Dec 11th, 12:00 AM

Designing Interactive Explainable AI Systems for Lay Users

Miguel Angel Meza Martínez
Karlsruhe Institute of Technology (KIT), miguel.martinez@kit.edu

Alexander Mädche
Karlsruhe Institute of Technology (KIT), alexander.maedche@kit.edu

Designing Interactive Explainable AI Systems for Lay Users

Completed Research Paper

Miguel Angel Meza Martínez
Karlsruhe Institute of Technology
Karlsruhe, BW, Germany
miguel.martinez@kit.edu

Alexander Maedche
Karlsruhe Institute of Technology
Karlsruhe, BW, Germany
alexander.maedche@kit.edu

Abstract

Explainability considered a critical component of trustworthy artificial intelligence (AI) systems, has been proposed to address AI systems' lack of transparency by revealing the reasons behind their decisions to lay users. However, most explainability methods developed so far provide static explanations that limit the information conveyed to lay users resulting in an insufficient understanding of how AI systems make decisions. To address this challenge and support the efforts to improve the transparency of AI systems, we conducted a design science research project to design an interactive explainable artificial intelligence (XAI) system to help lay users understand AI systems' decisions. We relied on existing knowledge in the XAI literature to propose design principles and instantiate them in an initial prototype. We then conducted an evaluation of the prototype and interviews with lay users. Our research contributes design knowledge for interactive XAI systems and provides practical guidelines for practitioners.

Keywords: Explainability, Interactive Systems, Design Science Research, Artificial Intelligence

Introduction

Due to the high performance that artificial intelligence (AI) systems have achieved across a wide range of applications, they are increasingly being deployed in high-stake domains with the expectation of improving decision-making quality and efficiency (Binns et al., 2018). Nevertheless, despite their success, it has been shown that AI systems are prone to replicate biases, which results in unfair decisions that can have considerable consequences for individuals (Pfeuffer et al., 2023). While humans can also fail in their judgment and provide biased decisions, asking them for a rationale and holding them accountable is possible (Binns et al., 2018). In contrast, AI systems building on machine learning (ML) models can be extraordinarily complex and difficult to understand or audit (Dodge et al., 2019). The effectiveness of AI systems is limited by their ability to explain their decisions to human users (D. Wang et al., 2019). In particular, for AI systems deployed in high-stake applications, lay users who can be potentially affected by AI systems' decisions must understand the logic behind these decisions to trust and accept them (Fernández-Loría et al., 2022). As a result, regulations such as the European Union General Data Protection Regulation (GDPR) have been implemented to ensure users' "right to explanations" of all decisions made or supported by AI systems (Goodman & Flaxman, 2017).

To address AI systems' lack of transparency, many researchers and practitioners have resorted to the field of explainable artificial intelligence (XAI). Research in XAI aims to support human understanding and trust in AI systems by developing models that explain AI decisions to lay users in non-technical terms (Diakopoulos et al., 2017). In recent years, extensive research in XAI has developed many innovative explainability methods to provide explanations of AI systems (Carvalho et al., 2019). However, most of these XAI methods have focused on providing static explanations, such as highlighting relevant features through static visualizations (Liu et al., 2021). However, providing static explanations represents a one-way communication from AI systems to users that limits the amount of information conveyed to them, which can, in turn, result in an insufficient understanding of how these systems make decisions (Cheng et al., 2019; Liu et al., 2021).

Therefore, researchers have argued for enhancing explainable AI systems by allowing users to explore explanations interactively (Lombrozo, 2006; Miller, 2019). Recently, there have been efforts to explore how interactive XAI systems should be designed to improve their transparency and users' understanding of these systems' decisions (Liu et al., 2021). However, most studies have rather focused on developing interactive XAI systems for data scientists or domain experts (e.g., Hohman et al., 2019; Spinner et al., 2020). The explanations provided by interactive XAI systems are often too complex for lay users, making them very challenging to understand (Cheng et al., 2019; Miller, 2019). Therefore, it is necessary to design interactive XAI systems that provide explanations for lay users that support their understanding of AI systems' decisions. As a result, we aim to address the following research question in our work:

How to design interactive explainable artificial intelligence (XAI) systems to help lay users to better understand AI systems' decisions?

We rely on the design science research (DSR) methodology and existing design knowledge in literature to design and develop an interactive XAI system prototype to address this research question. Our prototype shows explanations based on SHAP, an XAI method that provides feature attribution scores using Shapley values from game theory (S. M. Lundberg et al., 2017). In this paper, we present the result of Cycle 1 of our DSR project, in which we conducted the first evaluation of our prototype and interviews with lay users. Our work contributes to design knowledge for XAI systems by demonstrating how interactive explanations can give users more control over the information they receive and help them better understand how AI systems make decisions. Moreover, our research provides practitioners with guidelines for designing and developing interactive XAI systems for lay users.

Related Work

Contemporary AI systems can be extraordinarily complex and difficult to understand (Dodge et al., 2019; Janiesch et al., 2021). They are designed to process large amounts of data to perform a complex optimization process for a specific performance measure. As a result, AI systems are often considered "black boxes" where only their output is available to users (Rudin, 2019). This lack of transparency leaves the inner workings mechanisms behind their decisions unclear to users. The inability of AI systems to explain their decisions to users is a critical limitation to their adoption and effectiveness (D. Wang et al., 2019). In particular, it is very challenging for AI systems deployed in high-stake applications to scrutinize them and identify potential biases that can have considerable consequences for individuals (Binns et al., 2018).

Providing explanations of AI systems' decisions has been proposed as a helpful means to increase the transparency of AI systems and enable users to understand the reasons behind these systems' decisions (Binns et al., 2018; D. Wang et al., 2019). Multiple studies have shown that providing explanations improves users' trust (W. Wang & Benbasat, 2007; Yang et al., 2020). Moreover, prior work has found that explanations can increase the likelihood that users agree with AI systems' decisions (Liu et al., 2021; Yeomans et al., 2019). As a result, explainability is becoming a critical component of trustworthy AI systems (AI HLEG, 2020). According to Ribera & Lapedriza (2019), the requirements for explanations depend on the target audience. They argue that it is necessary to identify the target users, their goals, background, and relationship to the system to design adequate explanations that ensure proper understanding.

In this light, there has been a recent surge of interest in XAI among scholars and practitioners seeking to increase the transparency of AI systems (Miller, 2019). As a result of the extensive research performed in different communities over the last few years, many innovative XAI methods have been developed. For instance, some methods extract easily interpretable rules from the predictive model and present them to users as an explanation of the model's decision (e.g., Jian et al., 2000). Alternatively, others highlight regions of an image to indicate which pixels were influential in the model's prediction (e.g., Zhou et al., 2017). Several studies have surveyed the literature to provide a detailed overview of XAI by presenting the different developed methods (Adadi & Berrada, 2018; Carvalho et al., 2019; Guidotti et al., 2018).

So far, most XAI methods provide static explanations that only reveal pre-defined information about AI systems' decisions (Liu et al., 2021; Ribera & Lapedriza, 2019). For example, some methods highlight sections of an input text to indicate the importance of certain features towards the system's prediction (e.g., Bansal et al., 2021), while others present a set of influential training examples (e.g., Koh & Liang, 2017). Alternatively, other methods rely on static visualizations to show each feature's influence on the systems' decision (e.g., Ribeiro et al., 2016). These static explanations represent a one-way communication from AI

systems to lay users, which can limit the information conveyed and may result in an insufficient understanding of how these systems make decisions (Cheng et al., 2019; Liu et al., 2021). Specifically, studies have found that lay users perceive static explanations as not transparent enough as they do not allow them to investigate further the factors that influence a given decision (Sun & Sundar, 2022).

Interactivity has been identified in the literature as an essential component of XAI systems that can help to address the challenges posed by static explanations (Lombrozo, 2006; Miller, 2019). Providing interactive explanations allows users to explore the system's behavior, giving them more control over the information they receive and a sense of agency that can promote trust in AI systems (Sun & Sundar, 2022). Recently, there have been efforts to start exploring how to design interactive XAI systems. In particular, some studies have focused on incorporating research in information visualization as it excels at knowledge communication due to the extensive work investigating how to transform abstract data into meaningful representations over hundreds of years (Friendly, 2008). For instance, Hohman et al. (2019) developed an interactive XAI system for data scientists that allows them to explore the factors influencing the decision of an individual instance or a group of instances and to search and compare the decision for similar instances. Their system relies on generalized additive models (GAMs) (Friedman, 2001) to generate explanations represented by interactive plots for each feature that data scientists can explore to observe the feature's impact on the system prediction. Meanwhile, Spinner et al. (2020) developed an interactive XAI system to support users in developing and debugging ML models. Their system allows users to explore visual explanations from multiple XAI methods to support model understanding, diagnosis, and refinement.

However, there have been critiques that interactive XAI systems developed so far are designed for users with a solid understanding of statistical and ML concepts (Cheng et al., 2019; Miller, 2019). For example, some of these approaches rely on diagrams such as scatter plots, area under the curve (AUC), or precision-recall graphs, which are known to be hard to understand for lay users (e.g., Amershi et al., 2015; Cabrera et al., 2019). While data scientists are familiar with these concepts, lay users often do not have the necessary knowledge to understand these interactive XAI systems' explanations. Therefore, researchers have called for designing interactive XAI systems that consider lay users' needs to support their understanding of AI systems' decisions (Cheng et al., 2019; Liu et al., 2021; Miller, 2019).

Research Method

To design an interactive XAI system for lay users, we followed the DSR methodology by Peffers et al. (2007). This methodology allowed us to provide a solution for a real-world problem. Specifically, we proposed design principles for an interactive XAI system, instantiated them in a prototype, and evaluated it with lay users. Figure 1 presents the overall DSR project consisting of two cycles. The focus of this paper is on the finalized Cycle 1.

Design Process	Design Cycle 1	Design Cycle 2
Identify Problems and Motivation	Lack of interactive XAI systems utilizing local model-agnostic explanations designed for end-users	Analyze the evaluation of the first prototype
Define Objectives	Design an interactive XAI system prototype for end-users	Improve the first prototype of an interactive XAI system
Design & Development	Design principles	Adapt design principles based on first evaluation results
Demonstration	First prototype of an interactive XAI for end-users utilizing local model-agnostic explanations	Second prototype of an interactive XAI for end-users utilizing local model-agnostic explanations
Evaluation	Qualitative evaluation (lab experiment and interviews)	Quantitative evaluation (lab experiment)
Communication	Submission	

Figure 1. DSR Project

Our DSR project relied on previous work investigating how lay users engage with static explanations from XAI methods (e.g., Binns et al., 2018; Dodge et al., 2019; Hase & Bansal, 2020), as well as research exploring

how to design interactive XAI systems (e.g., Hohman et al., 2019; Spinner et al., 2020). In Cycle 1, we analyzed the results from these studies to comprehend how explanations help lay users understand AI systems' decisions. Furthermore, we identified several challenges lay users face when interacting with AI systems due to the lack of explicit interactive explainability design for them. Afterward, we derived two meta-requirements of interactive XAI systems for lay users. Then, we proposed four refined design principles to address these meta-requirements and suggested three design features based on these principles, which were implemented in an interactive XAI system prototype. As a last step, we conducted an evaluation study and interviews with lay users.

As part of our DSR project presented in Figure 1, we plan to conduct one additional cycle to further improve the design of our interactive XAI system prototype. In Cycle 2, after reviewing the evaluation results of Cycle 1, we plan to refine our design principles and develop a second prototype of our interactive XAI system. Then, we plan to evaluate our prototype in an experimental study to quantitatively analyze how the interactive XAI system affects lay users' understanding and trust.

Conceptualization

Problem Awareness and Meta-Requirements

The first meta-requirement (MR1) refers to offering lay users explanations they can really understand. Even though extensive research has focused on developing XAI methods, there is no sufficient empirical evidence on whether the explanations these methods provide are understandable to lay users (Cheng et al., 2019). There is strong criticism that most of these explanations are based on researchers' and practitioners' intuition instead of a deep understanding of what lay users need (Miller, 2019; Ribera & Lapedriza, 2019). As a result, many of these explanations require a deep technical understanding of statistical and ML concepts (Miller, 2019). Moreover, the quality of explanations generated by these methods is often evaluated using a mathematical definition of interpretability without any user evaluation (Doshi-Velez & Kim, 2017). Besides, most research efforts investigating how to design interactive XAI systems have focused on understanding the requirements these systems must fulfill to assist data scientists (e.g., Hohman et al., 2019; Spinner et al., 2020). Therefore, it is necessary to investigate which type of explanations can be integrated into interactive XAI systems to help lay users understand how the system makes decisions.

MR1*: An interactive XAI system should be able to provide lay users with understandable explanations that reveal in non-technical terms the reasons behind its decisions.*

The second meta-requirement (MR2) refers to the system's capacity to allow lay users to request additional information regarding its decision logic. Several studies have found that explanations are often insufficient for users to fully understand the logic behind the system's decisions (e.g., Hase & Bansal, 2020; Kaur et al., 2019). Each XAI method relies on a different approach to provide explainability of AI systems. As a result, due to how explanations are generated, they focus on describing certain aspects of a given decision (Cheng et al., 2019). Some studies have found that users can perceive the system as not transparent enough due to the limited information provided by some of these explanations (Sun & Sundar, 2022). Therefore, researchers have argued for enhancing XAI systems to give users more control over the explainability information they receive (Krause et al., 2016; Miller, 2019).

MR2*: An interactive XAI system should allow lay users to request additional information if explanations are insufficient to understand the decisions.*

Design Principles

To address the two derived meta-requirements (MRs), we propose design principles (DPs) for interactive XAI systems to help users understand AI systems' decisions. Interactive XAI systems should provide lay users with understandable explanations that reveal information about how the system makes decisions according to their needs (MR1). Regarding their scope, XAI explanations are classified as either global or local. Global explanations provide a comprehensive and holistic description of the model behavior across all instances for a given dataset (Guidotti et al., 2018). This type of explanation is better suited for researchers or practitioners trying to improve the predictive model's performance or domain experts looking to learn from the system to improve their decision-making (Doshi-Velez & Kim, 2017; Ribera & Lapedriza, 2019). In contrast, local explanations describe how a particular system's decision was made by

considering the vicinity of the instance to be explained (Molnar, 2020). Local explanations can help justify a system's decision to lay users, for whom this decision can have a personal or economic impact (Doshi-Velez & Kim, 2017; Ribera & Lapedriza, 2019).

DP1*: Provide local explanations that reveal to lay users how a specific system's decision was made.*

The type of explanation an interactive XAI system provides is another critical factor in delivering adequate information to help lay users understand AI systems' decisions (MR1). XAI methods rely on different approaches to generate explanations, which influences the information disclosed by explanations. Research has shown that explanations from some of these XAI methods might not be sufficient for lay users to understand the reasoning behind decisions (Dodge et al., 2019; Hase & Bansal, 2020). For instance, Binns et al. (2018) found that lay users get frustrated with counterfactual explanations as they do not reveal which features had more influence on a decision. In this line, Doshi-Velez & Kim (2017) argue that explanations should provide information regarding the factors used in a decision and their relative importance.

DP2*: Provide explanations that disclose the factors influencing each decision and their relative weights.*

How explanations are presented to lay users also plays an essential role in their cognitive process to analyze the information they contain (MR1). In XAI research, textual explanations and visual charts are the two main approaches to presenting explanations to users. Research has found that visual representations help lay users understand XAI explanations. For instance, Cheng et al. (2019) found that explanations in the form of interactive visualizations helped to improve lay users' objective comprehension of the logic behind the system's decisions. Furthermore, Szymanski et al. (2021) found that lay users prefer more visual explanations than textual explanations because these provide an easier way to obtain an overview of the factors influencing a decision.

Nonetheless, research has also found that users can often misinterpret visual explanations when they are too complex or lack details due to poor design (Kaur et al., 2019; Szymanski et al., 2021). Several studies have found that lengthy and complex explanations are harder to understand for users and can overload their cognitive abilities (Narayanan et al., 2018; Poursabzi-Sangdeh et al., 2021). To reduce the complexity of explanations, many researchers have resorted to limiting the number of factors presented to users by showing only the most relevant influencing a decision (e.g., Binns et al., 2018; Hase & Bansal, 2020). However, such strategies can also result in counterproductive effects as users would have only limited information on the system's inner workings (MR2). An alternative approach is to utilize interactive visualizations that provide an overview of the most relevant factors influencing a decision while allowing users to request details about the additional factors.

DP3*: Provide interactive explanation visualizations that provide an overview of the most important factors influencing a decision and allow lay users to request details regarding the additional factors.*

XAI systems should allow lay users to request additional information about its logic (MR2). Nonetheless, many of the XAI methods proposed in the literature generate only static explanations that are insufficient to help lay users understand how AI systems make decisions due to the limited information they provide (Cheng et al., 2019; Ribera & Lapedriza, 2019). An interactive user interface has been proposed to empower lay users to explore and investigate how an AI system makes decisions. One strategy utilized in the literature is allowing users to modify the input feature values to observe how the system's decisions and explanations change accordingly (e.g., Cheng et al., 2019; Hohman et al., 2019). This interactive interface can enable lay users to evaluate counterfactual scenarios that reveal a causal relationship between the feature changes and the model decision (Molnar, 2020). Studies have found that interactive interfaces implementing this strategy can improve lay users' understanding of how AI systems make decisions (Cheng et al., 2019).

DP4: *Provide an interactive user interface that allows lay users to explore how changes in the input features affect AI systems' decisions.*

Prototype Implementation

To design and implement an interactive XAI system prototype, we propose design features (DFs) that represent specific system capabilities that aim to satisfy the proposed design principles (Meth et al., 2015). To instantiate our prototype, we decided to develop an interactive XAI system for the bank loan application domain, which is commonly used in XAI research because it involves the notion of trust in AI systems

(Adadi & Berrada, 2018; Aggarwal et al., 2019; Binns et al., 2018; Chakraborty, Majumder, et al., 2020) and lay users are familiar with applying for a loan in a bank. In the following, we briefly present the domain and describe the dataset and model used by our interactive XAI system prototype. Afterward, we describe the DFs in detail and explain how they help to address the design principles.

We selected the bank loan application domain in a scenario where an AI system predicts the decision to approve or reject loan applications. This scenario has been widely used in XAI research because lay users are familiar with the process of applying for a loan at a bank and because it allows researchers to investigate the notion of trust in the system (Binns et al., 2018; Chakraborty, Peng, et al., 2020). Moreover, this is considered a high-stake domain, where the decisions made by an AI system can significantly impact loan applicants (Binns et al., 2018). To train the predictive model, we relied on a publicly available, open-source dataset with 1,000 instances of bank loan applications and their corresponding decision (700 approved and 300 rejected). Each loan application is represented by 20 features describing the details of the loan application and the applicant's financial and personal information. We modified the original dataset by adjusting the features' names and descriptions and removing the two sensitive features, "personal status and sex" and "foreign worker". A neural network trained using the Python library Keras (Chollet, 2015) was used as the predictive model. To address the class imbalance in the training data, we incorporated a Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). Categorical features were one-code encoded, and continuous features were min-max scaled. A grid parameter search was performed to find the best hyperparameters and architecture. The architecture with the highest score consisted of 2-hidden layers, each with 65 and 33 neurons. The predictive model had an accuracy of 0.77 and an f1-score of 0.83.

To satisfy DP1 and DP2, we incorporated explanations from the XAI method Shapley Additive Explanations (SHAP) (S. M. Lundberg et al., 2017), which provides local and global explanations that present feature attribution scores using the concept of Shapley values (Shapley, 2016). Shapley values, which have a solid theoretical foundation in game theory, compute how the influence on the model's prediction is fairly distributed among the features used by the model (Molnar, 2020). According to Lundberg et al. (2017), SHAP builds on the concept of the popular method LIME (Ribeiro et al., 2016) to build a linear regression model with Shapley values as weights, which indicate how much influence each feature had on the system's decision. SHAP explanations are contrastive because Shapley values are calculated from all the possible feature value collisions across all dataset instances (Molnar, 2020). As a result, the prediction of one instance can be compared against the predictive model's average prediction.

Moreover, SHAP is a model-agnostic method that can generate explanations for any underlying predictive model. In contrast to model-specific and model-class-specific methods that provide explanations to only one predictive model or a specific model family (Sokol & Flach, 2020), model-agnostic methods offer great flexibility and scalability in their implementation due to their decoupling of explainability from the prediction (Ribeiro et al., 2016). Despite model-agnostic methods' benefits, most studies investigating how to design interactive XAI systems have focused on developing and evaluating systems that utilize non-model-agnostic methods to provide explainability (e.g., Cheng et al., 2019; Guo et al., 2022; Sevastjanova et al., 2021).

DF1*: Provide local model-agnostic explanations based on the XAI method SHAP, which relies on the concept of feature importance to explain how features influence the system's decision.*

SHAP has gained popularity in research and practice due to the unique consistency and local accuracy of the attribution values it provides. For instance, SHAP has been implemented by explainability libraries such as AIX360 (Arya et al., 2019) and InterpretML (Nori et al., 2019). Furthermore, SHAP explanations have been incorporated in several research studies investigating how to provide explainability of AI systems (e.g., Jesus et al., 2021; Kaur et al., 2019; Weerts et al., 2019). The interpretation of the feature attribution scores provided by SHAP explanations depends on the ML task performed by the predictive model. When explaining a regression model, SHAP scores represent the contribution of each feature value to the model's predicted value compared to the average prediction value. Thus, the scores can be directly presented as an increment or decrement from the average predictive value with the same units of measure as the target variable. In contrast, when explaining a classification model, SHAP scores represent the contribution to the average predicted class probability of the model.

Figure 2 shows an example of a SHAP explanation for our selected binary classification scenario using the visualization of SHAP's open-source library (S. Lundberg & Lee, 2016). The model's average predicted probability is represented by the "base value". The feature attribution scores are represented by arrows that increase or decrease the prediction probability for the explained instance. Adding the base value and the scores results in the model's prediction probability represented by "f(x)".

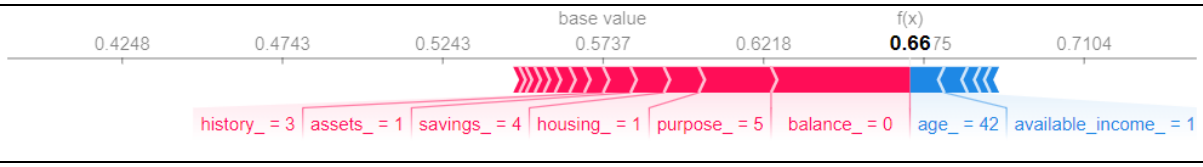

Figure 2. Example of a SHAP explanation using the original visualization.

To satisfy DP3, we designed two interactive visualizations for our system prototype, which provide an overview of each feature's influence on the model decision while only showing the details of the most influential features (i.e., name and attribution score). Nevertheless, lay users can hover over the explanation elements of the visualization to observe the details of the additional features.

***DF2**: Display an interactive visualization of SHAP's explanations highlighting the most influential features and allowing lay users to see details for the rest of the features.*

Figure 3 presents the two designs of the interactive visualizations integrated into our prototype. The system interface provides an overview of the loan application's details in the upper section by showing the features' values across the categories: financial information, personal information, and loan details. Lay users can hover over the information icon at the top of the screen to see a detailed description of each feature. The interface also shows the system's decision recommendation and prediction probability in the top right corner. The probability is presented to lay users as a confidence level in a percentage, together with a text legend indicating one of the system's three levels of confidence on the recommendation (low, medium, or high confidence).

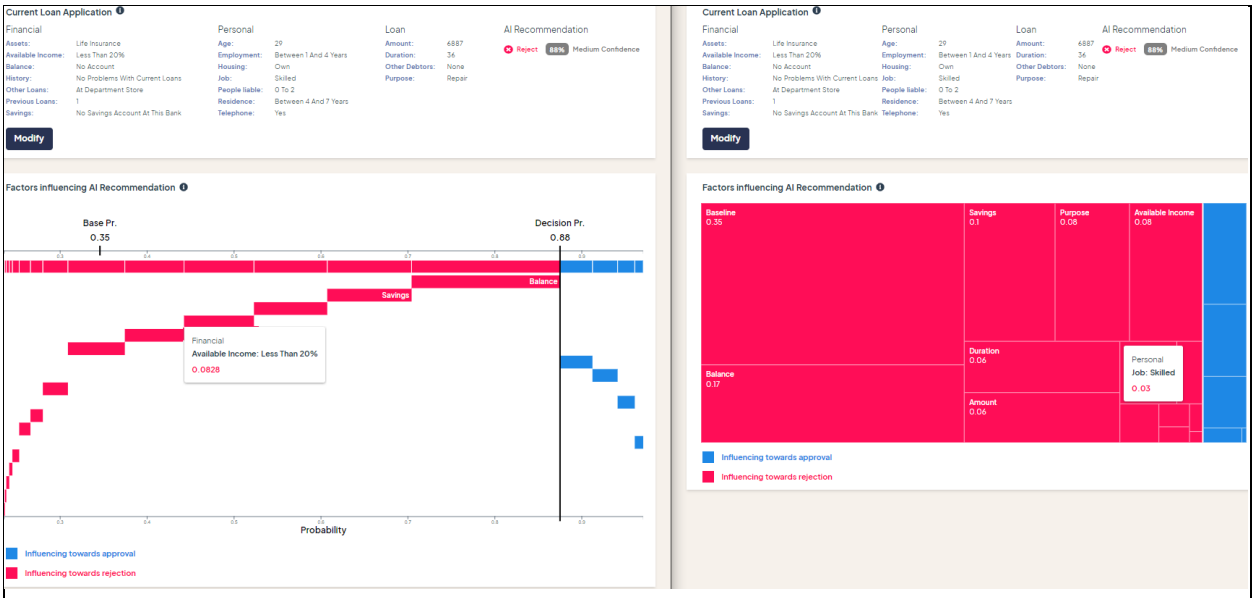

Figure 3. Design of interactive visualizations for SHAP explanations. The figure shows the cascade visualization on the left and the treemap visualization on the right.

The "cascade" visualization on the left side of Figure 3 is an adaptation of SHAP's original visualization. In the design of this visualization, we maintained the overview provided by the stacked bars from the original

visualization. Nonetheless, we included an individual bar for each feature below, and for the most influential features, we included their names on the bar. We maintained SHAP's original color coding to represent features contributing to approval with blue bars and rejection with red bars. We used the label "Base Probability" to indicate the model's average predicted probability and the label "Decision Probability" to show the prediction probability of the explained instance. We decided to show each class prediction probability instead of a complementary probability below 0.5. Thus, for approved instances, we show the features increasing the probability in blue and decreasing it in red, while for rejected instances, this was inverted (see Figure 4). The interactive visualizations allow lay users to hover over each bar to see the feature name, value, and corresponding attribution score.

The "treemap" visualization presented on the right side of Figure 3 was designed to provide a simpler visualization without the probability axis. In contrast to the cascade visualization, the treemap visualization uses boxes to represent the attribution scores. The box size representing each feature corresponds to the magnitude of their score. Moreover, this visualization utilizes the same color coding as the cascade visualization and shows the features' names and attribution scores for the most influential features. In contrast to the cascade visualization, the features influencing approval are always located on the right side of the visualization, while the features influencing rejection are on the left. Lay users can hover over the boxes to see the details of the corresponding feature. The treemap visualization includes the model's average predicted probability as an additional box with the description "Baseline".

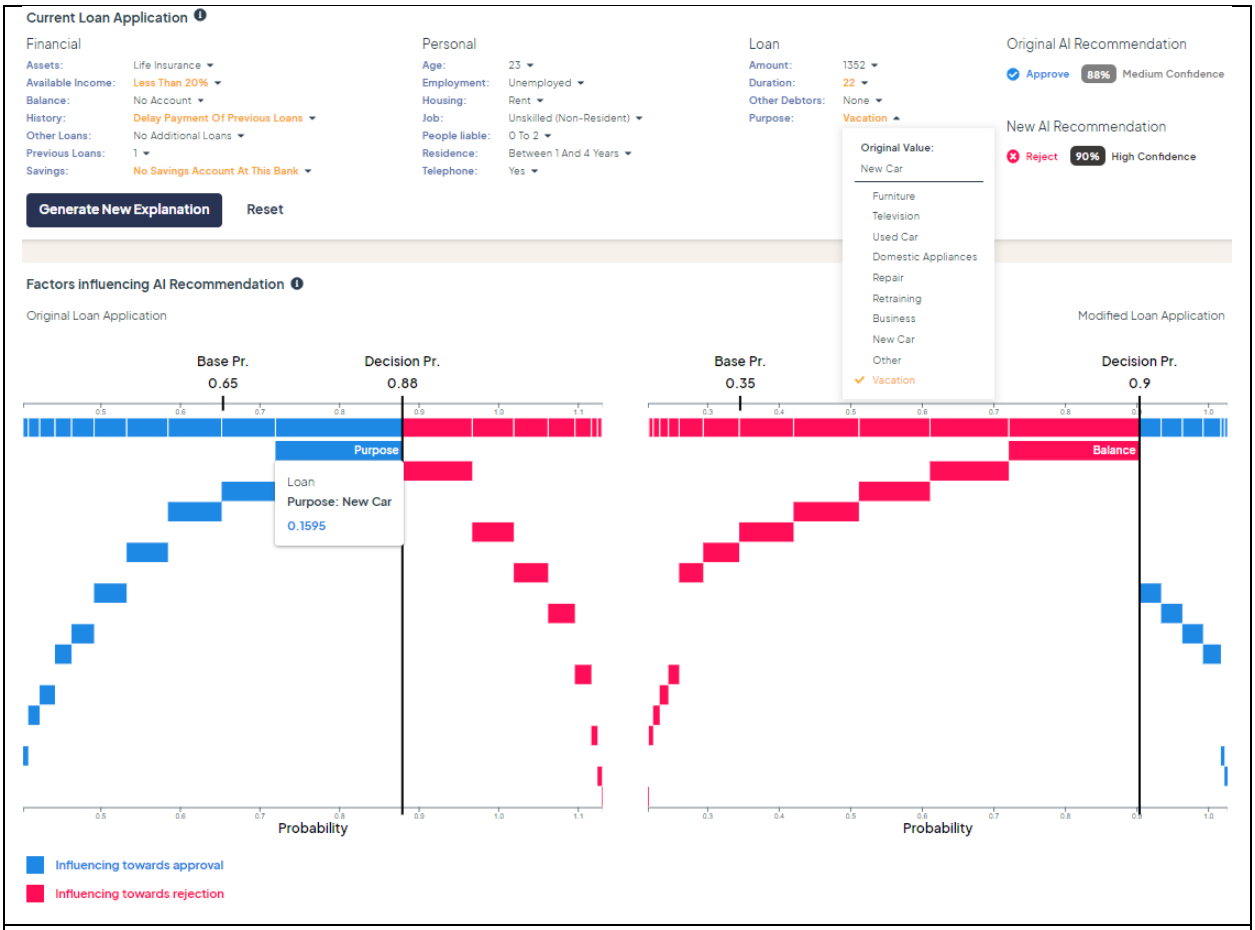

Figure 4. Design of interactive what-if functionality.

The base probability for both visualizations is shown according to the predicted class. Our predictive model's average prediction value is 0.35, representing the dataset's oversampled 300 rejected instances.

Thus, for rejected instances, a base probability of 0.35 is shown. In the case of approved instances, we show the complementary base probability of 0.65, representing the 700 approved instances.

To satisfy DP4, we instantiated a "what-if" interactive functionality that allows lay users to explore how modifications to the features' values of the instance being explained affect the system's decisions.

DF3*: Provide an interactive user interface that allows lay users to explore "what-if" scenarios by changing the features' values and observing the system's decision and corresponding explanation visualization.*

Following DP3, the what-if functionality is disabled by default. Lay users can activate it by clicking on the "Modify" button in the left part of the screen below the instance feature details (see Figure 3). When this button is clicked, all features display a caret-down icon next to the original value to indicate that modifying the values is now possible, as shown in Figure 4. The "Modify" button is replaced by a "Reset" button designed to revert any modifications and turn off the what-if functionality. Lay users can click on any feature value or its corresponding caret-down icon to open a drop-down menu that allows the modification of the original value. The drop-down menu lists valid values for categorical features and an adjustable slider for numerical features. To provide an overview of the features' values that have been modified, the interface highlights them using orange text. When there is at least one modified value, the interface displays the decision the system would make and its corresponding confidence level below the original decision on the right side of the screen. Moreover, the interface shows the "Generate New Explanation" button that provides the corresponding interactive visualization for the SHAP explanation of the modified instance, as shown in Figure 4.

Evaluation

We conducted an evaluation study to assess the design principles and the instantiated interactive XAI system prototype in Cycle 1 of our DSR project. In the evaluation study, participants interacted with one of five configurations of our prototype to understand better how they perceived the different design features. The five configurations were: what-if without visualization (what-if), cascade visualization without what-if (cascade), treemap visualization without what-if (treemap), cascade visualization with what-if (cascade-what-if), and treemap visualization with what-if (treemap-what-if).

The evaluation study consisted of four phases. First, participants were randomly assigned to one configuration and were introduced to the scenario and the features used. Second, participants were presented with information describing the system's interface, the visualization, and the what-if functionality according to their corresponding configuration. Third, participants were asked to interact with the system by reviewing eight loan applications (four approved and four rejected) with the corresponding system's decision recommendation, confidence level, and design features according to their group. Participants were asked whether they would approve or reject each loan application. Fourth, participants were asked to respond to questions regarding their demographics and their evaluation of the design features through self-reported measures. After the evaluation, semi-structured interviews were conducted with participants to discuss their perceptions of the system.

Through a research panel at our university, we recruited 21 students as proxies for lay users applying for a bank loan, 11 females and ten males. Eighteen participants were between 18 and 25, while three were between 26 and 30. Thirteen participants were coursing a bachelor's degree, seven a master's degree, and one a doctoral degree. Following Cheng et al. (2019), we asked participants about their familiarity with the task of credit scoring using a four-level scale (i.e., no experience, a little experience, some experience, a lot of experience). Nineteen stated they had no experience, and only two said they had little experience. Moreover, we asked participants to indicate their knowledge level of machine learning (Cheng et al., 2019). Seven had no previous knowledge, 11 had little knowledge, two had some knowledge, and one had a lot of knowledge.

The distribution of participants across the five groups was: what-if (4), cascade (4), cascade-what-if (4), treemap (5), and treemap-what-if (4). Each participant was paid 12.00€ for participating in the evaluation study and the interviews. The average duration for the evaluation study was 30.19 minutes (SD = 8.12) and 11.28 minutes (SD = 10.66) for the interviews. All interviews were recorded and transcribed for analysis.

We relied on an open coding strategy to analyze the transcripts and extract participants' evaluations of the elements of our interactive XAI system prototype (Myers, 2002).

The analysis of the semi-structured interviews revealed that participants had positive and negative feedback about the different design features of our prototype. Moreover, participants also made some suggestions for improvements that we plan to incorporate in Cycle 2. Table 1 summarizes the evaluation study across the most relevant elements of our prototype. The percentages shown in Table 1 relate to the number of participants that interacted with the corresponding elements of the prototype according to their assigned group.

Regarding their general perception of the system, a third of the participants indicated that it was fun to use the system. P1 and P11 indicated that it was fun because it was possible to see how the system works. Four participants indicated that the system was easy to use. However, eight participants raised concerns about the features used by the system to make decisions and the granularity of their categorical values. For instance, P9 mentioned that "*savings are only considered for this bank*", and P15 indicated that it would be good to "*get more customer data*". Four participants indicated that they would like to understand how features influence decisions for other instances, indicating that some participants would like to receive global explanations revealing the model behavior across all instances.

Moreover, three participants indicated that the confidence level was very helpful, allowing them to "*see when the system was not really sure*" about a particular decision (P8). Nonetheless, two participants said it was confusing that the system provided decisions with low confidence. P1 stated that "*low confidence [decisions] should be reviewed by a person*". In this line, although many participants liked the explanation visualizations and what-if functionality, fourteen participants indicated they would like decisions to be made in a human-system collaboration. P19 stated that it would be preferable to have the "*mathematical [reasoning]*" of the system, which can be "*very precise*", in combination with the "*human element*".

Element	**Positive Feedback**	**Negative Feedback**	**Suggested Improvements**
System	Fun to use (33.3%) Easy to use (19.0%)	Problems with features (38.1%)	Global explanations (19.0%)
Confidence level	Shows how sure it is (14.3%)	Some low-confidence decisions (9.5%)	Human-system decision (66.7%)
DF1: cascade visualization	Helps understand decisions (87.5%) Reveal feature importance (25.0%)	Not clear how weights are calculated (25.0%)	Clarify how weights are calculated (25.0%)
DF2: cascade visualization	Easy to understand (75.0%) Satisfying design (37.5%) Overview of features (25.0%)	Base probability is confusing (25.0%) Colors change depending on the decision (25.0%)	Do not change the position of colors (25.0%)
DF1: treemap visualization	Helps understand decisions (100%) Reveal feature importance (11.1%)	No information on how weights are calculated (66.7%)	Clarify how weights are calculated (66.7%)
DF2: treemap visualization	Easy to understand (77.8%) Overview of features (77.8%) Satisfying design (33.3%)	Baseline is confusing (44.4%) No control over baseline (11.1%)	Show values in all boxes (11.1%)
DF3: what-if functionality	Easy to use (75.0%) Helps understand decisions (66.7%) Possible to analyze alternative scenarios (50%)	Lack of feature importance (16.7%) Unrealistic modifications (8.3%)	Limit modifications to some features (8.3%) Input field beside slider (8.3%)

Table 1. Summary of evaluation of elements from our interactive XAI system prototype with the percentage of interviewees mentioning each point.

Regarding the evaluation of DF1, six participants interacting with the cascade visualization and all nine interacting with the treemap visualization indicated that the provided local explanations based on SHAP helped them understand how the system makes decisions. Expressly, four participants, two of each visualization, indicated that it was beneficial that the influence relevance of the features was shown. P21 stated about the cascade visualization that "*it gives you a good feeling [to know] about how important some aspects are*", while P4 said regarding the treemap that it is essential to know "*what factors are influencing [decisions]*". However, two participants in the cascade groups and six in the treemap groups mentioned that they would like the system to clarify how the influence weights of each factor are calculated. For instance, P20 stated, "*I wish more transparency … to see how it calculates [the weights]*".

Concerning DF2, two participants of the cascade groups and seven participants of the treemap indicated that the visualizations provided a good overview of the features' influence on the decision. For instance, P6 stated about the treemap visualization, "*It was very clear what factors were in favor and what factors were against the approval*". Similarly, for the cascade visualization, P1 stated, "*The graphic makes it quite clear what the system is doing … without the graphic, it would be … impossible to understand how the system works*". Moreover, three participants from the cascade group and three from the treemap group indicated that the design of the visualizations was good. For the cascade visualization, P1 stated that "*the design was really satisfying*", while P17 said it "*looked beautiful*". Meanwhile, for the treemap visualization, P20 stated that "*it is easy to understand because it uses complementary colors and has squares of different sizes*".

Nonetheless, there were also some critiques about the visualizations: Two participants of the cascade group and three of the treemap group indicated that the base probability and baseline were confusing. P4 said over the baseline that it is unclear whether "*it is a strategical decision*" and an "*influential factor that you do not have control over*". Likewise, P14 did not understand the base probability and why it was sometimes shown as 0.35 and others as 0.65. Moreover, for the cascade visualization, two participants indicated that it was confusing that the colors were inverted in the graph for approved and rejected instances. P7 said it took some time to get used to this change, while P1 said they should not change.

Regarding the evaluation of DF3, the what-if functionality was considered by nine participants as easy to use. Additionally, eight participants indicated that it was helpful to understand how the system makes decisions. For instance, P15 said that with the modifications, it was possible to see "*how to get [higher confidence]*". Moreover, six participants indicated that the what-if functionality allows them to analyze alternative scenarios. P1 said that "*the possibility to modify the criteria and see the new recommendation was useful*". P1 also mentioned that it was possible to compare the two graphics next to each other. However, two participants from the what-if group criticized that there was no information on the relevance of each feature for the decision. P3 said that "*it would be nice to know how much each component is weighted*", while P18 said that knowing how "*each variable affects the recommendation*" would be helpful. Additionally, P4 indicated that modifying certain features would not present a realistic scenario because applicants can not change some of their personal information. To address this, P4 suggested allowing only changes to some of the features. Finally, P3 suggested allowing writing the modification values on an input field in addition to the current slider shown in the drop-down menu of the numerical features.

The evaluation revealed that the design features utilized to instantiate our proposed design principles helped participants understand how the system makes decisions. Most participants considered that both the cascade and treemap visualizations proposed to represent SHAP explanations were easy to understand. Participants also appreciated that the provided explanations disclosed how much influence each feature had on the decisions. Moreover, participants indicated that the what-if functionality was easy to use and allowed them to analyze alternative scenarios to understand how features affect decisions.

However, participants also highlighted difficulties in understanding some aspects of our proposed interactive XAI system prototype. Regarding the visualizations proposed to represent SHAP explanations, most negative feedback was related to how the model's average prediction is displayed. Several participants indicated this concept was confusing and highlighted that they could not fully understand what this value

represented despite the detailed information received in the evaluation's introductions. Regarding the what-if functionality, some participants indicated that modifying certain features would be unrealistic because loan applicants cannot modify some aspects, such as how long they have been working in a company or living in their current address.

Discussion

Design Challenges

In our work, we designed and developed an interactive XAI system prototype to provide explainability for lay users as part of the first cycle of a larger DSR project. The evaluation of our prototype with lay users revealed several challenges in designing interactive XAI systems. First, lay users found that modifying certain features leads to unrealistic scenarios when analyzing counterfactual scenarios through the what-if functionality. Lay users felt they could not modify certain features to get approval if their loan application was rejected. Therefore, it seems that it might be helpful to restrict modifications for specific features in an interactive XAI system in certain scenarios.

Second, lay users had problems understanding the model's average prediction concept on both the cascade and the treemap visualization. For explanations of XAI methods that rely on weights from a regression model to represent each feature's influence on the decision, the intercept represents the model's average prediction across the dataset. Some XAI methods, such as LIME, do not incorporate this intercept as part of their explanation and instead only show the features' influence on each class. Nonetheless, this expected prediction value reflects the skew towards a given class in the dataset for classification tasks with imbalanced datasets such as the one used in our scenario. Thus, failing to disclose the average prediction value can result in contra-intuitive explanations that show more features influencing the opposite class than the one predicted by the system. In an ideal scenario, having an equal class distribution in the training dataset would result in an intercept value in the regression model that has an insignificant effect on the model's decision. However, there are many applications in which it is very challenging to achieve an equal class distribution in the training data because one class is significantly underrepresented (e.g., positive diagnostics in the health domain). In cases where an equal class distribution can be achieved, the baseline box could be removed from the treemap visualization simplifying the explanations by focusing only on each feature's influence on the system's decision. For the cascade visualization, the base probability would still be displayed with an approximate value of 0.5 for each predicted class.

Limitations and Future Work

There are limitations in our work conducted in Cycle 1 of our DSR project that need to be addressed by future research. First, our proposed design principles were instantiated in an interactive XAI system prototype evaluated in the bank loan application domain. This domain was selected as a representation of high-stakes domains where the decisions of AI systems can significantly impact individuals. Nevertheless, other domains can significantly differ in important factors that can have implications on how our proposed design principles are instantiated and how they are perceived by lay users. For instance, explanations might need to be adapted to account for the risk of disclosing proprietary information in highly sensitive domains. Therefore, it is necessary to investigate how our proposed design principles can be instantiated in interactive XAI systems developed for other domains with different characteristics. Moreover, these systems should be evaluated with targeted lay users in those domains to investigate if the explanations they provide can help them understand these systems' decisions.

Second, our prototype was implemented for a classification task using a tabular dataset with relatively few features. While our design principles provide guidelines that can be easily adapted to other tasks and types of datasets, our design features instantiated in our interactive XAI prototype might need to be adapted for different conditions. For instance, when dealing with a dataset with a significantly higher number of features, it might be very challenging to present the influence of all features on the decision, as the boxes or bars representing the influence of the least relevant features might not be visible at all. One possible way to address this challenge would be to aggregate features below a certain threshold and display them together in the visualizations. Lay users could then display the details of these features by utilizing a drill-down functionality. Likewise, the actual implementation of the what-if functionality might not be appropriate to

allow lay users to change the inputs for text or image data. Instead, a suitable interactive user interface would need to be designed to allow input modifications for these data types.

We plan to address these limitations in Cycle 2 of our DSR project by instantiating our proposed design principles across two application domains and tasks. Thus, we plan to develop two independent interactive XAI system prototypes that provide explanations for different datasets to investigate how our derived meta-requirements and proposed design principles can be generalized. Furthermore, we plan to conduct quantitative evaluations with a larger sample size by recruiting target lay users of the corresponding application domains.

Theoretical and Practical Implications

As AI systems increasingly support decision-making in high-stake applications, lay users affected by these systems' decisions must have access to explanations that help them understand the reasons behind these decisions to trust and accept them (Fernández-Loría et al., 2022). To address these requirements, research in the field of XAI has proposed models that provide explanations in non-technical terms to support lay users understanding (Diakopoulos et al., 2017). Nonetheless, despite these research efforts, it has been shown that most of the developed XAI methods provide static explanations that limit the amount of information conveyed to lay users, resulting in insufficient understanding (Cheng et al., 2019; Liu et al., 2021).

Our study contributes design knowledge for interactive XAI systems by demonstrating how explanations in the form of interactive visualizations can give lay users more control over the information they receive. Interactive visualizations can transform abstract data into meaningful representations, which help to provide an overview of the most relevant factors influencing a decision to lay users. Additionally, these interactive visualizations allow lay users to explore details about additional factors not presented in an overview. As a result, these interactive explanations give lay users a sense of agency that can promote understanding and trust in AI systems (Sun & Sundar, 2022). Moreover, our study provides insights into how lay users interact with different elements of interactive XAI systems and how these elements can help them understand AI systems' decisions.

Furthermore, our study derived meta-requirements from existing knowledge in the literature and then proposed four design principles to address them. Afterward, we proposed three design features to instantiate the proposed design principles into an interactive XAI system prototype. Through these proposed design principles and design features, our study offers practical guidelines for researchers and practitioners in designing interactive XAI systems. We also provide a GitHub open-source repository with the implementation of our system prototype and the software architecture design.[1] Therefore, researchers and practitioners can rely on our work to continue exploring how to design interactive XAI systems and investigate how they help lay users understand their decisions.

Conclusion

In our work, we argue about the importance of designing interactive XAI systems for lay users to assist them in understanding AI systems' decisions. Relying on the DSR methodology and existing design knowledge provided by the XAI literature, we performed the first design cycle of our DSR project. We derived two meta-requirements for interactive XAI systems designed for lay users. To address these meta-requirements, we proposed four design principles that we then instantiated into an interactive XAI system prototype. An evaluation of our prototype and interviews with lay users revealed that our proposed design features implemented in our initial prototype could help users understand how the system makes decisions. However, we also identified several potential improvements that we plan to address in Cycle 2 of our DSR project. Our work contributes to the XAI literature by identifying design knowledge for developing interactive XAI systems to increase lay users' understanding and trust. Furthermore, with the development of our interactive XAI system prototype instantiating our design principles, we provide researchers and practitioners with guidelines on giving users more control over the information they receive to help them better understand how AI systems make decisions.

[1] https://github.com/miguelmezamartinez/interactive_xai_system

Acknowledgments

The authors would like to thank Mario Nadj, Felix Hasse, Nicolas Kiefer, Isabelle Konrad, and Tilio Schulze for their support on this research project and the development of the Interactive Explainable AI System.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2019). Black Box Fairness Testing of Machine Learning Models. Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 625–635. https://doi.org/10.1145/3338906.3338937

AI HLEG. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). https://doi.org/10.2759/002360

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 337–346. https://doi.org/10.1145/2702123.2702509

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2019). One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. ArXiv Preprint ArXiv:1909.03012.

Bansal, G., Fok, R., Ribeiro, M. T., Wu, T., Zhou, J., Kamar, E., Weld, D. S., & Nushi, B. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3411764

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3173574.3173951

Cabrera, A. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). Fairway: A Way to Build Fair ML Software. Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 654–665. https://doi.org/10.1145/3368089.3409697

Chakraborty, J., Peng, K., & Menzies, T. (2020). Making Fair ML Software using Trustworthy Explanation. 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE), 1229–1233.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3290605.3300789

Chollet, F. (2015). Keras. In Keras. GitHub. https://github.com/fchollet/keras

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., & Zevenbergen, B. (2017). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. FAT/ML. https://www.fatml.org/resources/principles-for-accountable-algorithms

Dodge, J., Vera Liao, Q., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. Proceedings of the 24th International Conference on Intelligent User Interfaces, 275–285. https://doi.org/10.1145/3301275.3302310

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. ArXiv Preprint ArXiv: 1702.08608.

Fernández-Loría, C., Provost, F., & Han, X. (2022). Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach. MIS Quarterly, 46(3), 1635–1660.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232.

Friendly, M. (2008). A Brief History of Data Visualization. Handbook of Data Visualization, 15–56. https://doi.org/10.1007/978-3-540-33037-0_2

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." AI Magazine, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5). https://doi.org/10.1145/3236009

Guo, L., Daly, E. M., Alkan, O., Mattetti, M., Cornec, O., & Knijnenburg, B. (2022). Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. 27th International Conference on Intelligent User Interfaces, 22, 537–548. https://doi.org/10.1145/3490099.3511111

Hase, P., & Bansal, M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5540–5552.

Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300809

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine Learning and Deep Learning. Electronic Markets, 31(3), 685–695. https://doi.org/10.1007/S12525-021-00475-2/TABLES/2

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How Can I Choose an Explainer? An Application-grounded Evaluation of Post-hoc Explanations. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 805–815. https://doi.org/10.1145/3442188.3445941

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics, 4(1), 53–71. https://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401_04

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2019). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3313831.3376219

Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. Proceedings of the International Conference on Machine Learning, 1885–1894.

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 5686–5697. https://doi.org/10.1145/2858036.2858529

Liu, H., Lai, V., & Tan, C. (2021). Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 45. https://doi.org/10.1145/3479552

Lombrozo, T. (2006). The structure and function of explanations. Trends in Cognitive Sciences, 10(10), 464–470. https://doi.org/10.1016/J.TICS.2006.08.004

Lundberg, S., & Lee, S. (2016). Shap. https://github.com/slundberg/shap

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 4765–4774.

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. Journal of the Association for Information Systems, 16(9), 2. https://doi.org/10.17705/1jais.00408

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Myers, M. D. (2002). Qualitative Research in Information Systems: A Reader. SAGE.
Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. ArXiv Preprint ArXiv:1802.00682.
Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. ArXiv Preprint ArXiv:1909.09223.
Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302
Pfeuffer, N., Baum, L., Stammer, W., Abdel-Karim, B. M., Schramowski, P., Bucher, A. M., Hügel, C., Rohde, G., Kersting, K., & Hinz, O. (2023). Explanatory Interactive Machine Learning. Business & Information Systems Engineering, 1–25. https://doi.org/10.1007/S12599-023-00806-X
Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–52. https://doi.org/10.1145/3411764.3445315
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778
Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. Proceedings of the IUI Workshops, 2327, 38.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
Sevastjanova, R., Jentner, W., Sperrle, F., Kehlbeck, R., Bernard, J., El-Assady, M., Jentner, W., Sperrle, F., Kehlbeck, R., & El-Assady, M. (2021). QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3–4), 1–38. https://doi.org/10.1145/3429448
Shapley, L. S. (2016). 17. A Value for n-Person Games. In Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press. https://doi.org/10.1515/9781400881970-018/HTML
Sokol, K., & Flach, P. (2020). Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 56–67. https://doi.org/10.1145/3351095.3372870
Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2020). ExplAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. IEEE Transactions on Visualization and Computer Graphics, 26(1), 1064–1074. https://doi.org/10.1109/TVCG.2019.2934629
Sun, Y., & Sundar, S. S. (2022). Exploring the Effects of Interactive Dialogue in Improving User Control for Explainable Online Symptom Checkers. Conference on Human Factors in Computing Systems - Proceedings, 1–7. https://doi.org/10.1145/3491101.3519668
Szymanski, M., Millecamp, M., & Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. International Conference on Intelligent User Interfaces, Proceedings IUI, 109–119. https://doi.org/10.1145/3397481.3450662
Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-driven User-Centric Explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3290605.3300831
Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. Journal of Management Information Systems, 23(4), 217–246. https://doi.org/10.2753/MIS0742-1222230410
Weerts, H. J. P., van Ipenburg, W., & Pechenizkiy, M. (2019). A Human-Grounded Evaluation of SHAP for Alert Processing. ArXiv Preprint ArXiv:1907.03324.
Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? Proceedings of the 25th International Conference on Intelligent User Interfaces, 20, 189–201. https://doi.org/10.1145/3377325.3377480
Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. Journal of Behavioral Decision Making, 32(4), 403–414. https://doi.org/10.1002/bdm.2118
Zhou, J., Arshad, S. Z., Luo, S., & Chen, F. (2017). Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making (pp. 23–39). Springer, Cham. https://doi.org/10.1007/978-3-319-68059-0_2

Appendix

Guide for Semi-structured Interviews

General Perception of the System

- How did you feel in general about your interaction with the AI system?
- Do you think the system was reliable?
- Do you think the system provided fair recommendations?
- How would you feel if bank employees used this system when deciding on loan applications you apply for in a bank?

Perception of System Functionality

- Do you think that the system's functionality helps you understand why the system made a certain decision recommendation for a loan application?
- How would you feel if the system didn't provide functionality that allows you to understand how it makes decision recommendations?

Perception of What-if Analysis

- What do you think about the system's functionality that allows you to modify the attributes of the bank loan application to observe how the system's decision recommendation would change?
- Do you think that this functionality is useful? Does this function help you to understand how the system makes decision recommendations?
- Was this functionality easy to understand? Was it easy to use?
- Would you change anything regarding this functionality to make the system better?
- Could you imagine another way the system could provide you with an alternative functionality to help you more?

Perception of Explanations

- What do you think about the system's explanations for how it made each decision recommendation?
- Were the explanations clear and easy to understand?
- Do you think that providing such explanations helps you understand why the system made a certain decision recommendation for a loan application?
- How would you feel if the system didn't provide explanations for its decision recommendations?
- Would you change anything regarding the explanations to make the system better?
- Could you imagine another way the system could provide you with an alternative explanation to help you more?