principles we propose into a finance-related interface. We describe the usage scenario in Section 4.3.1 and the XUI illustrating the implementation of the principles we propose in Section 4.3.2.

4.3.1 | Usage scenario: loan application

We apply the principles we propose in a solvency evaluation interface. In this considered scenario for the usage of the interface we propose, a user connects to a platform and starts applying for a loan by providing several pieces of information such as the the desired loan settings (loan amount, duration, installment rate), bank information

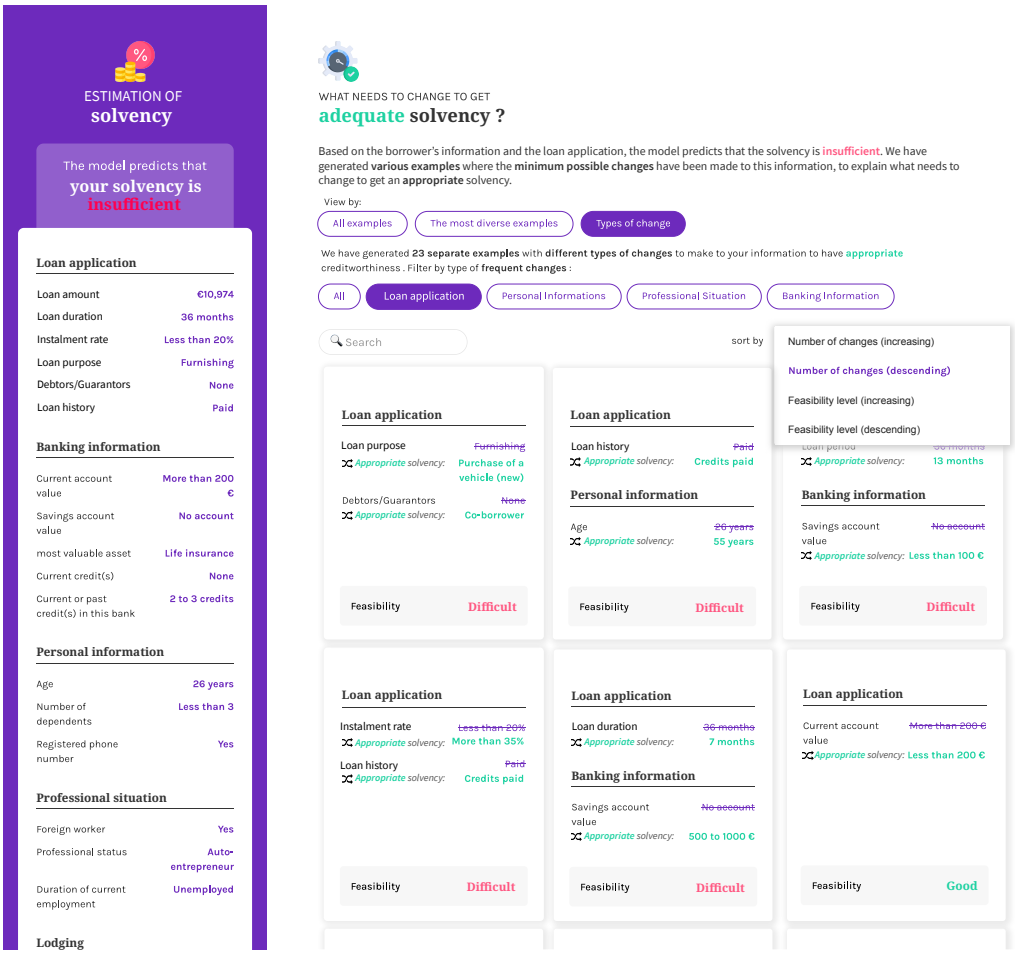


Figure 4.4: Implementation of plural counterfactual examples and comparative analysis principles in a fictitious finance-related scenario. Left: estimated solvency for the considered loan application. Right: provided explanations with grid presentation of the cards associated with each counterfactual example. The highlighted singularities are implemented as enhanced changes and added expert knowledge on feasibility scores. The guided comparison is implemented with contextual filtering and sorting options on top of the grid. *Note: The interface has been translated from the original language used for the evaluation.*

(bank account and savings values, current and/or loan history), personal information (age, number of dependents, phone number), professional situation (current job occupation and duration, foreigner worker status), as well as current lodging situation. This information is usually required by financial organizations to evaluate the solvency of the applicant according to each individual risk for the payment and reimbursement of the loan.

We consider that the names used in this form define non-technical labels the user understands, as he/she fills them: they thus constitute the labels used in the expla-

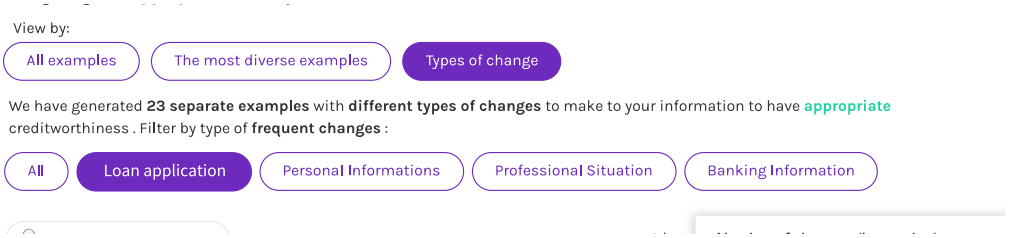


Figure 4.5: Application of the guided comparison principle: for "Types of change" button, an additional row of button appears and offers different filtering options of the cards display, according to frequency of similar changes. *Note: The interface has been translated from the original language used for the evaluation.*

nation interface.

A ML model uses this information to estimate the solvency of this user. The aim of the XAI interface is to present the estimated solvency to the user, together with explanations to help him/her understand how the provided information impact the evaluation.

4.3.2 | Proposed XUI interface

The implementation of the explanations in the form of plural counterfactual examples, as well as the comparative analysis XAIO principles we propose to enhance such explanations is illustrated in Figures 4.4 and 4.5. On the left, this interface presents the solvency predicted for the considered loan application whose characteristics are supposed to have been inputted to the system in a preliminary step. Similarly to the proposed XUI in previous chapter, we provide the user with transparency on the ML system's scope and basic operations above the explanations, as we demonstrate that it can help the users understand how the model works and how to read the following explanations. We describe in the following paragraphs the design of these explanations with the implemented principles.

Implementing plurality We use the grid of cards design approach presented in Section 4.1 to display the rich set of counterfactual examples. The search bar and the "sort by" are offered to the user above the grid, as illustrated in Figure 4.4, to allow users to search for specific information and sort the cards. We implement the filtering options into buttons on top of the grid as well, as presented below.

Highlight singularities Each example-associated card contains the two complementary pieces of information described in Section 4.2.1: highlighted information

about the counterfactual change on the top and highlighted value of the example as compared to others with a feasibility score on the bottom.

On each card, for all features modified by the counterfactual example, we display the initial value, striking it through, and we highlight, in bold and green color, the new counterfactual value. We also add the legend "Good solvency" next to this value, which is the opposite class. We do so to highlight that this change (or the combination of these changes for counterfactual examples with plural changes) would have made the model to predict the opposite class.

In addition, we add the feasibility score on each card as described in Section 4.2.1. For each feature, a domain expert defines three levels of feasibility (i.e., easy, moderate or difficult to do in real life). For example, for the feature "Current account value", the changes proposing to decrease the value are defined with a "good" score of feasibility, while those proposing to increase the value are defined with a "difficult" one. In the user interface, we add a color code for the three levels of feasibility (green for feasible, orange for moderate and red difficult) in order to ease the visual screening of the examples for non-expert users.

Guided comparison We design filter buttons above the list of the feature-associated cards (see Figure 4.5), allowing users to change the ordering and/or the filtering of the cards to better compare them. As described in Section 4.2.2, a first button allows to display the example as generated by the XAI approach. Another button allows to filter the cards and the display most diverse examples. A last button allows to filter the cards by frequency of changes. When selected, it activates an additional row of buttons below, offering different filtering options of the cards display, according to frequency of similar changes. In this context, the most frequent types of changes are for the "loan application" settings, and the least frequent are for the "Banking information".

4.4 | Experimental evaluation

To evaluate the effectiveness of the XAI principles we propose, we describe in turn below the prototype we build and the evalulation method we used to conduct a monitored study at the INSEAD-Sorbonne University Behavioural Lab. 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 4.4.2.

4.4.1 | Prototype

We develop an interactive prototype of the proposed XUI for the evaluation, as described in Section 4.4.3. We discuss in turn below the data set we use to train a ML model for the estimation of the solvency for a prospective loan customer and the method we use to extract diverse counterfactual explanations.

We develop an interactive prototype for a solvency estimation service, as described in Section 4.3.1. We use the German Credit dataset (Hofmann, 1994) which is a public dataset downloaded from the UCI Machine Learning Repository [1]. It contains the description of 1000 loan applicants on 20 descriptive features and their labels as having a good or bad solvency.

We use this data set to train a ML model to compute a predicted solvency for each user, namely a Random Forest trained with default parameters with the sklearn tool[2]. On the estimated solvency we get for one instance, we use the *DiCE* method[3] (Mothilal et al., 2020) to generate diverse counterfactual examples to explain the given output. To obtain diversity in this set, *DiCE* requires the number of desired examples (defined as `total_CFs=23` and `desired_class="opposite"`), the weight on distance and sparsity (defined as `proximity_weight=1.5` and `diversity_weight=1.0`), as well as the definition of user knowledge such as the list of features that can be modified and their associated range of accepted variation. For the instance we select from the training set and the chosen DiCE configuration, that excludes modifying "foreigner worker status" and "phone number", DiCE generates 23 counterfactual examples that suggest changes on at most two descriptive features.

4.4.2 | Hypothesis testing

In this work, we aim at studying the presentation of such explanations in an XUI for non-expert users. We also study the enhancement of these explanations with comparative analysis features. More precisely, the aim is to examine how effective they are to improve the explanation quality for these users. Similarly to the previous chapter and discussed in more details in Section 4.4.3, we consider two components for this explanation quality, distinguishing between objective understanding and satisfaction. More precisely, the study is driven by the following research questions and hypotheses:

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[2] https://scikit-learn.org/stable/index.html
[3] we follow the authors' implementation guidelines as documented on https://github.com/interpretml/DiCE

- **RQ1** : How effective are plural examples for improving understanding and satisfaction of counterfactual explanations for non-expert users?
  - H.1.1 : Plural counterfactual examples improve understanding, as compared to one example.
  - H.1.2 : Plural counterfactual examples improve satisfaction, as compared to one example.
  - H.1.3 : Comparative analysis on plural counterfactual examples improves understanding, as compared to one example only.
  - H.1.4 : Comparative analysis on plural counterfactual examples improves satisfaction, as compared to one example only.
- **RQ2** : How effective is comparative analysis for improving understanding and satisfaction of plural counterfactual explanations for non-expert users?
  - H.2.1 : Comparative analysis improves understanding of plural counterfactual explanations with plural examples, as compared to having plural counterfactual examples only.
  - H.2.2 : Comparative analysis improves satisfaction of plural counterfactual explanations with plural examples, as compared to having plural counterfactual examples only.

4.4.3 | Method

We describe in turn the participant recruitment, the evaluation material, the study procedure and the method to analyze the collected results. The method has been approved by the INSEAD Institutional Review Board (IRB). We pre-tested it with 2 participants to validate the understanding of the XAI interfaces and questionnaires presented in this section, and to adjust the vocabulary used in the questions.

4.4.3.1 | Participant recruitment

We recruited 112 participants from a large open network of volunteers at the INSEAD-Sorbonne University Behavioural Lab (in Paris, France), filtered to meet the requirements of our experiments, i.e., participants with little to no basic knowledge in AI nor in finance. Participants were aged from 19 to 39 (on average 25.5 ± 5.3), 73 were women and 39 were men, and there were diverse demographics (e.g., job position, level of study, previous experience in loan application).

To ensure the participants were non-experts in both AI and finance, we asked them to self-report their literacy for both topics on a 5-point Likert scale. We excluded the data of 1 participant who reported literacy scores between 4 to 5 at the end of the experiment, despite the initial filtering. After checking the data collected, we also excluded 2 participants who answered all open-response questions with in total less than five words.

The results analyzed in the next sections thus rely on the evaluation collected from 109 participants, randomly and evenly distributed across the three versions of the interfaces we propose (see Section 4.4.3.2). The participants were distributed in independent groups in a between-subjects setting, allowing us to compare (RQ1) the impact of the plural condition, and (RQ2) the impact of comparative analysis features, on the objective understanding scores and the satisfaction rates. All participants received a 6-euro compensation at the end of the experiment.

4.4.3.2 | Material

In this work, we adopt the similar approach as the one used in the previous chapter and adapt the material that we use for the user study. We describe in turn below the three tested interfaces, the questionnaires for objective understanding and satisfaction evaluation and the additional collected data.

Tested interfaces In this monitored experiment, we use three versions of our interface corresponding to the three conditions required for the hypothesis testing. More precisely, the different versions are designed as follows:

- Interface A is the baseline interface. It simply displays one counterfactual example with the card-based design described in Section 4.1.3. Figure 4.6 shows a screenshot of this version.
- Interface B is the interface with plural examples. It adds to interface A plural counterfactual examples, as described in Section 4.3. None of our proposed design principles are applied in this version. Figure 4.7 shows a screenshot of this version.
- Interface C, presented in Section 4.3.2, is the interface offering the features of comparative analysis on plural counterfactual examples. It adds to interface B the two principles of comparative analysis described in Section 4.2. Figures 4.4 and 4.5 show screenshots of this version.

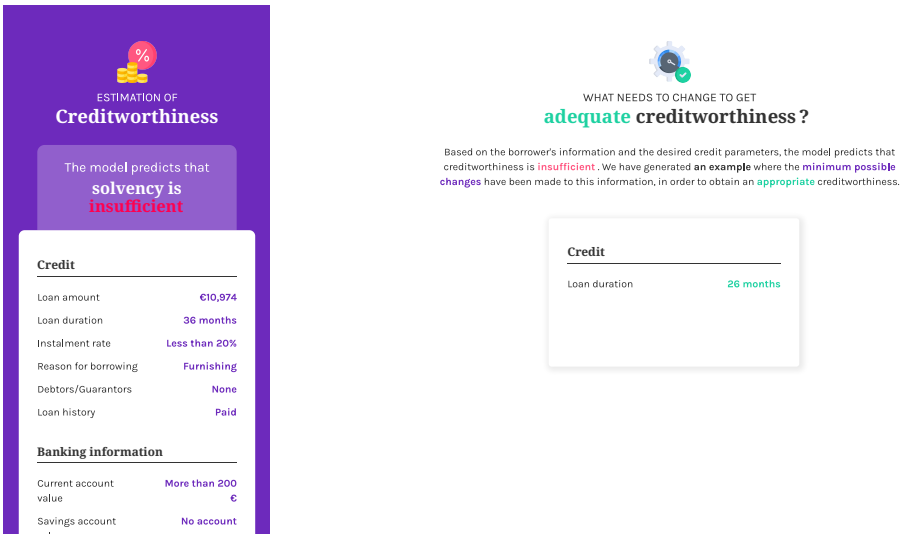


Figure 4.6: Interface A: baseline version with a single counterfactual example.

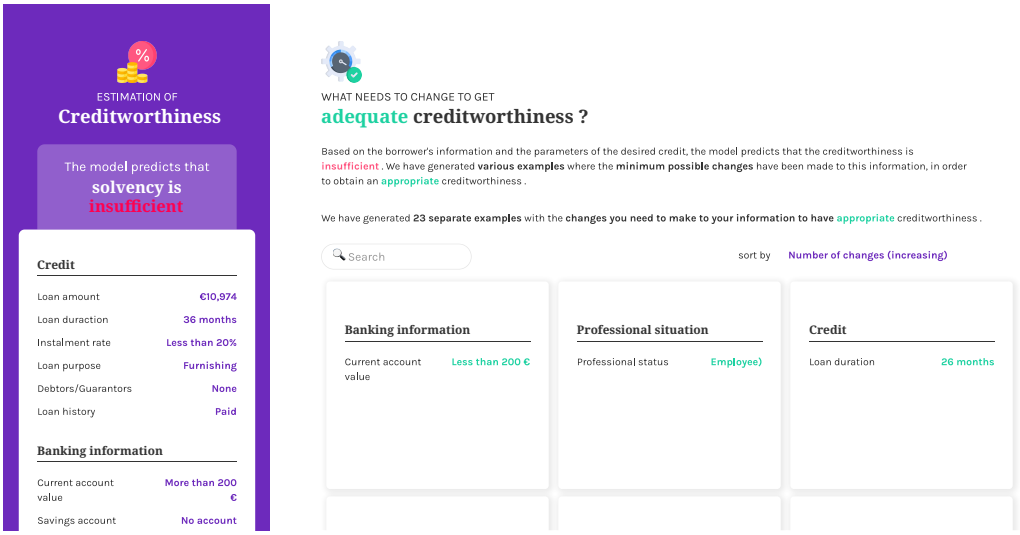


Figure 4.7: Interface B: proposed XUI for implementing counterfactual explanations with plural examples.

Objective understanding questionnaire We design a questionnaire with 14 statement questions (see questionnaire in Appendix B), for which users can either answer "I agree", "I disagree" or "I don't know". This questionnaire is an improved version of the one used in our previous work presented in Chapter 3, and aims to be generic for the assessment of objective understanding. We propose three types of questions to capture different components of user understanding when evaluating the intelligibility of XAI interfaces:

- *(i) Explanations' nature* questions measure the extent to which users understand what type of explanations is provided by a counterfactual example. More precisely, we check whether participants understand that the provided information is a counterfactual example. *i.e, whether they agree with the statement "The interface provided examples that suggest changes on the initial values that would have made the model predict a different solvency."*

- *(ii) Explanations' effects* questions measure the ability of users to understand how to interpret the explanation towards the predicted outcome. In our experiment, we measure participants understanding of the changed value as compared to the initial values. *e.g., "The model would have predicted a good solvency if the loan duration was reduced by 10 months."*

- *(iii) Explanations' specificity* questions measure the users' understanding of one complex component specific to the explanation provided. In our experiment, we measure participants understanding of the diversity in the generated counterfactual examples and how they compare them. Thus, these questions apply only to participants using interfaces with plural explanations. *e.g., "It is easier to reduce the loan amount than to change job position."*

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 questionnaire We use the same self-reporting questionnaire as the one used in the previous chapter, adapted from the Explanation Satisfaction Scale (Hoffman et al., 2018) in order to assess users' satisfaction.

Open-response questions In addition, we ask participant two open-response questions to qualitatively measure the intelligibility of the provided explanations. For the objective understanding, we ask participants *"what examples would they select to explain the predicted outcome"*. For the satisfaction, we ask participants *"if they are satisfied with the provided explanations"*. Participants can also share their insights and comments on the study in open-response questions. We perform a thematic analysis on answers for both questions and comments (Clarke et al., 2015).

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 finance, using 6-point

Likert scales, from “Not familiar at all” to “Strongly familiar”, to ensure that participants are indeed non-expert users.

Finally, we collect basic demographic information such as age, gender, education level and current occupation. We also ask participants their experiences with loan applications.

4.4.3.3 | Study procedure

We conduct the user study in a lab setting at INSEAD-Sorbonne University Behavioural Lab.

After giving written consent and prior to the experiment, participants are introduced to the following experimental scenario, translated and summarized from original language : "26-year old freelance graphic designer, Swann will be moving to a new place to work and live in Bordeaux, France. Swann is applying for a loan to the bank in order to fully furnish and equip this new apartment. Swann has previous experiences with loans (for studies first and travel then) and is confident that it will be accepted. Yet, Swann’s solvency is estimated as being not acceptable on the XAI platform used to submit the loan application and some explanations are provided".

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. While interacting with the interface, they take the objective understanding questionnaire, answer the subjective satisfaction questionnaire, and then answer the open-response questions and demographics.

4.4.3.4 | Data analysis

We use three versions of the interface described in Section 4.4.3.2 to answer these research questions presented in the previous section. More formally, we consider null hypotheses of the form "the considered condition or enhancement provides no significant improvement of the considered metric". To answer RQ1, we compare the two scores and answers (for objective understanding and satisfaction) for each of the two enhanced interfaces (interface B with plural examples, and interface C with comparative analysis on plural examples) as compared to the baseline interface (interface A with a single counterfactual example). To answer RQ2, we compare again the two scores and answers between the two enhanced interfaces (B and C).

As the preprocessings of the collected data show that it is normally distributed, we use one-way ANOVA to analyze (RQ1) the impact of the plural condition, and (RQ2) the impact of comparative analysis features. Table 4.2 displays the results for the scores and rates obtained in the experiment.

To answer the first research question towards the effectiveness of having plural examples (see Section 4.4.2), we use the seven questions from the objective understanding questionnaire presented in Section 4.4.3.2 that are relevant for this comparison, both on the nature and the effects of the explanations. As we compare the intelligibility of the explanations between participants having one counterfactual example (interface A) and participants having plural examples (interfaces B and C), we need to ask questions all participants can answer with the provided information. Thus, we focused on one counterfactual example that is provided on all interfaces. The first score of objective understanding score can vary from 0 to 7 corresponding to the number of correct answers for the 7 related questions of the questionnaire.

To answer the second research question towards the effectiveness of comparative analysis features (see Section 4.4.2), we use all the questions from the objective understanding questionnaire presented in Section 4.4.3.2. We compare the intelligibility of the explanations between participants having plural counterfactual examples (interface B) and participants using comparative analysis on plural counterfactual examples (interface C). This second score of objective understanding can vary from 0 to 14 corresponding to the number of correct answers for the 14 questions of the objective understanding questionnaire.

Finally, the user's satisfaction is reported from 1 to 6 corresponding to the average score over the eight satisfaction's dimensions presented in Section 4.4.3.2.

We use one-way ANOVAs to compare the difference between the independent groups. The significance level is defined as $\alpha = .05$. We use the Tukey post-hoc test to get adjusted p-values for multiple pairwise comparisons. Table 4.2 shows descriptive statistics for each group and their statistical significant differences.

4.5 | Results

We use the results presented in Table 4.2 and Figure 4.8, to answer the two research questions we consider regarding the plural condition in Section 4.5.1 and the comparative analysis in Section 4.5.2.

	Means (sd)	Interface A	Interface B	Interface C
RQ1	Objective understanding	4.16 (±1.5)	5.14 (±1.3)**	5.0 (±1.1)*
	ANOVA (as compared to A)	-	+.98 (p=.003)	+.84 (p=.01)
	Satisfaction	2.1 (±1)	2.5 (±1.1)	2.9 (±0.7)***
	ANOVA (as compared to A)			+.8 (p=.0009)
RQ2	Objective understanding	-	9.78 (±1.4)	9.66 (±1.4)
	Satisfaction rate	-	2.5 (±1.1)	2.9 (±0.7)

Table 4.2: Descriptive analysis of the results for the two objective understanding scores and the satisfaction rates, as well as the results of one-way ANOVAs (only when significant differences). For RQ1: we compare the scores and rates obtained for group B and group C to the ones for group A. For RQ2: we compare the scores and rates obtained between groups B and C.
Significance code: *** p<.001 ; ** p<.01 ; * p<.05

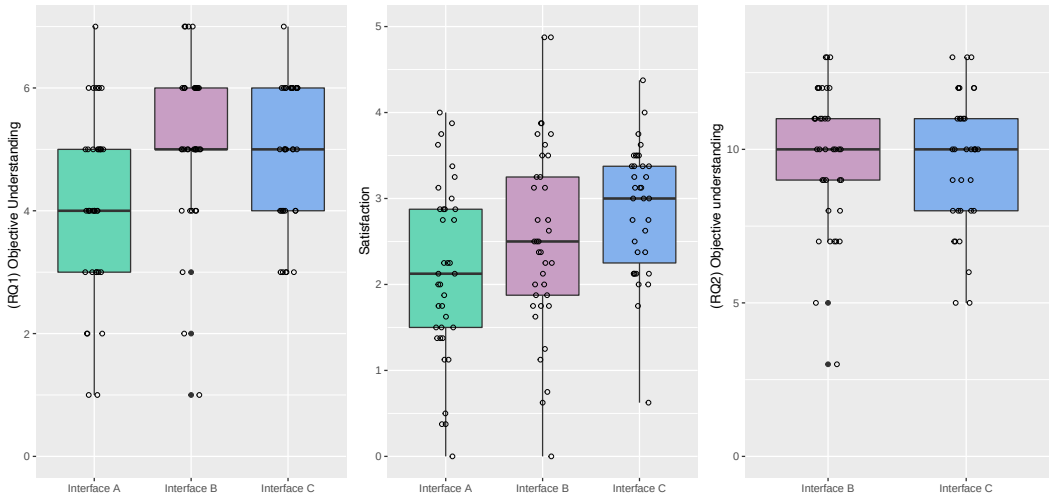


Figure 4.8: Measuring the intelligibility of the different versions of the interface: overview of (left) the objective understanding scores for evaluating the effect of the plural condition, (middle) the satisfaction rates for evaluating the effect of the plural condition and comparative analysis features, and (right) the objective understanding scores for evaluating the effect of the comparative analysis features.

4.5.1 | Plural condition

We measure the significant effectiveness of having plural counterfactual examples on users' objective understanding scores and satisfaction rates. The analysis of Table 4.2 leads to two main observations commented in turn below. First, having plural examples improves significantly objective understanding. Second, it also improves users satisfaction but there is only a significant difference when there are comparative analysis features.

Having plural examples improves significantly objective understanding Table 4.2 shows that interface B (plural counterfactual examples) has the highest improvement in objective understanding with an average score of 5.14 correct answers out of 7, i.e., .98 point more than interface A (one counterfactual example only). The one-way ANOVA shows that this difference is significantly higher (f(1)=9.18; p=.003). The Tukey post-hoc test also reveals significant pairwise differences between interfaces A and B (p=.005). In addition, we observe that participants interacting with interface C (plural examples paired with comparative analysis) obtain also higher scores for objective understanding with an average score of 5 out of 7, i.e., which is .84 point higher than for interface A. This difference is statistically significant (f(1)=6.76; p=.01) and the Tukey post-hoc test also reveals significant pairwise differences between interfaces A and C (p=.03).

Thus, we reject the null hypotheses as the scores for interfaces with plural counterfactual explanations are greater than the claimed value and conclude that **having plural examples in counterfactual explanations significantly improves objective understanding of non-expert users,** both with (H1.3) and without (H1.1) comparative analysis features.

Having plural examples improves satisfaction We observe that participants interacting with interface B give higher satisfaction rates regarding the provided explanations with an average rate of 2.5 out of 5, which is .4 point higher than participants interacting with interface A. Yet, this difference is not statistically significant. Participants interacting with interface C (plural examples paired with comparative analysis) also give higher satisfaction rates with an average rate of 2.9 out of 5, i.e., which is .8 point higher than for interface A. This difference is statistically significant (f(1)=11.82; p=.006) and the Tukey post-hoc test also reveals significant pairwise differences between interfaces A and C (p=.005).

Based on these observations, we fail to reject the null hypothesis and are not able to demonstrate the positive effect of having plural examples only on counterfactual explanations on users satisfaction (H1.2). Yet, we reject the null hypothesis as the average rate for the interface with comparative analysis features is higher than claimed value and conclude that **having plural examples when paired with comparative analysis significantly improves satisfaction of non-expert users** (H1.4).

4.5.2 | Comparative analysis

We measure the significant effectiveness of comparative analysis of plural counterfactual examples, as compared to plural counterfactual examples only, on users' objective understanding scores and satisfaction rates. We use here all 14 questions in the objective understanding questionnaire described in Section 4.4.1, and the same satisfaction rates as for RQ1. The analysis of Table 4.2 and Figure 4.8 leads to two main observations commented in turn below. First, having plural examples does not improve the objective understanding of plural counterfactual explanations. Second, it improves satisfaction but this difference is not statistically significant.

Comparative analysis on plural counterfactual explanations does not improve objective understanding Table 4.2 shows that participants using comparative analysis features (interface C) have slightly lower scores of objective understanding, with an average score of 9.66 out of 14, i.e., which is .12 point lower than for participants without these features (interface B). This difference is not statistically significant. Yet, when analyzing Figure 4.8, we observe that the minimum score for interface C is 2 point higher than for interface B.

Thus, we fail to reject the null hypothesis and are not able to demonstrate the impact of comparative analysis for plural counterfactual explanations on users' objective understanding (H2.1).

Comparative analysis on plural counterfactual explanations improves satisfaction but the difference is not significant When comparing the average rates for satisfaction among participants interacting with plural counterfactual explanations, we can see on Table 4.2 that those who are using comparative analysis features (interface C) rate their satisfaction higher, with an average rate of 2.9 out of 5, i.e., which is .4 point higher than the average rate of participants interacting with interface B. Yet, this difference is not statistically significant.

Again, we fail to reject the null hypothesis and are not able to demonstrate the positive effect of comparative analysis for plural counterfactual explanation on users satisfaction (H2.2).

4.5.3 | Qualitative analysis

In combination with the statistical analysis done on the participants' scores and rates, we also analyze their answers for the two open-response questions presented in Section 4.4.3.2. We conduct a thematic analysis with an iterative coding process (Clarke

et al., 2015): for each question separately, we analyze in an iterative process the answers without knowing the version of the interface that they were associated with, and identify codes. Then, we analyze the codes by versions of the interface and define themes for both objective understanding and satisfaction. We then use some participants answers to illustrate our observations (e.g., C2 refers to participant 2 who is interacting with interface C).

4.5.3.1 | Objective understanding

For the objective understanding open-response question, we identify 11 codes, and define 4 themes discussed in turn below.

Interpretation of counterfactual examples We identify codes related to the level of understanding of the counterfactual examples presented to participants. For each version of the interface, the same number of participants (between 12 and 13 in each group) understand that with the minimum suggested changes, the predicted outcome would have been different. Similarly, when having plural examples (interfaces B and C), the same number of participants (respectively 11 and 10) partially understand counterfactual examples. Most of these participants do not refer to the change values when suggesting modifications to the input values to get the loan accepted. Indeed, they all suggest features to change but only one participant in interface C provides the new value as suggested on the examples (C2 says "to lower the loan duration to 26 months, to lower the bank account value to 200 euros and to change the investment rate for 20% to 25%" as suggested on different examples). Finally, we observe 2 participants interacting with interface A who do not understand that the example suggested can be used to explain the predicted outcome, as well as for 1 participant interacting with interface C.

Personal beliefs We also identify codes related to participants' personal beliefs. Regarding interface A, most participants (21 out of 35) propose alternative explanations based on their own beliefs to explain the predicted class. In addition to the suggested change on the one example provided, participants propose additional changes based on the input data they have (e.g., A6 says that in addition to have a shorter loan duration, the applicant should "open a saving account, find a stable position and find a warrant for the loan"). For interfaces B and C, there are less participants who suggest personal beliefs' based explanations for the predicted class (15 participants for interface B, and 11 for interface C).

Feasibility of the examples For participants interacting with plural counterfactual examples (interfaces B and C), we identify codes related to the assessment of the feasibility for each suggested example. When disposing of comparative analysis (interface C), 10 participants are capable of selecting the most feasible examples to explain the predicted class. For interface B, only 4 participants are able to do so.

Association of different examples Finally, for participants interacting with plural counterfactual examples (again, interfaces B and C), we identify codes related to the ability of the participants to differentiate among the counterfactual examples. They understand that the examples can be used to explain the reject of the loan, yet most of them believe that the suggested changes from different examples can be associated (8 participants with interface B; 7 participants with interface C). For example, participant B1 believes that the best changes that would have made the model to accept the loan application are "to lower the amount of the loan, to find a new position and to wait 10 years", which are three changes on three different examples in the provided set of counterfactuals.

Review Overall, having plural examples seems to increase the intelligibility of counterfactual explanations and to reduce the inference with personal beliefs. Yet, it also can increase the risk of believing that the proposed changes can be associated. Adding comparative analysis features on counterfactual explanations with plural examples may reduce this risk, and allows users to better assess the feasibility of each suggested change. Thus, these observations lead us to believe that both having plural examples and comparative analysis are promising features to increase objective understanding of counterfactual explanations.

4.5.3.2 | Satisfaction

Similarly for satisfaction, we identify 20 codes and define 4 themes discussed in turn below.

Dissatisfaction First, we identify codes related to participants expression of satisfaction. More specifically, we identify three levels of satisfaction. The first level and most observed is expressed dissatisfaction of the provided explanations. Among the 65 participants who report they were unsatisfied, 27 of them are interacting with interface A, 19 with interface B and 19 with interface C. Participants report that the reasons why they are not satisfied are either because there is no explanation according to them (A20 says "I am not satisfied because there are no explanations provided")

or because the information provided is incomplete (A31 says "it misses more justifications, explanations and contextual information"). Some participants also blame the complexity of the explanations (e.g., C22 did not know "how to interpret" the examples, despite them being "very clear and detailed").

The second level expresses partial satisfaction. In total, 20 participants report they were partially satisfied with the explanations (7 for interface A, 7 for interface B and 6 for interface C). Most participants appreciate the clarity of the provided interfaces, but still believe that the explanations are too complex (B28 suggests to introduce better how to interpret the examples "so that it could be easier to understand why this value should change" for the model to accept the loan application).

Finally, the last level expresses satisfaction. Among the 24 participants who reported they are satisfied with the explanations, only 3 of them are interacting with interface A, 11 with interface B and 10 with interface C. In particular, participants report they like to get actionable changes (e.g., B10 says "We immediately understand which values we can change so that we can get the loan application to be accepted"). Overall, there are more participants satisfied with the explanations in interfaces B and C as compared to interface A.

Missing explanations We identify codes related to missing content in the presented explanations. For participants interacting with interface A, 11 of them feel that there is no explanations provided (as mentioned above, A20 says “there are no explanations provided"). For interfaces B and C, the number of participants who share similar opinion is lower (2 for B, and 2 for C).

Expressed needs Also, we identify codes related to needs explicitly expressed. Whether participants are satisfied or not with the explanations and no matter which interface they are interacting with, 36 participants say they need further information or details about the provided information. For example, participant C16 says that "there should be more detailed information for some examples [...] that are counterintuitive". In addition, 12 participants say they need more contextualization of the provided explanations. Participant A32 says that "this information could be further explained and detailed so that the borrower understands what aspect of his/her application is problematic". Finally, participants interacting with interface C express some additional needs, such as the need to have human contact (1 participant), more transparency over the model (1 participant), and other explanations such as the weight of features on the predicted solvency (1 participant).

Perceived complexity We identify codes related to the complexity of the provided information. The expressed complexity is different from one version of the interface to another. For interface A, 9 participants say that the provided example is either "difficult to understand", "not intelligible" or "not feasible" according to them.

For interface B, 17 participants report that the explanations are complex. Most reported complexity regards the organization of the provided examples. For example, participant B28 says that "the 23 examples are a bit scattered all over the interface and could have been grouped by category (bank information, duration of loan...)". These insights are particularly valuable to us as we aim at addressing this issue with the comparative analysis in interface C. Others report that the examples are also difficult to understand, and that all examples are not always feasible.

For interface C, also 17 participants report that the explanation are complex. More specifically, 7 participants report that the explanations are difficult to understand completely because of lack of knowledge in the applied domain. Participant C12 says that "without previous knowledge in loans, it is difficult to understand the reasons why some examples are proposed." Also, 2 participants say that it is difficult to understand how to interpret the explanations because of the plurality of examples (for example, participant C23 says that "the accumulation of cards make it difficult to use the platform"). Moreover, 2 participants say that the explanations are counter-intuitive: participant C16 says that some "suggested changes" are "unclear and counter intuitive". Finally, other participants question the feasibility of the suggested examples and report that the latter are not well organized on the interface.

Review Overall, these observations lead us to believe that participants are mostly unsatisfied with the provided explanations. They are even more unsatisfied when there is only one example suggested. Reasons are multiple: the provided information does not act as explanations, there is a need for more details and justification about the suggested examples, as well as for contextualized information, and it is difficult to interpret these examples. We believe that participants are more inclined to consider one counterfactual example does not constitute an explanation. Yet, the complexity seems to be increased when there are plural examples. Participants suggest that the explanations would need to be better organized, and that they would need to be guided on how to analyze and interpret them. These insights encourage us to believe that comparative analysis features are promising tools to improve the intelligibility of such explanations.

4.6 | Conclusion

In this work, we investigate the intelligibility of explanations expressed in the form of plural counterfactual examples that are presented to the non-expert users. This work contributions are, first, a process for designing and evaluating an XUI for such explanations. We investigate (i) if plural counterfactual examples are indeed better than having a single one, and (ii) if we can mitigate the users' confusion through a comparative analysis enhancement when there is a high number of examples. We propose an implementation of such enhanced explanations in an XUI for a financial scenario related to a loan application. We perform quantitative and qualitative evaluations of the collected data. In the quantitative analysis, the results show that having plural examples does improve significantly the objective understanding of counterfactual explanations, as compared to having one example only. It does also improve the satisfaction, but this difference is significant only for the interface offering comparative analysis features. On the contrary, the comparative analysis features do not appear to improve significantly neither the objective understanding nor the subjective satisfaction of plural counterfactual explanations. Yet, the qualitative analysis of the collected open-response answers shows that they may reduce the inferences with personal beliefs and help the users to better assess the feasibility of the suggested changes. These observations lead us to believe that the comparative analysis features are promising tools to improve the intelligibility of counterfactual explanations for non-expert users. These results are of course dependent on the quality of the explanations generated by the machine learning explainer model in the first place, prior to the question of the presentation.

While we show in this chapter that having plural examples and offering comparative analysis feature can improve the intelligibility of counterfactual explanations, there are some limitations to our work which are important to mention.

We acknowledge that the DiCE method (Mothilal et al., 2020) is not adapted to the Random Forest model we trained with sklearn tool, resulting in presenting the participants with surprising counterfactual explanations. For instance, according to one of the provided example, a huge change in the loan amount would be needed to yield a positive outcome which is quite unrealistic. This might explain why some participants are unsatisfied with the explanations and find it difficult to interpret them. We believe there is room for improvement on that point and consider as future works conducting additional user experiments applying the same protocol with other configuration of the explainer as well as other explainer models.

Moreover, in the considered settings, DiCE generates 23 counterfactual examples

based on the objective to provide users with maximum diversity in the built explanation. We believe that another experiment with fewer counterexamples on a similar enhanced XUI would be also an important topic to address.

Future works will also aim at investigating new modalities to evaluate the objective understanding and the satisfaction, in particular extending the conducted study with qualitative methods for analyzing the collected results, such as interrater reliability. We believe a qualitative evaluation might help to have move comprehensive view of what the users understand or not about the provided explanations, as well as their points of satisfaction or disappointment when using the XUI. Other directions for refining the conducted study will focus on other possible effects of interest. The latter for instance include a possible correlation between objective understanding and subjective satisfaction, the scores per type of questions we ask to the participants for the evaluation of the objective understanding, or the users demographics.

5

An ontology of inconsistencies in ML explanations

At a fundamental level, this chapter consider the issue of inconsistency within ML explanations from a theoretical point of view. Several issues have been reported in the literature in an ad hoc manner: the information provided by XAI approaches does not always have the expected impact on users' understanding, meaning they do not necessarily provide consistent and faithful explanations of the predicted outcome or the ML model behavior. For instance in the previous chapter, we observe the possible link between poor explanations and explanations that are unconvincing. We believe that these inconsistencies in ML explanations can come from both the ML system itself and the inferences the users make with the presented explanations.

We propose to identify different sources of inconsistencies in ML explanations and their potential impact on their interpretation. We aim at understanding the various forms they can take and we propose to organize them within an ontology, as presented in Section 5.2 and summarized in Tables 5.1 and 5.2. This ontology is based on a literature review of recent papers on XAI that have identified limitations to the intelligibility of ML explanations.

This chapter is structured as follows: in Section 5.1, we discuss limitations for the intelligibility of ML explanations and present and overview of the ontology that we propose; then, we present in details the various inconsistencies we propose to distinguish for this ontology in Sections 5.3 and 5.4. Section 5.5 concludes the paper and discusses its potential implications.

5.1 | Motivations

As presented in Section 2.2 and illustrated in Figure 2.7, the XAI framework puts into play two actors of a different nature, namely a machine learning system and a human user interacting with the latter. The quality of this framework output depends both on the ability of the machine to provide a relevant explanation, and on the ability of the user to interpret it. There are thus two potential sources of limitations for the quality of ML explanations, as discussed in details in the next sections,

First, it appears that the information extracted by XAI approaches is not always coherent, which can make it hard for AI practitioners to build intelligible explanations and for designers to build consistent XUIs. Such a consistency depends on the ability of the machine to provide a relevant explanation, which is a complex notion that has driven many discussions (Holzinger et al., 2020; Zhou et al., 2021; Jesus et al., 2021). We focus in particular on the dimensions of faithfulness and accuracy of explanations in relation to the predictive model. Hence, we investigate the technical limitations of the ML system itself for the generation of such explanations, either from the model or/and from the implemented explanation method.

Second, even in the case where the generated explanations are coherent and faithful to an accurate prediction, users can make different interpretations. An explanation will not be interpreted the same way from one user to another, because each of them is unique based on prior knowledge, past experiences, specific needs for example. Hence, we investigate the explanatory limits that are specific to the user.

Overall, there can be various types of such "explanation failures", which we call inconsistencies, when generating and displaying ML explanations, that can lead to confusion, mistrust or erroneous interpretation and conclusion by the end user. We believe that inconsistencies in ML explanations is an important topic to tackle when considering one of the main challenge in XAI towards the intelligibility of explanations, and to avoid explanations pitfalls.

5.2 | Overview of the proposed ontology

The methodology applied to establish the proposed ontology of explanation failures relies on a literature review of recent papers on XAI approaches, interfaces and evaluations: it first collected the issues they point at, listing the limitations to the intelligibility of ML explanations they highlight. We then proposed to structure them in a three-level hierarchical structure, graphically represented in Figure 5.1 that provides

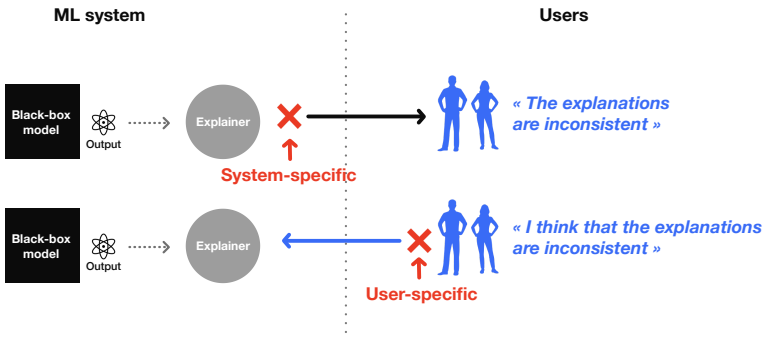


Figure 5.1: Inconsistencies in ML explanation. In the context of an interaction between a ML system and users, we propose to distinguish between inconsistencies that come from the technical limitations of the ML system (system-specific inconsistencies) and inconsistencies that come from the inferences the users make of the explanations provided by the ML system (user-specific inconsistencies). Schematically, issues occur at the red cross position, either on the ML side or the user side.

a visual representation of the first level, whose two types are then detailed in Figures 5.2 and 5.3 and discussed in turn in the next sections. In each case we provide information on the potential effects for the intelligibility of such inconsistent explanations with examples from the literature. We believe this structured view can be helpful to improve XAI approaches and avoid explanation pitfalls.

As illustrated in Figure 5.1, we first propose to classify explanation inconsistencies into two categories: the first one, we call system-specific, come from technical limitations of the ML system, the second one, we call user-specific, come from the inferences users make about the provided explanations, that are specific to each user.

First, system-specific inconsistencies may come from all components of the ML system: from the ML model (e.g., when the prediction is inaccurate), from the explainer (e.g., when there are issues with the implementation of the explainer) or from their combination the ML system (e.g., when the provided explanations are not faithful). We summarize these limitations in Table 5.1, and discuss them in details in Section 5.3. We believe that the system-specific inconsistencies are in fact a design problem of ML system and that further research should be conducted to minimize the risks of confusion in users' perception of such explanations.

On the other hand, we propose to identify common misinterpretations of ML explanations by the users, that we define as user-specific inconsistencies. When a black-box ML model provides an accurate prediction and an explainer provides a faithful explanation, the latter can still be misinterpreted by the users. We summarize these users' inconsistent inferences and their effects in Table 5.2, and discuss them in de-

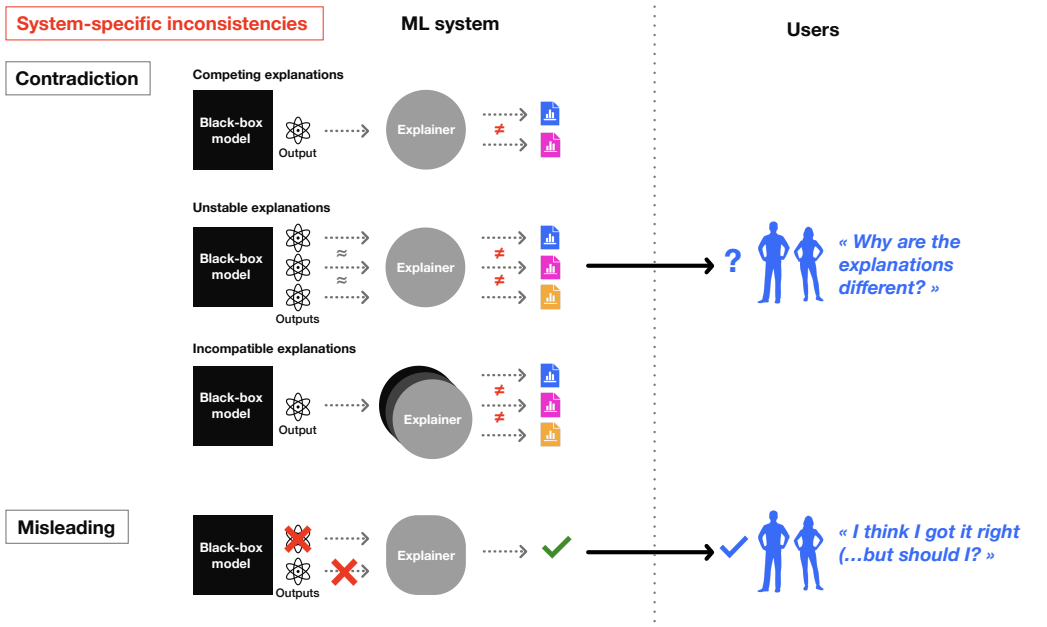


Figure 5.2: Two types of system-specific inconsistencies in ML explanations: information contradictions (competing, unstable or incompatible explanations) coming from the explainer; and misleading explanations coming from the ML system.

tails in Section 5.4 illustrating them with examples from the literature. We believe that the user-specific inconsistencies should be known so XAI designers can understand the users' mental model processes and support their interpretation of the provided explanations.

5.3 | System-specific inconsistencies

Regarding the explanation failures that come from the system itself, illustrated in Figure 5.2 and summarized in Table 5.1, we propose to distinguish between two types, respectively called *contradictory* and *misleading* explanations, and discuss them in turn in the next subsections.

5.3.1 | Contradictory explanations

The first system-specific inconsistencies we propose to identify are related to conflicting information provided by an explainer, or several explainers, that are supposed to provide faithful information with respect to the ML model to explain. More pre-

cisely, these inconsistencies occur when the ML system provides several explanations or explanation parts that differ one with another, as illustrated in the next sections: one consequence of this kind of limitation is that the users do not understand the differences in this conflicting information, and wonder *"Why is it different?"*.

We propose to distinguish between three types of such contradictions, depending on the origin of the explanation variety, that are discussed in details in the following. In the first case, that we call *competing explanation*, the explanation generated by a single explainer for a single instance is made of several components that may contain contradiction. In the second case, that we call *unstable explanations*, the contradiction comes from explanation generated by a single explainer for several instances. In the third case, that we call *incompatible explanations*, the contradiction comes from explanations generated from multiple explainers for a single instance.

5.3.1.1 | Competing explanations

The first case of contradictions we identify occurs when the explanation generated by a single explainer for a single instance is made of several components that contain contradictions. It is represented visually on Figure 5.2 with a black box model outputting a single prediction (depicted as an atom) that is transferred to a single explainer (represented as a disk) that generates several results (depicted as histograms). Note that the same principle applies to the case of global explanations that provide information about the ML model's behavior, when these explanations also break down to several components.

This case can for instance occur when the explanation is composed of multiple, and usually diverse, counterfactual examples (Mothilal et al., 2020; Rodriguez et al., 2021; Suffian et al., 2022). The latter are then competing by definition: they can suggest contradictory modifications to the considered instance so as to change the associated predictions, either involving different features or different variations on similar features. Depending on the set of involved features, the modification of the values for one of them may even apply to opposite directions. They may make their interpretation confusing for end users. For instance in the previous chapter, DiCE (Mothilal et al., 2020) generates some inconsistent counterfactual examples, suggesting contradictory directions for the amount of savings (i.e., increasing or decreasing their value). In another example for image data, current methods for generating multiple counterfactual explanations are limited to small contiguous regions of features with high influence on the target model outcome (Rodriguez et al., 2021). In doing so, there are increased chances for these explainers to generate competing changes on the input,

resulting in generating more confusion for the end users (Suffian et al., 2022).

In the case of explanation in the form of feature importance scores, the computed weights often consider the features independently one of another (Ribeiro et al., 2016). Yet, some of these weights can be impacted by hidden correlations between features (Slack et al., 2020), which can result in competing feature weights between some features. For example in Chapter 3 for a smart insurance pricing scenario, the SHAP method can be used to provide explanations on the predicted price. For a car insurance for a 52 year-old mother and her 21 year-old daughter, the explanations provided gives a significant weight for the mother's age and an insignificant weight (close to zero) for the daughter's age (although the lack of experience in driving should have a higher impact). Hence, the expected influence of the daughter's age has potentially been shifted to the one of the first driver (i.e., the mother). In this case, this contradiction can be explained by the current limit of such interpretability method that cannot handle correlation between features.

5.3.1.2 | Unstable explanations

The second case of contradictions we propose to identify are so-called unstable explanations, it may be encountered for local explanations but not global ones. In this case, a single explainer, that provides single faithful explanations, is applied to several similar instances with similar outcomes. It can be expected that the explanations should be similar, which is known as a robustness requirement of the explainer (Mishra et al., 2021; Alvarez Melis and Jaakkola, 2018). Yet, it can be the case that the explainer does not meet this requirement and extracts different pieces of information from one instance to another, which results in user confusion.

In particular, experimental comparison of explanation generated by an explainer under small perturbations of the input values of an instance have shown that popular methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) can lead to such unstable explanations. It has also been demonstrated that choosing an adequate sampling strategy for generating the instances used to fit such a surrogate model has a major impact on the quality of the approximation of the local black-box decision boundary and thus on the accuracy of the generated explanation (Laugel et al., 2018b).

When the model input remains unchanged, it has been observed as well that current model-agnostic methods can provide different explanations when modifying the underlying model (e.g., adversarial model manipulation) (Heo et al., 2019). A related issue is that of explanation fairwashing (Dimanov et al., 2020; Mishra et al., 2021;

Description	Examples	Effects	References
Contradiction (location: explainer)			
Competing explanation For a single instance, one explainer method can generate competing explanations. These explanations are all faithful but vary at some levels, creating information contradiction	Multiple counterfactual examples can suggest competing variations for the same attribute's value (Suffian et al., 2022)	Users can find it confusing and it can limit their ability to build an accurate mental model of the ML system	Slack et al. (2020) Rodriguez et al. (2021) Suffian et al. (2022) Bove et al. (2022)
Unstable explanations An explainer can extract different information for two similar instances with similar ML predictions, making the explanation process unstable.	LIME and SHAP can generate different explanations for one instance, due to small perturbations (Alvarez Melis and Jaakkola, 2018)		Alvarez Melis and Jaakkola (2018) Heo et al. (2019) Anders et al. (2020) Dimanov et al. (2020) Mishra et al. (2021) Visani et al. (2022)
Incompatible explanation For one instance, various types of explainers (e.g., feature importance, rules, example-based) can generate explanations that do no agree	Local and global feature importance scores can be different for a single instance (Harel et al., 2022)		Olah et al. (2018) Wang et al. (2019) Barredo Arrieta et al. (2020) Harel et al. (2022)
Misleading (location: ML system)			
The ML system provides users with a convincing explanations that do not reflect potential issues within this system.	The ML model outputs an erroneous prediction and the explainer gives faithful but misleading explanation (Papenmeier et al., 2019)	Users can be fooled by these explanations, misleading their decision making processes.	Guo et al. (2017) Guidotti et al. (2018) Papenmeier et al. (2019) Jacovi and Goldberg (2020) Dimanov et al. (2020) Ye and Durrett (2022)
	The ML model gives an accurate prediction but the explainer extracts erroneous, yet convincing information (Ye and Durrett, 2022)		

Table 5.1: Ontology of common system-specific inconsistencies in ML explanations, due to technical limitations of the ML system. Each identified type of inconsistency is described and illustrated with examples and references from the literature, with their potential effects on the intelligibility of the ML explanations

Anders et al., 2020), that raises ethical issues about the explanation generation task: experiments have shown that is is possible to train a model that has the same prediction as a reference one, and thus the same accuracy, but that leads to explanations hiding the bias, i.e., the role of sensitive features.

5.3.1.3 | Incompatible explanations

The third case of contradiction we propose to identify are called incompatible explanations and arise in a different setting, when several explainers are used to generate explanations for a given instance (local explanations) or machine learning model (global explanations). Now there is no guarantee that these explanations agree one with another. The notion of agreement may be not straightforward to define in the case where the explanation types differ, for instance for counterfactual examples and feature importance vectors, but they are for a given type. As the variety of explainers within a type aims at offering different properties, they usually provide different results. For instance the weight scores computed by LIME and SHAP do not always lead to the same feature importance ordering.

On one hand, providing users with interfaces that present various types of explanations can be particularly important when the decision to make may have a considerable impact. For example, for a medical diagnosis AI-based tool, some doctors have reported that they need to understand both the weight of each symptom on the predicted disease, and they also need to confirm one diagnosis by comparing it with similar and alternative diagnosis for similar instances (Wang et al., 2019). Yet, current methods have been studied in isolation, meaning that the explanations they provide are not necessarily compatible (Olah et al., 2018; Barredo Arrieta et al., 2020). For example, Harel et al. (2022) show that local and global feature importance scores can be different for a single instance.

5.3.2 | Misleading explanations

We also propose to identify **misleading explanations** as another system-specific inconsistency in ML explanations as described in Table 5.1 and illustrated in Figure 5.2. The inconsistency concerns one of the two cases marked with a cross: either regarding the prediction or the implementation of the explainer. Hence, we consider two scenarios where the ML system is dysfunctional. In the first one, an explainer provides accurate explanations for an erroneous prediction given by the ML model. In the second one, the prediction is accurate and the explainer gives a convincing but false explanation. Either way, the ML system retrieves an explanation that can be

misleading (e.g., users believe they can trust the explanation and use it to make a decision), which increases the risk of misinterpretation (e.g., the explanation is false and should not used by users).

An important criteria for the explanations' quality is their faithfulness to the ML model (Guidotti et al., 2018). However, ML models can output confident but incorrect predictions (Sanchez et al., 2022), and an explainer can be faithful to this model and generate accurate explanations. In such context, these explanations can be useful for calibration (Guo et al., 2017; Ye and Durrett, 2022), but they can also be also harmful for the users with low levels of awarness (Jacovi and Goldberg, 2020; Papenmeier et al., 2019).

On the other hand, an explainer can also generate convincing but false explanations for an accurate prediction. For example, it has been demonstrated that the explanations generated by large language models can be unreliable, even for a very simple synthetic dataset (Ye and Durrett, 2022). When measuring model's fairness, it has also been demonstrated that explanation attacks can mask a model's discriminatory use of a sensitive feature (Dimanov et al., 2020). Again, these explanations can also be harmful for the users who will not be able to perceive their inaccuracy.

5.4 | User-specific inconsistencies

In the ontology that we propose, we also identify inconsistencies coming from the users, as illustrated in Figure 5.3. As opposed to the system-specific inconsistencies, we assume that the ML system here is functional: a ML model (depicted as a black-box) provides an accurate prediction (depicted as an atom) and an explainer generates faithful explanations (depicted as a disk). In other words, we make the assumption that there are no system-specific inconsistencies. However, each user makes his/her own interpretation of the ML explanations, and not all will have the same difficulties. This can lead to inconsistencies that come from either the inferences the users make of the explanations (depicted as a cross next to the users) or a mismatch in objectives (depicted as a cross between the users and the ML system). We propose to distinguish between three types of user-specific inconsistencies: in the two first case, which we propose to call counter-intuitive inferences, biased reasoning and mismatch inferences respectively, and discuss in turn in Sections 5.4.1, 5.4.2 and 5.4.3.

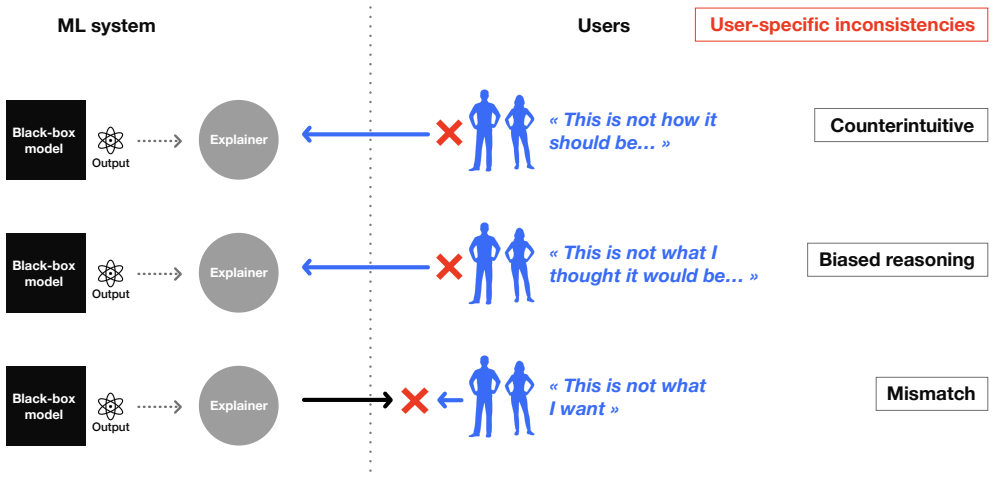


Figure 5.3: User-specific inconsistencies in ML explanations: counter-intuitive explanations, biased reasoning and mismatch explanations. On one side, the ML black-box model gives an accurate prediction and an explainer generates faithful explanations. On the other side, the users receive the explanations. Here, users are perceiving inconsistencies in ML explanations despite the accuracy of the ML system.

5.4.1 | Counter-intuitive explanations

First, we propose to identify **counter-intuitive explanations** when the provided information does not support what users have learned or experienced in the past. There are many definition of prior knowledge in the literature, in various domain such as cognitive psychology and artificial intelligence (e.g., see the discussion proposed by Dochy and Alexander (1995)). As we analyze the users' perceptions of ML explanations, we consider here the following definition: "stored knowledge about the world that have been acquired by an individual" (Brod et al., 2013). In a ML context, the explanations can differ from the users' acquired knowledge and past experiences, which can make the latter question or reject these explanations.

First, the users' acquired knowledge can have an impact on the perception of the explanations provided by the ML system (Nourani et al., 2022). We refer to acquired knowledge for the objective information and/or skills that one individual has learned in the past, such as domain expert knowledge. In the literature, it has been demonstrated that domain experts sometimes disagree with these explanations. In the environmental area, some experts were skeptical about using Black-Box ML models when the explanations for one prediction do not correspond to their expert knowledge (Palaniyappan Velumani et al., 2022). In another study, it has been argued that fraud agents or actuaries can disagree with a prediction when the explanations pro-

vided by the ML system are counter-intuitive as compared to their knowledge (Collaris et al., 2018). Moreover, past experiences can also have an impact on the perception of ML explanations. For example, users may expect that there can be intuitive changes in counterfactual examples because they have experienced the same logical path in real life (e.g., in a loan application, if the users can increase the income, then they can increase the average credit card usage and increase the mortgage too (Suffian et al., 2022). In social sciences, researchers argue that people tend to ignore an information that is inconsistent with their beliefs from past experiences (Thagard, 1989; Nickerson, 1998). Researchers in neuroscience have also identified patterns of brain activity that underline the users' ability to interpret based on past experiences (Sohn et al., 2019). Thus, we believe that the users may perceive some explanations as being absurd as they do not correspond to what they would have expected based on past experiences. Some works in XAI that have identified this problem propose to fix it by integrating the users' knowledge directly in the generation of explanations (Ustun et al., 2019; Mahajan et al., 2019; Jeyasothy et al., 2022).

5.4.2 | Biased reasoning

Second, we propose to identify **biased reasoning** when the users make inaccurate inferences of accurate and faithful ML explanations. In this context, we consider that users do not have prior knowledge and build their understanding of the explanations during the interaction with the ML system. In the literature, researchers demonstrate that despite having faithful explanations, some users with no prior knowledge and no past experiences can misinterpret the explanations (Cheng et al., 2019; Cai et al., 2019; Zhou et al., 2021), and overrate the depth of their knowledge when interacting with such system (Chromik et al., 2021). Other works also show how people's biases can have an impact on explanations interpretation Miller (2019); Bertrand et al. (2022); Nourani et al. (2021); Gajos and Mamykina (2022); van der Waa et al. (2021).

When lacking prior knowledge, users build a mental model of a ML system based on the interaction they have with this system. Yet, their mental models can be inaccurate despite the faithfulness of the explanations provided. In cognitive sciences, it has been demonstrated that people often form an inaccurate understanding of complex systems and often overrate the depth of their knowledge Mueller et al. (2019). For a university admission explanation interface, both the objective and the self-reported understanding of non-expert participants are measured, and the results show differences in both scores: some are higher for the self-reported understanding than for the objective one (Chen et al., 2021b). Similar results are found in Chapter 3 for the

enhanced XUI we propose for the non-expert users. It has also been demonstrated that people can fall into the illusion of explanatory depth when interacting with ML explanations: they can form false or incomplete interpretations of the explanations and believe they understand better than what they actually do (Rozenblit and Keil, 2002; Chromik et al., 2021). Overall, these findings implies that people might over- or under trust ML explanations (Rudin, 2018; De Visser et al., 2020; van der Waa et al., 2021)

These disparities in understanding of explanations may also result from the impact of cognitive differences between users. People's cognitive biases on the interpretation of ML explanations have also been studied (Miller, 2019; Bertrand et al., 2022). In particular, it is argued that explanations can bias the users and impair their decision-making process, and that these biases vary from one individual to another, making it challenging for expecting a uniformed interpretation of information provided.

We argue it is important to understand the reasons of the gaps between the objective and perceived understanding so researchers can improve XAI techniques accordingly.

5.4.3 | Mismatching explanations

Finally, we propose to identify **mismatch explanations** when the format of the extracted information is not meeting the users' expectations towards the ML system. These expectations can vary depending on the context of the interaction, the background knowledge of the users or the output of the model, which makes it challenging to select the adequate XAI methods.

Depending on the context of the interaction, users have different needs and questions regarding the ML system, as discussed in Section 2.2.4.2. Hence, the explanations generated may vary from one approach to another, and not all of them answer the same question. User questions usually focus on specific needs, and only a limited number of XAI approaches can be used to answer them. During the co-design workshop for an AI-based diagnosis tool with doctors (Wang et al., 2019), many of the latter reported that they would prefer alternative hypotheses (e.g., counterfactual examples) rather than factual explanations (e.g., local feature importance). These user questions may also vary depending on the model's output. For example, the Expectation Confirmation Model (Bhattacherjee, 2001) postulates that user satisfaction and acceptance of a system is directly related to the difference between initial expectations and their actual experience. For an imperfect AI powered email scheduling assistant

Description	Example	Effect	References
Counter-intuitive Explanations do not support what the users have learned or experienced in the past	The provided information is contrary to what the domain expert would expect (Collaris et al., 2018)	The users may reject the explanations, or question the trustworthiness of the ML system	Thagard (1989) Nickerson (1998) Collaris et al. (2018) Sohn et al. (2019) Nourani et al. (2022) Palaniyappan Velumani et al. (2022)
Biased reasoning Users with no prior knowledge make inaccurate interpretations of the explanations.	People can fall into the illusion of explanatory depth (Chromik et al., 2021) The explanations can bias the users and impair their decision-making process (Bertrand et al., 2022)	Users with no prior knowledge may overtrust or undertrust the explanations, independently from the quality of the ML system	Rudin (2018) Cai et al. (2019) Miller (2019) Mueller et al. (2019) De Visser et al. (2020) Chen et al. (2021b) Zhou et al. (2021) Nourani et al. (2021) Chromik et al. (2021) van der Waa et al. (2021) Bove et al. (2022) Bertrand et al. (2022) Gajos and Mamykina (2022) Rozenblit and Keil (2002) Suffian et al. (2022)
Mismatch The extracted information do not correspond to need of the users in terms of explanations.	Depending on their needs, users can have different why-question regarding the ML system that requires specific XAI techniques for each	Users reject the ML system as it does not deliver the expected performances.	Wang et al. (2019) Liao et al. (2020) Riveiro and Thill (2021)

Table 5.2: Ontology of common user-specific inconsistencies in ML explanations, due to user inferences. Each identified type of inconsistency is described and illustrated with examples and references from the literature, with their potential effects on the intelligibility of the ML explanations.

(Riveiro and Thill, 2021), it has been demonstrated that adjusting the explanations according to users' expectations improves satisfaction and acceptance. Background knowledge of the users may also be important to consider when selecting an XAI approach. For example, users who lack skills in AI can find it hard to interpret some visual explanations like partial dependency plot (PDP) or individual conditional expectation (ICE) graphs. Recent work also demonstrates that lexical alignment improve the understanding an explanation provided by a conversational agent (Srivas-

tava et al., 2023).

Overall, understanding both the users needs and adapting the explanations accordingly would help avoiding mistmatch explanations.

5.5 | Conclusion

In this chapter, we propose an ontology of common inconsistencies in ML explanations. We review recent works in XAI and identify various limitations for generating intelligible explanations to the end-users. We propose to distinguish two sources of inconsistencies: system-specific inconsistencies due to technical limitations of the ML system, and user-specific inaccurate inferences the users make of the ML explanations. We propose to identify various types of inconsistencies from both sources, and describe them in details with examples from the literature and potential effects on the users' understanding. We believe this ontology can help AI practitioners better understand current limitations when generating ML explanations for a ML system and avoid explanation pitfalls.

We also discuss some challenges in XAI to improve the quality of ML explanations. Future works include a deeper understanding of each type of inconsistency with additional use cases and user studies to identify their effects on the users' understanding. We also believe that studying the combination XAI techniques would be beneficial to mitigate gaps in explanations and help to build more complete ones for the users. Finally, user studies are needed to understand users' variation in interpretation of ML explanations.

6

Conclusion and perspectives

Summary of the contributions

In this thesis, we investigate the intelligibility of Machine Learning in the following context: on one side, an opaque classifier or a regressor provides a prediction, and an XAI post-hoc approach generates pieces of information as explanations; on the other side, users receive both the prediction and the explanations.

We address several issues that can arise in such a context, and that might limit the quality of the explanations: the lack of contextual information in ML explanations, the unguided design of functionalities or the user's exploration, as well as confusion that could be caused when delivering too much information. All these issues make it difficult for users to have an accurate interpretation of the explanations, especially when they have little to no knowledge in AI nor in the applied domain. To address these issues, we develop an experimental procedure to design and evaluate intelligible ML explanations. We identify several opportunities for XAI enhancements, and turn them into generic XAI principles that are implemented into explanation user interfaces. The latter are then used to evaluate intelligibility of these explanations through user studies, measuring both objective understanding and subjective satisfaction. In particular, we investigate two types of XAI approaches: local feature importance in Chapter 3 and plural counterfactual examples in Chapter 4.

For local feature importance, we propose XAI principles for contextualization and exploration. Moreover, we propose an implementation of these principles into an XUI for an insurance scenario. Finally, we use this enhanced XUI to conduct a user study in a monitored lab setting with 80 non-expert participants and evaluate the effectiveness of such enhancements on the two dimensions of the intelligibility. The quantitative analysis of the results demonstrate that contextualization principles sig-

nificantly improve user's satisfaction and are close to significantly improve user's objective understanding. Also, the results show that the exploration principles significantly improve user's satisfaction.

Similarly, we adapt this experimental process for a second type of explanations in the form of plural counterfactual examples. We propose XAI principles for integrating comparative analysis tools, and their implementation into an XUI for a financial scenario. We use this enhanced XUI to conduct another user study with 112 non-expert participants. We evaluate the effectiveness of the plural condition and comparative analysis principles on the two dimensions of the intelligibility of counterfactual explanations. The method used for the evaluation is similar to the one proposed in the first case, and adapted to this context. The quantitative analysis of the results shows the effectiveness of the plural condition, both on objective understanding and satisfaction scores, as compared to having a single counterfactual example. The qualitative analysis shows that the proposed comparative analysis features are promising approaches to improve the intelligibility of such explanations, even if the participants partially report they are not satisfied by counterfactual explanations, as they perceive them as incomplete and too complex.

At a fundamental level, we consider the issue of inconsistency within ML explanations from a theoretical point of view. Several issues have been reported in the literature in an *ad hoc* manner. We propose an ontology that structures the most common inconsistencies in XAI. In particular, we propose to distinguish between two types of inconsistencies: those coming from the Machine Learning system, and those coming from the users' misinterpretations. We believe this ontology can help AI practitioners better understand the current limitations when generating ML explanations for a ML system and avoid explanation pitfalls.

Future works

The contributions of this thesis open several opportunities for further works. Beyond perspectives discussed in the chapters' conclusions, these include prospective works on the proposed experimental process for designing XAI principles and evaluating XUIs, as well as mitigating limitations for providing qualitative ML explanations. Four main research directions are identified and developed in the following sections. First, we discuss perspectives to extend the conducted work to other XAI approaches and applications. Then, we discuss new design enhancements opportunities for local explanations. Next, we discuss some fundamental challenges that we have identi-

fied in the explanation process and did not study. Finally, we identify perspectives opening the discussion for future Human-Computer Interaction contributions in AI research.

A transversal experimental process

A first perspective would be to develop the experimental procedure of this thesis for different XAI approaches, application contexts and kinds of users. This would allow us to have a holistic view of the intelligibility of various XAI methods.

First, one perspective is to investigate design enhancements for new types of explanations. In this thesis, we are interested in the intelligibility of local explanations in the form of counterfactual examples and feature importance for non-expert users. As discussed in Chapter 2, there is a great variety of explanations: according to hierarchical level (local or global), natures (e.g., feature importance, counterfactual examples, rules) and XAI approaches (e.g., LIME, SHAP, DiCE). We can study the extent to which this information can improve the objective understanding on the model's behavior. Also, studying several modalities of explanations would allow us to compare the intelligibility of these modalities for a specific group of users. Likewise, studying the same explanation modality for different groups would make it possible to compare them and better understand the disparities between these groups.

Another approach consists to apply the design enhancements we propose in a different application context. In this thesis, we have proposed several implementations of these principles in XUIs for two insurance scenarios. As reported in the user studies, insurance companies can be perceived negatively, and consequently this can have an impact on the obtained results. Implementing these principles in a different application setting with a well-established trust before interacting with the ML system (e.g., a medical tool for health) would allow to confirm the transversality of these design principles.

Deepening the study on local explanations

A second perspective is to continue investigating potential design enhancements for human-centered local explanations. In particular, we discuss in turn below two directions opportunities: personalizing the explanations and presenting them in conversational interfaces.

The proposed interfaces and conducted experiments show the interest of displaying personalized information to the users. It appears that part of this information can

be defined beforehand through common knowledge about users preferences. For instance, some variants of counterfactual examples, used in this thesis, suggest changes that are considered to be feasible, and others that are not. In a "human-in-the-loop" paradigm, it would be interesting to go deeper into personalization. By collecting users' needs in an automated way, the explanations can be tailored to each of them. For example, when the objective of an explanation is to optimize the prediction, a counterfactual example may propose a change that is usually feasible, but a user may not be able to do it because of his/her personal situation. Personalizing the explanations would thus allow to offer users with only the information they need. The question of how to collect these needs arises. We believe that investigating new interactive and dynamic design enhancements to collect such needs in XUIs would contribute to provide personalized and useful ML explanations.

Another interesting topic to study is the representation of explanations through the interface. In this thesis, we explore the usefulness of visual interfaces to present ML explanations but there are other interaction modalities (see Chapter 2), in particular the conversational mode. Indeed, users have progressively become familiar with conversational AIs (e.g., from voice assistants like Alexa to chatbots like ChatGPT). One perspective would be to enrich the representations explored in this thesis by providing explanations in the form of text, i.e., presented in conversational interfaces. The provided explanations may thus become a dialogue (i.e., questions from the users and corresponding answers from the machine) and take a whole new dimension: the interaction modality is natural for users; there is a narrative logic that allows temporizing the current information; and users may have the ability to express their needs with their own words. We believe that studying such a modality for the display of explanations would allow to better understand users' processes for analyzing and understanding ML explanations.

Investigating fundamental challenges in ML explanations

Other directions of research consist in extending the study of fundamental challenges in ML explanations, that have been identified but not addressed in this thesis. We focus on intelligibility and other criteria need to be considered for the quality of an explanation, such as trust and consistency.

The notion of trust represents a real challenge for explainability, as the two are related concepts: a "trustworthy" AI is often perceived as an "explainable" one. Trust is a factor that plays an important role in the adoption of any sort of systems. It is a complex topic that has given rise to many definitions in a wide variety of research

areas, as well as many works in XAI. In the context of an interaction between a user and a ML system, trust can be impacted by the application domain. In this thesis, we observed that many people perceive negatively insurance companies, which can have a direct impact on the understanding of an automated system provided by the latter. In such a context, one perspective is to study the construction and evolution of trust. We believe this would help AI practitioners to create more qualitative XAI approaches and to minimize the risks of over or under-trust (e.g., by comparing discrepancies between objective and subjective metrics). It would also help build XUIs that users can appropriately rely on to support their decision-making (e.g., in finance, users can rely on a trustworthy robot advisor for their assets management).

Besides, the ontology we propose identifies and categorizes common inconsistencies in ML explanations. It allows to know how to work on these issues. In particular for the quality of explanations, we should investigate first the inconsistencies that are specific to the ML system, and then the ones that are specific to the users. For example, it is likely to imagine that users may have not just one question but multiple ones regarding the ML system. Hence, building XUIs with multiple types of explanations may be useful to them. Yet, not all XAI approaches are compatible with each other. The same type of approach (e.g., feature importance) can generate contradictory information depending on whether the explanations are local or global. This is also true for the combination of several types of approaches (e.g., feature importance and conterfactual example). In such a context, we believe that mitigating these inconsistencies would be beneficial for the design of XUIs combining several XAI approaches. Another perspective oriented towards the users is to study their perceptions and understandings of potential discrepancies between diverse types of explanations, so as to better design such XUIs.

Human-Computer Interaction in AI

Beyond the scope of XUIs, we identify other relevant topics for Human-Computer Interaction contributions in AI.

When interacting with XUIs, users perceive the system as more "intelligent" than it is. This problem can be observed in the more general framework of AI. In particular with generative AIs (e.g., ChatGPT), the natural interaction modalities make such systems playful and invite the users to request for more complex tasks than what these AIs are able to do. There is thus a risk that users overtrust them due to their increased sophistication. These risks are even more fundamental than those mentioned in the ontology. A research perspective is to study the perception of these new

interaction modalities in order to see how to help the users having an appropriate understanding of the capabilities and limitations of these AIs.

Moreover, we believe another direction is to consider the collaboration between humans and AI. Generative AIs can blend into everyday tasks, making it possible to imagine that expert users can team up with these AIs to improve their performance on given tasks. For example, such AIs can be used in creative tasks (e.g., assisting game designers in the modeling of an environment). It requires that the expert users trust the AI agents for such tasks, as discussed in a work perspective previously developed. Hence, studying the trust building in such AIs and analyzing what types of tasks are easier to be delegated to an AI would be relevant to improve the adoption and relationship between the end-user and an AI agent.

A

Proposed XUI for local feature importance: evaluation materials

In this appendix, we present in details the evaluation material used for the evaluation of the proposed XUI described in Chapter 3: the usage scenario presented to participants during the study, the experimental setup at the INSEAD Behavioural Lab, the detailed questionnaire for measuring objective understanding, and the pilot study.

A.1 | Usage scenario for participants

Participants are presented with the same basis scenario as follow:

Marianne (47 year old), lives with her daughter, Lucille (21 year old) in Bourg la Reine, near Paris (France). Lucille will start her internship in Marne-la-Vallée (France) next month and will have to take her mother's car to go to work. Marianne would like to change her insurance policy to provide better protection against accidents and damages because her daughter is a young driver.

The information that Marianne has filled in to take out a new car insurance policy:

- *Her personal information (gender, age, year of license, place of residence)*
- *Her daughter Lucille's personal information (title, age, year of license, place of residence)*
- *Her vehicle's license plate (information retrieved from the INSEE database: Make and model of the vehicle, engine power, maximum speed, type of power supply, cubic capacity, age, start and end dates of the model...)*
- *Information about her insurance history (her Bonus/Malus, her contract)*

- *Her insurance coverage preferences (intermediate +), payment frequency (annual) and insurance options (second driver option for Lucille)*

After filling in all this information, Marianne discovers the price that this insurer is offering online (based on all the information she has provided) and the explanations that will allow her to understand this price. Does this explanation allow her to make an informed decision? Your role will be to look at the explanations given for this price and give your opinion to Marianne.

A.2 | Experimental setup in lab

Figure A.1: Testing room at INSEAD Behavioural Lab composed of 12 isolated desks

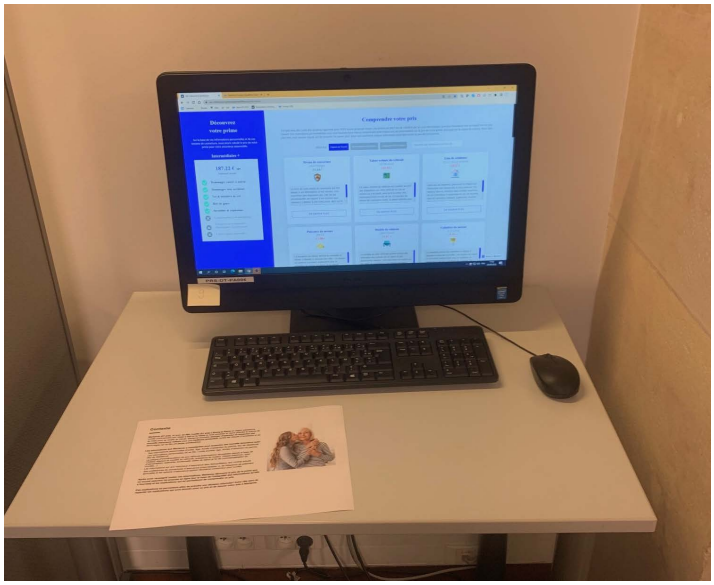


Figure A.2: On each isolated desktop, participant has a printed scenario, and a computer with the assigned version of the interface

A.3 | Objective understanding questionnaire

This section lists the questions asked to the user-lab experiment participants to evaluate the proposed interface, translated from the original language. In all cases, except for the open-response question, the participant must choose between three answers:

- I agree with Marianne
- I disagree with Marianne
- I don't know

A.3.1 | Explanations' scope

- Marianne thinks that the fact that both drivers are women has an impact on the proposed premium price. Do you agree with her?
- Marianne believes that the relationship between the two drivers has an impact on the proposed premium price. Do you agree with her?
- Marianne thinks that her age has an impact on the proposed premium price. Do you agree with her?
- Marianne thinks that the maximum speed of her vehicle has an impact on the proposed premium price. Do you agree with her?
- Marianne thinks that the level of coverage she has chosen has an impact on the proposed premium price. Do you agree with her?
- Marianne believes that her choice not to take the pay-per-mile option has an impact on the proposed premium price. Do you agree with her?
- Marianne believes that her vehicle information is used to calculate the proposed premium price. Do you agree with her?
- Marianne believes that only the 1st driver's information is used to calculate the proposed premium price. Do you agree with her?

A.3.2 | Explanations' effect

- Marianne believes that her bonus/malus percentage has less of an impact on the quoted premium price than her total historical insurance period. Do you agree with her?

- Marianne thinks that the make of her vehicle has less of an impact on the price of her premium than her model. Do you agree with her?
- Marianne thinks that Lucille's age has less of an impact on the proposed premium price than her age. Do you agree with her?
- Marianne believes that the use she and her daughter have of their vehicle does not affect the proposed premium price. Do you agree with her?
- Marianne thinks that the level of coverage she has chosen increases the price of the premium quoted. Do you agree with her?
- Marianne believes that the appraised value of her vehicle increases the proposed premium price. Do you agree with her?
- Marianne believes that her age decreases the proposed premium price. Do you agree with her?

A.3.3 | Explanations' locality

- Marianne thinks that the impact of the information she provided on the proposed premium price would be the same for sure for all other people who have similar information to her and Lucille. Do you agree with her?
- Marianne thinks that the proposed premium price would be the same for sure for all other people who have similar information to her and Lucille. Do you agree with her?
- Marianne thinks that if she and her daughter had another vehicle, the proposed premium price would remain the same for them. Do you agree with her?
- Marianne thinks that if she had a higher bonus/malus percentage, the proposed premium price would probably be different. Do you agree with her?
- Marianne thinks that if the value of her vehicle was estimated to be lower, the impact on the proposed premium price would surely be different.

A.4 | Pilot study

This section presents the first experimental evaluation in an online settings to evaluate the evaluation material. In this pilot study, we evaluate the intelligibility of a first

version of contextualization principles we propose and their implementation into an XUI for the same insurance scenario, as presented in Chapter 3. We describe in turn the questionnaires, the study design and the results for this pilot study. We use the results of this pilot study to correct and finalize the experimental procedure we develop.

A.4.1 | Pilot questionnaires

We use a simplified version of the evaluation material described in Section 3.6.1. We measure both the objective understanding and satisfaction.

Objective understanding We propose four types of questions to check the user objective understanding. The details of the questionnaire are provided in A.

- (i) **Feature Importance Questions** measure the extent to which the user understands the relative influence of the attributes on the prediction, e.g., *"Does feature X impact more the prediction than feature Y ?"*.
- (ii) **ML Information Questions** measure the user's effective understanding of what the ML system is and how it works, e.g., *"Are the explanations provided based on the average prediction?"*.
- (iii) **Local Explanation Questions** measure the user's understanding of the difference between the influence of his/her attributes and global explanations, e.g., *"Will the prediction remain for sure the same even if feature X is different?"*.
- (iv) **Interpretation Questions** measure the extent to which the user processes the explanations provided to understand the price, e.g., *"Does this information/event influence the prediction?"*.

We design two quiz questions for each of the four types. For statement questions, three answer options are provided: "true", "false" and "I don't know"; for one-choice questions, lists of possible answers are offered as well as an "I don't know" option. We measure the answer correctness and time to answer each question. The 8 questions are as follows:

Question 1: The model of your vehicle influences more your price than the number of children you have at charge.

True

False

I don't know
Question 2: What is the influence of the gearbox of your car on your price?
It increases my price
It doesn't change my price
It decreases my price
I don't know
Question 3: Even if you were older, you would get the same price for sure.
True
False
I don't know
Question 4: If you were living in another city, you would probably get a different price.
True
False
I don't know
Question 5: Which one of your information doesn't influence your price?
My age
My vehicle's power supply
My job occupation
My residence area
I don't know
Question 6: Again, which one of your information doesn't influence your price?
The model of my vehicle
The number of children at my charge
My gender
My job occupation
I don't know
Question 7: Your price is calculated based on an average price of 15.5€
True
False
I don't know
Question 8: Your information increases your price by 1.15€.
True
False
I don't know

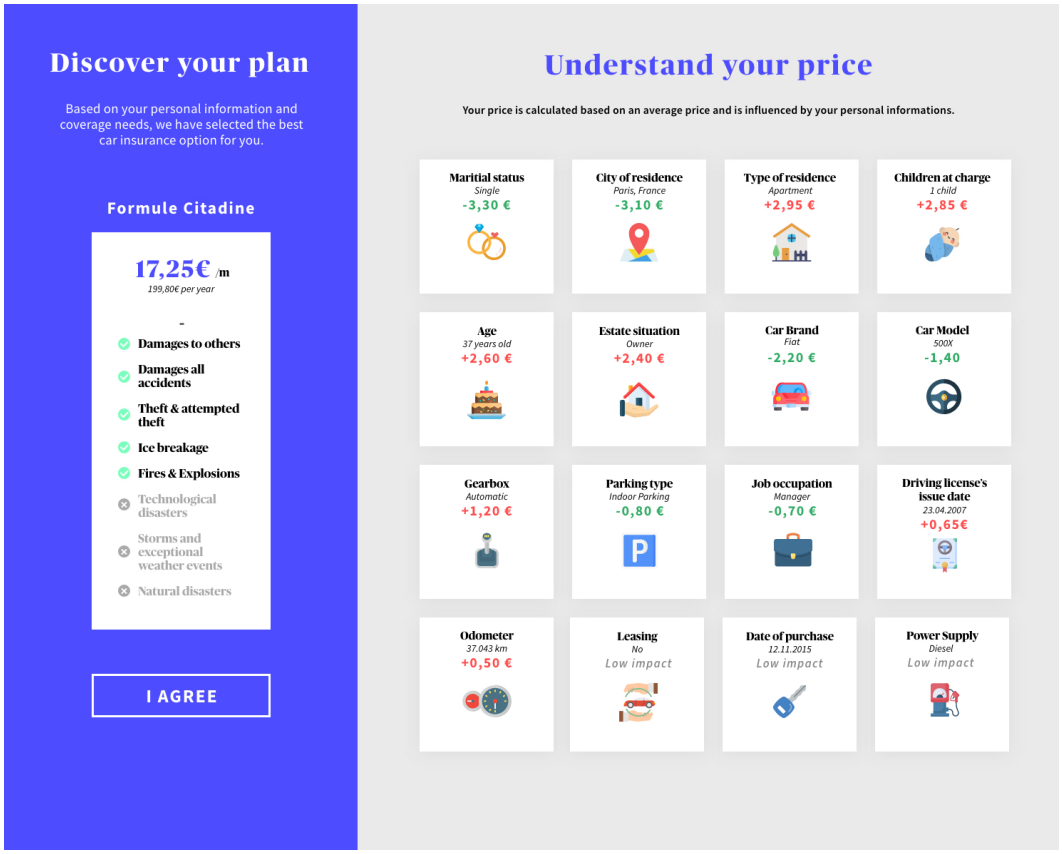


Figure A.3: Interface A without contextualization principles

Satisfaction We use two questions from the Explanation Satisfaction Scale Hoffman et al. (2018), to assess the perceived understanding and usefulness of explanations. The first question asks the participant his/her perceived understanding of the explanations, the second one his/her perceived usefulness of the explanations.

Additional questions In addition, the 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". We also ask for basic demographic information such as age and education level. Finally, participants can share their insights and comments on the study in an open response question.

A.4.1.1 | Experimental Design

A/B Testing We conduct an A/B testing to compare the results obtained for the baseline interface (A) with the enhanced one (B). Interface A, displayed in Figure A.3,

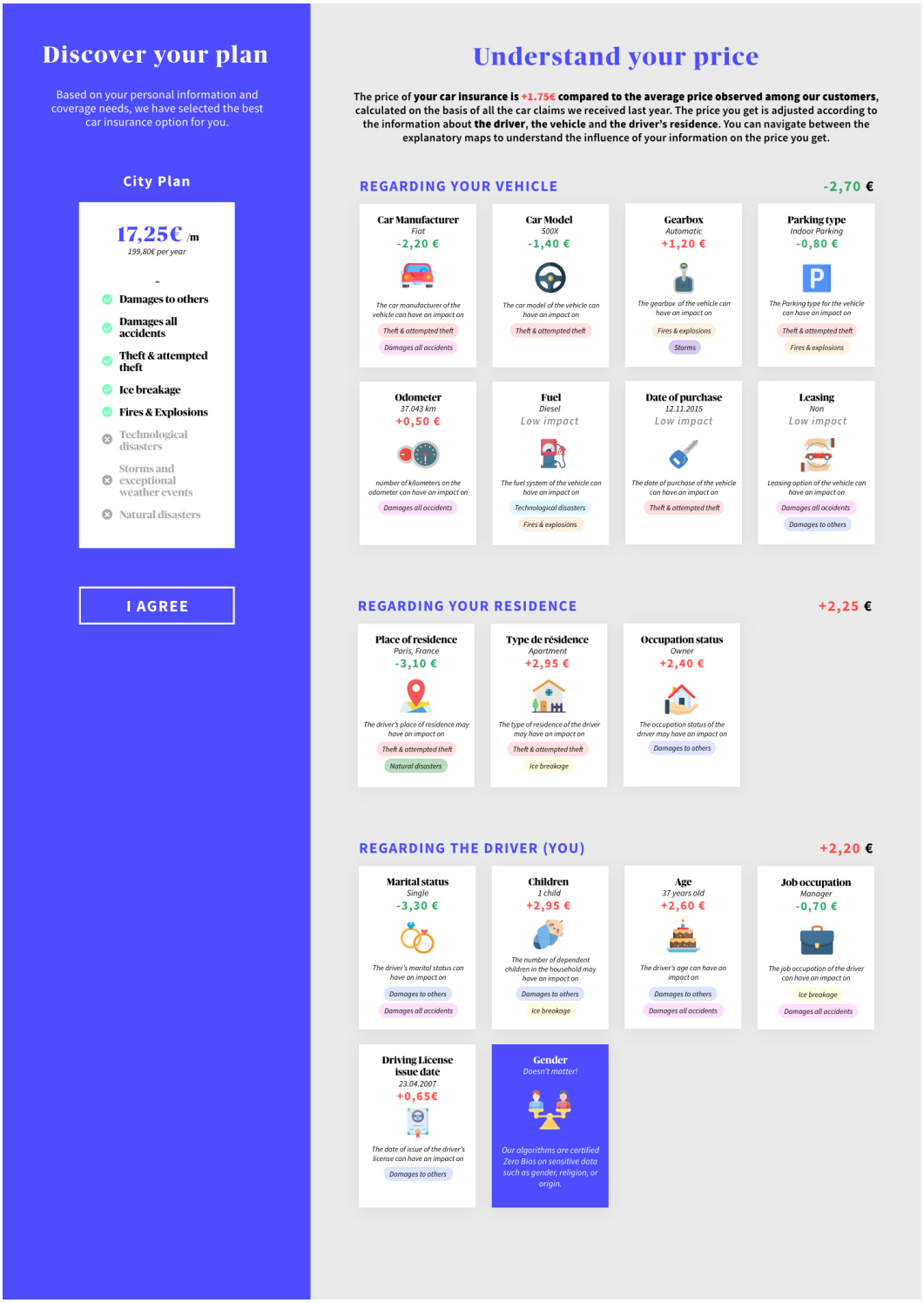


Figure A.4: Interface B with contextualization principles

	Objective understanding	Feature Importance Questions	Interpretation Questions	Local Explanation Questions	ML Information Questions
Interface A	**0.73 (±0.20)**	0.71 (±0.24)	0.71 (±0.24)	0.71 (±0.39)	0.14 (±0.35)
Interface B	**0.88 (±0.16)**	0.63 (±0.26)	0.63 (±0.26)	1.00 (±0.00)	0.83 (±0.41)
	Self-reported understanding	**Self-reported usefulness**			
Interface A	**0.71 (±0.28)**	**0.63 (±0.37)**			
Interface B	**0.87 (±0.16)**	**0.91 (±0.10)**			

Table A.1: Obtained results for the two interfaces, interface A without contextualisation and interface B with contextualisation: for objective questions, average and standard deviation of the percentage of correct answers (overall and for each question type), for self-reported questions, average and standard deviation of the scores on the Likert scale

presents local feature importance explanations as extracted from SHAP, as described in Section 3.6.2.1. Interface B includes a first experimental implementation of our propositions, and it is displayed in Figure A.4.

Pilot procedure For the pilot experiment, participants are randomly assigned to one version of the interface. Each participant acts as a female persona with a given set of 16 feature values related to the driver, the vehicle and the residence of the driver. Prior to the evaluation, participants are introduced to the persona and her need to understand the price she gets. We also explain the platform uses an algorithm to determine a personalized price based on her personal information. The evaluation starts with the objective understanding question quiz, which is displayed next to the interface to allow participants to look for the answers. Then, the subjective understanding questions and demographics information questions are asked.

Because of the COVID-19 situation, we were unable to conduct the pilot study in lab. Thus, we conducted the pilot on Useberry[1]. 20 participants were recruited from university and professional social network.

A.4.2 | Results

The obtained results are displayed in Table A.1. For objective questions, the results are defined as the percentage of correct answers; for the subjective questions, the results are the average scores on the Likert scale, normalized to the [0,1] interval. We did not collect enough data to perform further analytics.

The data of 6 participants were not exploitable as they dropped off from the survey at the start. We also excluded the data of one more participant, who completed

[1]https://www.useberry.com/

the test in an abnormally short time and who appeared not to scroll through the explanations to look for the answers. Out of the 13 remaining participants, 7 were assigned to interface A and 6 to interface B. Participants assigned to interface A (resp. interface B) are 29.6 years old on average (resp. 29.8) and reported an average artificial intelligence literacy score of 0.71 (resp. 0.37) and an average insurance literacy score of 0.47 (resp. 0.60).

Participants assigned to interface B obtain overall higher scores for the objective understanding questions (0.88) as compared to the ones using interface A (0.73). When considering the different types of questions, it appears that interfaces A and B lead to comparable results for the feature importance and local explanation questions. This could mean that providing local feature importance is enough for a non-expert user to correctly answer these questions, even without any contextualization. The results hint that there is an improvement for participants assigned to interface B for the interpretation and ML information questions, which hints that contextual information on the ML system may help non-expert users to interpret the explanations.

In this pilot study, participants who used interface B report higher satisfaction (0.87) compared to version A (0.71) and also rate higher the usefulness of the explanations (0.91 for interface B and 0.63 for interface A).

Finally for the feedback collected through the open-question, it appears that participants using interface A, without contextual information, report more uncertainty regarding their answers and their understanding, as two of them explicitly state. On the other hand, 1 participant using interface B reports that the explanations are "pleasantly surprising and help choosing among different insurance plans", while another participant states that the explanations are clear.

A.4.3 | Discussion

Overall, we observe that the contextualization elements of interface B provide an improvement for all considered evaluation metrics: +0.14 for objective understanding, +0.15 for self-reported understanding and +0.29 for self-reported usefulness. These preliminary results indicate that interface B seems to improve the explanation understanding thanks to the three levels of added contextual information. Yet, the sample size is too small to provide strong and reliable insights backed up with statistical tests. However, the results confirm that contextualization principles represent a promising approach to improving the intelligibility of local explanations. Moreover, this pilot study allows us to adjust the evaluation method for a larger scale study. In particular, it allows to adapt the objective understanding questionnaire, to adapt eight dimen-

sions of the Explanation Satisfaction Scale proposed by Hoffman et al. (2018), and to favor in lab settings for the user study to be able to monitor the experiment and avoid participants' drop-off.

B

Proposed XUI for counterfactual examples: evaluation materials

In this appendix, we present the details materials for the evaluation of the proposed XUI describe in Chapter 4 and adapted from the materials used in previous work presented in Chapter 3 and in Appendix A.

B.1 | Usage scenario for participants

Participants are presented with the same basis scenario as follow:

"Swann is a 26 year old graphic designer who lives in Bordeaux and has recently started working for him/herself. Swann dreams of having a comfortable and spacious home because his/her current rental is not adapted to work from home. Swann has just found the perfect place: a small apartment with a balcony in the center of town. But he/she still needed money to move in, furnish and decorate it to his/her liking. Swann contacted his/her bank to obtain a consumer loan. After talking to his/her bank advisor, Swann knew that it was possible to borrow up to 15,000€ at a rate of less than 20%, with repayments spread out over 36 months. To apply for a loan, Swann goes to his/her bank's website, applies for a loan and enters his/her information. She/He would need to borrow 10,974€ to finalize his project. Having already taken out a few loans in the past (notably to move to Paris for his/her studies, and then to travel after graduation), Swann thinks his/her application will be accepted. Unfortunately, the site informed him/her that his/her creditworthiness was insufficient, and his/her application was denied. To understand why the service considered him/her creditworthiness to be insufficient, explanations show him/her changes that could have been made to his/her data to improve the credit rating."

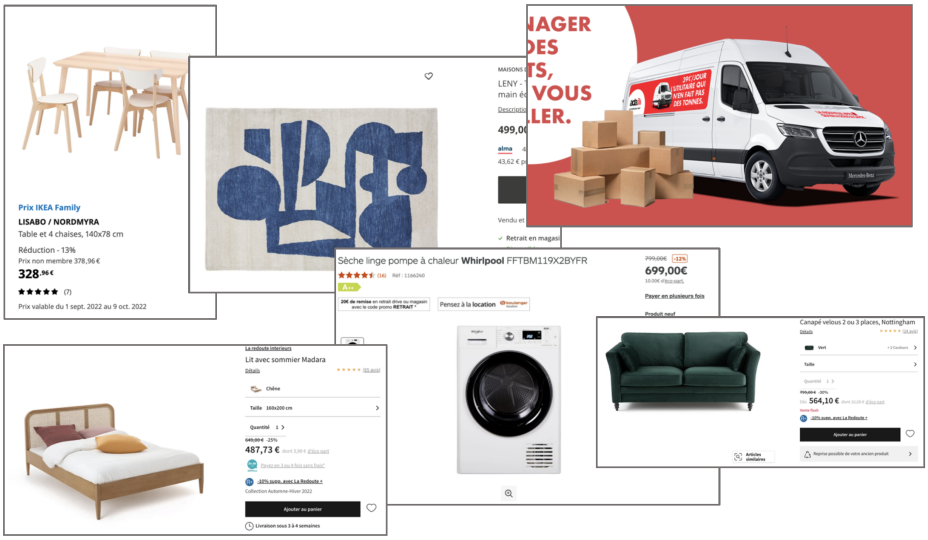


Figure B.1: Onboarding illustration on the scenario for participants, in relation with Swann's project for moving into a new apartment

We also provide images to accompany the scenario, as illustrated in Figure B.1.

B.2 | Experimental setup in lab

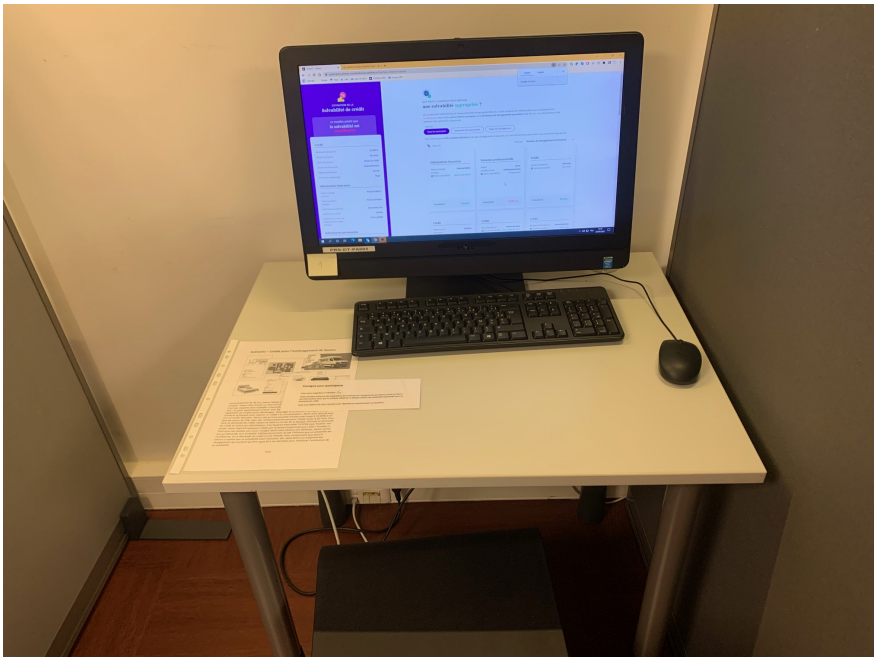


Figure B.2: On each isolated desktop, the participant has a printed scenario, a printed instruction notice, a computer with the assigned version of the interface

B.3 | Objective understanding questionnaire

This section lists the questions asked to the user-lab experiment participants to evaluate the proposed interface, translated from the original language. In all cases, except for the open-response question, the participant must choose between three answers:

- I agree with Swann
- I disagree with Swann
- I don't know

B.3.1 | Explanations' nature questions: counterfactual examples

- Swann thinks that the information displayed indicates what variations can be made on his/her information, to be predicted as having an adequate solvency.
- Swann thinks that the proposed changes are always on the parameters of his/her credit application (amount, duration, loan rate, and so on).
- Swann thinks that all the provided information needs to be changed, in addition to the proposed changes, to be predicted as having an adequate solvency.
- Swann thinks this system proposes changes to be predicted as having an adequate solvency.

B.3.2 | Explanations' effects questions: if...then...

- Swann thinks that if the loan term was 20 months instead of 36 months, his/her solvency would have been predicted as adequate.
- Swann thinks that for the solvency to be predicted as adequate, the loan term could be reduced by 10 months.
- Swann thinks that with a loan rate of 22%, the solvency would be predicted as adequate.
- Swann believes that his/her solvency would have been predicted as adequate if there had been a co-borrower.
- Swann thinks that the solvency would have been predicted as adequate if he/she was not a foreigner.

B.3.3 | Explanations' specificity question: plurality

- Swann believes that the only way to be predicted as having an adequate solvency would be to reduce the loan duration to 26 months.
- Swann thinks that the least common changes suggested concern the employment status.
- Swann thinks that among all the proposed changes, some are more feasible than others.
- Swann thinks that he/she would have to be at least 54 year old in order to be predicted as having an adequate solvency.
- Swann believes that for all the examples provided, only one or two pieces of information would need to be changed as indicated in order to be predicted as having an adequate solvency

B.3.4 | Open-response question

What strategy/changes would you recommend to Swann to make his/her solvency to be predicted as adequate?

B.4 | Satisfaction questionnaire

This section presents the self-reporting questionnaire we propose, adapted from the Explanation Satisfaction Scale Hoffman et al. (2018) in order to assess users' satisfaction (translated from the original language). Participants are required to answer on a 6-point Likert scale, from "Strongly disagree" (1) to "Strongly agree" (6).

B.4.1 | Explanation Satisfaction Scale adapted

- In your opinion, the explanations for obtaining appropriate creditworthiness are understandable
- In your opinion, the explanations for obtaining appropriate credit are satisfying
- In your opinion, the explanations for obtaining appropriate credit are sufficiently detailed
- In your opinion, the explanations for obtaining appropriate credit are complete
- In your opinion, the proposed explanations indicate how they should be interpreted to fully understand how to obtain appropriate credit
- In your opinion, the explanations for obtaining appropriate credit are useful in helping you make an informed decision
- In your opinion, the explanations for obtaining appropriate credit are accurate
- n your opinion, the explanations for obtaining appropriate credit are trustworthy

B.4.2 | Open-response question

How satisfied are you with the explanations the interface provides to achieve appropriate creditworthiness?

References

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

Amir Ahmad and Shehroz S. Khan. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7:31883–31902, 2019.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *Int. Conf. on Machine Learning*, pages 314–323. PMLR, 2020.

Charles Antaki and Ivan Leudar. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2):181–194, 1992.

Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proc. of the AAAI Conf. on Artificial Intelligence*, pages 5277–5285, 2022.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv preprint:1909.03012*, 2019.

Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

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

Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, and Cecilia Zanni-Merk. Une terminologie pour une IA explicable contextualisée. In *EGC 2022 Workshop EXPLAIN'AI*, 2022.

Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proc. of the 2022 AAAI/ACM Conf. on AI, ethics, and society, AEIS 2022*, pages 78–91, 2022.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proc. of the 2020 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2020*, pages 648–657, 2020.

Anol Bhattacherjee. Understanding information systems continuance: An expectation-confirmation model. *MIS quarterly*, pages 351–370, 2001.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI Workshop on eXplainable AI (XAI)*, pages 8–13, 2017.

Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proc. of the 27th Int. Conf. on Intelligent User Interfaces*, IUI 2022, 2022.

Garvin Brod, Markus Werkle-Bergner, and Yee Lee Shing. The influence of prior knowledge on memory: a developmental cognitive neuroscience perspective. *Frontiers in behavioral neuroscience*, 7:139, 2013.

Sylvain Bromberger. Why-questions, 1966.

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

Ruth MJ Byrne and Alessandra Tasso. Counterfactual reasoning: Inferences from hypothetical conditionals. In *Proc. of the 16th Annual Conf. of the Cognitive Science Society*, pages 124–130. Routledge, 2019.

Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proc. of the 24th Int. Conf. on Intelligent User Interfaces, IUI 2019*, pages 258–262, 2019.

John M Carroll and Judith Reitman Olson. Mental models in human-computer interaction. *Handbook of human-computer interaction*, pages 45–65, 1988.

Leiyu Chen, Shaobo Li, Qiang Bai, Jing Yang, Sanlong Jiang, and Yanming Miao. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22), 2021a.

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021b.

Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 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 2019*, page 1–12. ACM, 2019.

Michael Chromik and Andreas Butz. Human-XAI interaction: a review and design principles for explanation user interfaces. In *Proc. of the 18th TC13 Int. Conf. on Human-Computer Interaction, INTERACT 2021*, pages 619–640. Springer, 2021.

Michael Chromik and Martin Schuessler. A taxonomy for human subject evaluation of black-box explanations in XAI. In *IUI Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies (ExSS-ATEC)*. CEUR, 2020.

Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 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 2021*, pages 307–317. ACM, 2021.

Victoria Clarke, Virginia Braun, and Nikki Hayfield. Thematic analysis. *Qualitative psychology: A practical guide to research methods*, 222:248, 2015.

Dennis Collaris, Leo M Vink, and Jarke J van Wijk. Instance-level explanations for fraud detection: A case study. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *Int. Conf. on Parallel Problem Solving from Nature*, pages 448–469. Springer-Verlag, 2020.

Maartje De Graaf and Bertram Malle. How people explain action (and autonomous intelligent systems should too). In *Proc. of the 2017 AAAI Fall Symposium Series*, 2017.

Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

Daniel C Dennett. *The Intentional Stance*. MIT press, 1989.

Daniel C Dennett. *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company, 2017.

Michael DeVito, Jeffrey Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 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 2018*, pages 1–6. ACM, 2018.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *Proc. of the 24th European Conference on Artificial Intelligence, ECAI 2020*, volume 325 of *Frontiers in Artificial Intelligence and Applications (FAIA)*, pages 2473–2480. IOS Press, 2020.

Filip JRC Dochy and Patricia A Alexander. Mapping prior knowledge: A framework for discussion among researchers. *European Journal of Psychology of Education*, pages 225–242, 1995.

Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proc. of the 24th Int. Conf. on Intelligent User Interfaces*, IUI 2019, page 275–285. ACM, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint:1702.08608*, 2017.

Christophe Dutang and Arthur Charpentier. Package 'CASdatasets', 2020.

Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conf. on Human Factors in Computing Systems*, CHI 2021. ACM, 2021.

Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. Human-centered explainable AI (HCXAI): beyond opening the black-box of AI. In *Extended Abstracts of the 2022 CHI Conf. on Human Factors in Computing Systems*, CHI 2022, pages 1–7, 2022.

Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proc. of the 8th ACM Conf. on Recommender Systems*, RecSys '14, page 161–168. ACM, 2014.

Frank Emmert-Streib and Matthias Dehmer. Evaluation of regression models: Model assessment, model selection and generalization error. *Machine Learning and Knowledge Extraction*, 1(1):521–551, 2019.

Shi Feng and Jordan Boyd-Graber. What can AI do for me? Evaluating machine learning interpretations in cooperative play. In *Proc. of the 24th Int. Conf. on Intelligent User Interfaces*, IUI'19, pages 229–239, 2019.

M. Fernández-Delgado, M.S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande. An extensive experimental survey of regression methods. *Neural Networks*, 111:11–34, 2019.

Krzysztof Z Gajos and Lena Mamykina. Do people engage cognitively with AI? impact of AI assistance on incidental learning. In *27th Int. Conf. on Intelligent User Interfaces, IUI 2022*, pages 794–806, 2022.

Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Vice: visual counterfactual explanations for machine learning models. In *Proc. of the 25th Int. Conf. on Intelligent User Interfaces, IUI'20*, pages 531–535. ACM, 2020.

Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *IEEE Visualization Conf., VIS 2022*, pages 31–35. IEEE, 2021.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Nastacia L Goodwin, Simon RO Nilsson, Jia Jie Choong, and Sam A Golden. Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. *Current opinion in neurobiology*, 73:102544, 2022.

Brandon M. Greenwell, Bradley C. Boehmke, and Andrew J. McCarthy. A simple and effective model-based variable importance measure. *arXiv eprint:1805.04755*, 2018.

Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

David Gunning. Explainable artificial intelligence (XAI). *Defense advanced research projects agency (DARPA)*, 2(2):1, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proc. of the 34th Int. Conf. on Machine Learning*, volume 70 of *Proc. of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part I: Causes. *The British journal for the philosophy of science*, 2005.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems*, CHI 2018, page 1–14. ACM, 2018.

Nimrod Harel, Ran Gilad-Bachrach, and Uri Obolski. Inherent inconsistencies of feature importance. *arXiv preprint:2206.08204*, 2022.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *Association for Computational Linguistics (ACL)*, 2020.

Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery*, 28:1503–1529, 2014.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems, NeurIPS*, 32, 2019.

Germund Hesslow. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 11–32, 1988.

Denis J Hilton. Logic and causal attribution. In *Proc. of the Annual Conf. of the British Psychological Society*. New York University Press, 1988.

Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107 (1):65, 1990.

Denis J Hilton. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4):273–308, 1996.

Denis J Hilton and McClure L John. The course of events: counterfactuals, causal sequences, and explanation. In *The psychology of counterfactual thinking*, pages 56–72. Routledge, 2007.

Denis J Hilton and Ben R Slugoski. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1):75, 1986.

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

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 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 2019*, pages 1–13. ACM, 2019.

Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. 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*, pages 1–8, Cham, 2018. Springer International Publishing.

Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. *KI-Künstliche Intelligenz*, 34(2):193–198, 2020.

David Hume. *An enquiry concerning human understanding: A critical edition*, volume 3. Oxford University Press on Demand, 2000.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. How can I choose an explainer? an application-grounded evaluation of post-hoc explanations. In *Proc. of the 2021 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2021*, pages 805–815, 2021.

Adulam Jeyasothy, Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Integrating prior knowledge in post-hoc explanations. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'2022)*, volume 1602, pages 707–719. Springer, 2022.

John R Josephson and Susan G Josephson. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, 1996.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Towards causal algorithmic recourse. In *ICML Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 139–166, Cham, 2022. Springer.

Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In *Proc.of the Int. Joint Conf. on Artificial Intelligence, IJCAI-21*, pages 4466–4474. International Joint Conferences on Artificial Intelligence, 2021.

Frank C Keil. Explanation and understanding. *Annual Review of Psychology*, 57:227–254, 2006.

Harold H Kelley. Causal schemata and the attribution process. *American Psychologist*, 28:107, 1987.

Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294:103459, 2021.

René F. Kizilcec. How much information? Effects of transparency on trust in an algorithmic interface. In *Proc. of the 2016 CHI Conf. on Human Factors in Computing Systems*, CHI 2016, page 2390–2395. ACM, 2016.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4), 2019.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI 2016*, pages 5686–5697. ACM, 2016.

Matev Kunaver and Toma Porl. Diversity in recommender systems a survey. *Knowledge-Based Systems*, 123(C):154–162, 2017.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

Michael T. Lash, Qihang Lin, Nick Street, Jennifer G. Robinson, and Jeffrey Ohlmann. Generalized inverse classification. In *Proc.of the 2017 SIAM Int. Conf. on Data Mining, SDM2017*, pages 162–170. SIAM, 2017.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *Proc. of the 17th Int. Conf. of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2018*, pages 100–111. Springer, 2018a.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. In *ICML Workshop on Human Interpretability for Machine Learning (WHI)*, 2018b.

Thibault Laugel, Marie Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proc. of the Int. Joint Conf. on Artificial Intelligence, IJCAI'19*, pages 2801–2807, 2019.

Thibault Laugel, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Achieving diversity in counterfactual explanations: a review and discussion. In *Proc.of the 2023 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2023*. ACM, 2023.

Thai Le, Suhang Wang, and Dongwon Lee. Grace: Generating concise and informative contrastive sample to explain neural network model's prediction. In *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, pages 238–248. ACM, 2020.

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

D Lewis. Causal explanation. *Philosophical Papers*, pages 214–240, 1986.

David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

Q Vera Liao and Kush R Varshney. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint:2110.10790*, 2021.

Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI 2020*, pages 1–15. ACM, 2020.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

Zachary C. Lipton. The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, page 36–43. ACM, 2016.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proc. of the 19th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 623–631, 2013.

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

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *NeurIPS Workshop on CausalML*, 2019.

Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press, 2006.

Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655, 2021.

Kyle Martin, Anne Liret, Nirmalie Wiratunga, Gilbert Owusu, and Mathias Kern. 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*, pages 309–324. Springer-Verlag, 2019.

Ann L McGill and Jill G Klein. Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, 64(6):897, 1993.

Christian Meske and Enrico Bunde. Design principles for user interfaces in AI-based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers*, pages 1–31, 2022.

John Stuart Mill and John M Robson. A system of logic: The collected works of john stuart mill. In *University of Toronto Press Routledge*, volume 7, page 353. Kegan Paul, 1973.

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

Yao Ming, Huamin Qu, and Enrico Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on visualization and computer graphics*, 25(1):342–352, 2018.

Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint:2111.00358*, 2021.

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

Christoph Molnar. Interpretable machine learning - a guide for making black box models explainable, 2020.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proc.of the 2020 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2020*, pages 607–617. ACM, 2020.

Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint:1902.01876*, 2019.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *arXiv preprint:2201.08164*, 2022.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. ACL, 2018.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Donald Norman, Dedra Gentner, and AL Stevens. Mental models. *Human-computer Interaction*, pages 7–14, 1983.

Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proc. of the 26th Int. Conf. on Intelligent User Interfaces, IUI'21*, pages 340–350, 2021.

Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric D Ragan, and Vibhav Gogate. On the importance of user backgrounds and impressions: Lessons learned from interactive AI applications. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–29, 2022.

Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27:393–444, 2017.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Jeroen Ooge, Leen Dereu, and Katrien Verbert. Steering recommendations and visualising its impact: Effects on adolescents' trust in e-learning platforms. In *Proc. of the 28th Int. Conf. on Intelligent User Interfaces, IUI'23*, pages 156–170, 2023.

James A Overton. Scientific explanation and computation. *Explanation-aware Computing ExaCt 2011*, page 41, 2011.

Reshika Palaniyappan Velumani, Meng Xia, Jun Han, Chaoli Wang, ALEXIS K LAU, and Huamin Qu. AQX: Explaining air quality forecast for verifying domain knowledge using feature importance visualization. In *Proc. of the 27th Int. Conf. on Intelligent User Interfaces, IUI'22*, pages 720–733, 2022.

Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust. *IJCAI Workshop on Explainable Artificial Intelligence*, 2019.

R. S. M. Lakshmi Patibandla and N. Veeranjaneyulu. Survey on clustering algorithms for unstructured data. In Vikrant Bhateja, Carlos A. Coello Coello, Suresh Chandra Satapathy, and Prasant Kumar Pattnaik, editors, *Intelligent Engineering Informatics*, pages 421–429. Springer Singapore, 2018.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society*, pages 344–350. ACM, 2020.

Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable AI. *arXiv preprint:1810.00184*, 2018.

Yanou Ramon, Tom Vermeire, Olivier Toubia, David Martens, and Theodoros Evgeniou. Understanding consumer preferences for explanations generated by XAI algorithms. *arXiv preprint:2107.02624*, 2021.

Stephen J Read and Amy Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3): 429, 1993.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " Why should I trust you?" explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining*, pages 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proc. of the AAAI Conf. on Artificial Intelligence*, Palo Alto, CA, United States, 2018. AAAI Press.

Maria Riveiro and Serge Thill. "That's (not) the output I expected!" on the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence*, 298:103507, 2021.

Pau Rodriguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pages 1056–1065, 2021.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proc. of the 39th Int. Conf. on Machine Learning*, volume 162, pages 18770–18795. PMLR, 2022a.

Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable AI: User studies for model explanations. *arXiv preprint:2210.11584*, 2022b.

Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562, 2002.

Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *stat*, 1050: 26, 2018.

Téo Sanchez, Baptiste Caramiaux, Pierre Thiel, and Wendy E Mackay. Deep learning uncertainty in machine teaching. In *Proc. of the 27th Int. Conf. on Intelligent User Interfaces, IUI 2022*, pages 173–190, 2022.

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

James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your AI. In *Proc. of the 24th Int. Conf. on Intelligent User Interfaces, IUI'19*, pages 240–251, 2019.

Daniel Schwarcz. Transparently opaque: Understanding the lack of transparency in insurance consumer protection. *UCLA Law Review*, 61:394, 2014.

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

Alana Semuels et al. The internet is enabling a new kind of poorly paid hell. *The Atlantic*, 23, 2018.

Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. Why am I not seeing it? understanding users' needs for counterfactual explanations in everyday recommendations. In *Proc.of the 2022 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2022*, page 1330–1340. ACM, 2022.

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

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society*, pages 180–186, 2020.

Hansem Sohn, Devika Narain, Nicolas Meirhaeghe, and Mehrdad Jazayeri. Bayesian computation through cortical latent dynamics. *Neuron*, 103(5):934–947, 2019.

Kacper Sokol and Peter Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34(2):235–250, 2020.

Kacper Sokol and Peter A Flach. Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *Proc.of THE Int.*

Joint Conf. on Artificial Intelligence, IJCAI-18, pages 5868–5870. Int. Joint Conf on Artificial Intelligence, 2018.

Sumit Srivastava, Mariët Theune, and Alejandro Catala. The role of lexical alignment in human understanding of explanations by conversational agents. In *Proc. of the 28th Int. Conf. on Intelligent User Interfaces*, IUI 2023, page 423–435. ACM, 2023.

Muhammad Suffian, Pierluigi Graziani, Jose M. Alonso, and Alessandro Bogliolo. FCE: Feedback based counterfactual explanations for explainable AI. *IEEE Access*, 10:72363–72372, 2022. doi: 10.1109/ACCESS.2022.3189432.

Maxwell Szymanski, Vero Vanden Abeele, and Katrien Verbert. Explaining health recommendations to lay users: The dos and dont's. In *IUI Workshop on Adaptive and Personalized Explainable User Interfaces (APEx-UI)*, IUI 2022. CEUR, 2022a.

Maxwell Szymanski, Katrien Verbert, and Vero Vanden Abeele. Designing and evaluating explainable AI for non-AI experts: challenges and opportunities. In *Proc. of the 16th ACM Conf. on Recommender Systems*, pages 735–736, 2022b.

Paul Thagard. Extending explanatory coherence. *Behavioral and brain sciences*, 12(3):490–502, 1989.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proc. of the AAAI Conf. on Artificial Intelligence*, pages 6021–6029, 2020.

Tom Trabasso and Jake Bartolone. Story understanding and counterfactual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5):904, 2003.

S. Umadevi and K. S. Jeen Marseline. A survey on data mining classification algorithms. In *Proc. of the 2017 Inter. Conf. on Signal Processing and Communication, ICSPC'17*, pages 264–268, 2017.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proc. of the 2019 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2019*, pages 10–19, 2019.

Jeroen Van Bouwel and Erik Weber. Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*, 32(4):437–449, 2002.

Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint:2010.10596*, 2022.

Giulia Vilone and Luca Longo. A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. *Frontiers in Artificial Intelligence*, 4, 2021.

Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for lime: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022.

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

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable AI. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI 2019*, page 1–15. ACM, 2019.

Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 8981–8990, 2020.

Xinru Wang and Ming Yin. 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 2021, page 318–328. ACM, 2021.

Greta Warren, Mark T Keane, and Ruth MJ Byrne. Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. *IJCAI-ECAI Workshop on Cognitive Aspects of Knowledge Representation*, 2022.

Daniel Karl I. Weidele, Justin D. Weisz, Erick Oduor, Michael Muller, Josh Andres, Alexander Gray, and Dakuo Wang. AutoAIviz: Opening the blackbox of automated artificial intelligence with conditional parallel coordinates. In *Proc. of the Int. Conf. on Intelligent User Interfaces, IUI'20*, pages 308–312, 2020.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020.

Christopher D Wickens, Justin G Hollands, Simon Banbury, and Raja Parasuraman. *Engineering psychology and human performance*. Psychology Press, 2015.

James Woodward. Sensitive and insensitive causation. *The Philosophical Review*, 115(1):1–50, 2006.

Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proc. of the Int. Conf. on Intelligent User Interfaces, IUI'20*, pages 189–201. ACM, 2020.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 2022.

Wencan Zhang and Brian Y Lim. Towards relatable explainable AI with the perceptual process. In *Proc.of the 2022 CHI Conf. on Human Factors in Computing Systems*, CHI 2022. ACM, 2022.

Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, 14(1):1–101, 2020.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proc. of the 2020 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2020*, pages 295–305, 2020.

Xun Zhao, Yanhong Wu, Dik Lun Lee, and Weiwei Cui. iForest: Interpreting random forests via visual analytics. *IEEE Transactions on visualization and computer graphics*, 25(1):407–416, 2018.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.