Impact of Explanation Techniques and Representations on Users Comprehension and Confidence in Explainable AI

ANONYMOUS AUTHOR(S)*

Local explainability, an important sub-field of eXplainable AI, focuses on describing the decisions of AI models for individual use cases by providing the underlying relationships between a model's inputs and outputs. While the machine learning community has made substantial progress in improving explanation accuracy and completeness, these explanations are rarely evaluated by the final users. We therefore evaluate the impact of various explanation and representation techniques on users' comprehension and confidence. Through a user study on two different domains, we assessed three commonly used local explanation techniques—feature-attribution, rule-based, and counterfactual—and explored how their visual representation—graphical or text-based—influences users' comprehension and trust. Our results show that the choice of explanation technique primarily affects user comprehension, whereas the graphical representation impacts user confidence.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence.*

Additional Key Words and Phrases: Machine Learning, Interpretability, Explainability, User Studies

ACM Reference Format:
Anonymous Author(s). 2024. Impact of Explanation Techniques and Representations on Users Comprehension and Confidence in Explainable AI. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (CSCW '24).* ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Artificial Intelligence (AI) algorithms have become ubiquitous for decision-making, even in high-stakes domains such as law [5, 76] and healthcare [13, 25]. This has raised numerous critical questions and concerns. One of those concerns arises from the fact that current AI algorithms can be incredibly complex, which makes algorithmic decision-making opaque—*i.e.,* the algorithms behave like black boxes [68]. One approach to tackling this challenge is to make AI algorithms more explainable. This is the main goal of the field of eXplainable AI (XAI). By improving the transparency of AI systems, the XAI research community aims to increase people's confidence and comprehension in AI systems, and thereby facilitate their adoption [29, 42, 63].

Over the last five years, the XAI community has primarily focused on developing methods to compute local explanations for AI models. These approaches explain the reasoning of an AI system when applied to an individual case, *i.e.,* a **target instance**, and can be categorised into three broad 'explanation families': feature-attribution, rule-based, and counterfactual [9, 26, 28, 35]. A large number of explanation methods exist, with several of them being widely adopted by data practitioners [27, 48, 63, 64]. Despite the rise of various XAI methods, numerous works have pointed out a lack of end-user involvement in the assessment of XAI methods [1, 4, 24, 65]. For instance, Adadi et al. [1] found that across 381 XAI articles, only 5% of articles explicitly evaluated the proposed methods through a user study. This implies that novel explanation methods are frequently

CSCW '24, October 09–13, 2024, San José, Costa Rica

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

published without a clear understanding of how the intended end-users perceive or interpret these explanations.

In contrast to the majority of work found in the XAI and ML communities, user studies on AI explanations have been commonplace within the wider Human-Computer Interaction (HCI) community [16, 43, 77]. This line of work underscores the importance of evaluating the impact of explanations on comprehension (i.e., do users better understand the AI system thanks to the explanation?), and confidence (i.e., to which extent explanations increase or decrease users' confidence in AI recommendations?). Nevertheless, existing studies typically focus on specific use cases, for example in a distinct domain, involving a single explanation type, or considering a small and very specific cohort (*e.g.*, CS students). Furthermore, these studies usually rely on hand-crafted explanations rather than explanations generated by real-world AI systems. This creates a barrier to the adoption of these results to other XAI scenarios, and is also unable to provide comparative evidence of the suitability of the different explanation techniques that are used in the real world. In this paper, we seek to address this limitation by studying the impact of feature-attribution, rule-based, and counterfactual explanations on users' comprehension and confidence in AI-based recommendations. Given the known effect of visual representations on human perception of information [16, 77], our investigation also includes a comparison of the effect of the explanation's visual representation on users' comprehension and confidence.

Our investigation consists of a user study on a cohort of 280 crowd-workers who were given an AI-assisted prediction task across two high-stakes use cases: prediction of the risk of obesity and recidivism. The AI agents operate on tabular data and were enhanced with explanations. Those explanations were computed using established explanation methods, *i.e.*, LIME [63], Anchors [64], and Growing Fields [21, 45]. The contributions of our work include:

(1) Two user studies evaluating the impact of (a) the three aforementioned explanation techniques, and (b) two visual representations (graphical vs. text) on users' confidence and comprehension.
(2) A methodological framework for user studies designed to measure the impact of AI explanations on users' confidence and comprehension;

Our study shows that the explanation technique impacts primarily users' comprehension, whereas the choice of a graphical representation has a greater impact on users' confidence. Graphical representations are perceived as more trustworthy, whereas rule-based explanations are most effective at conveying the relevant features of an AI's decision process. The results of our studies inform a set of recommendations for XAI practitioners and researchers.

2 RELATED WORK

Our work lies at the intersection of eXplainable AI, HCI, and data visualisation. Thus, we first review the most prominent local XAI techniques that motivate this research. Next, we discuss user evaluations of XAI systems. This is followed by a survey of existing guidelines and metrics used to conduct user studies on XAI tools.

2.1 XAI Techniques for Local Explanations

An AI model is an agent that takes an instance x as input and returns an output. The instance x is composed of features, *e.g.*, attributes of a person for tabular data, image pixels, or words in a text. The output can be a class, *e.g.*, low risk vs. high risk, or a number, *e.g.*, a price estimation. An explanation is an expression that describes, in an understandable manner, the relationships between the AI model's inputs and outputs [47]. Explanations can be computed via a post-hoc explainability module, or extracted directly from the model (for white boxes). When the explanation focuses on a single instance, we say it is a *local explanation*. These have lately received more attention from researchers in machine learning (ML) [35]. Based on prominent XAI surveys [9, 28], we can categorise these explanations into three main types (see Figure 2):

Feature-attribution explanations. These explanations provide the contribution of the input features to a black box's output on a target instance. The magnitude of the contribution informs us of the importance of the feature for a particular prediction outcome, whereas the sign denotes a positive or negative correlation with that outcome. Besides classical white boxes such as linear regression, there exists a range of methods that can compute such scores from black-box models in a post-hoc manner. Some of them work for specific models, such as neural networks [71, 74], whereas others such as LIME [63] and SHAP [48] are model-agnostic. This has made them popular among researchers and practitioners. While we used LIME in our study, SHAP could have also been a viable alternative.

Rule-based explanations. Approaches such as Anchors [64] and LORE [27] compute explanations under the form of decision rules on the input features. Anchors is model- and data-agnostic and resorts to a bandit exploration to compute a single general and accurate decision rule that mimics the black box's behaviour on the target instance [64], whereas LORE operates on tabular data and learns a decision tree trained on artificial instances that resemble the target instance [27]. Explanation rules can, therefore, be extracted from this decision tree. We chose Anchors for our experiments since it provides a single explanation rule without additional computation.

Counterfactual explanations. These explanations convey the minimum adjustments required in the target instance to modify the AI model's prediction. They, therefore, identify the most *sensitive* features within the AI agent's decision process. Counterfactual explanations are similar to adversarial examples as they both perturb an instance in order to change a model's prediction. However, their objectives differ. Adversarial examples aim to deceive the model to test the robustness of ML models and, therefore, rely on non-perceptible perturbations in the input data [36]. Counterfactual explanations, on the other hand, do not have this constraint because they aim to be actionable and understandable. Methods such as Growing Spheres [45], FACE [62] or DICE [59] perturb the target instance, *i.e.*, they create new instances by increasingly modifying various attributes in the target instance until they identify an instance that changes the model's prediction. Our experiments use the Growing Fields algorithm [21], an extension of the Growing Spheres [45] that supports both continuous and categorical attributes. We opted for this algorithm because of its simplicity. Contrary to other approaches [59, 62], it does not impose additional constraints on the counterfactuals (e.g., diversity, likelihood), whose evaluation lies beyond the goal of our study.

2.2 Evaluating Explainable AI Systems

Explainability is an inherently human-centric property. Consequently, Miller argues that the development of effective explanation modules requires the joint effort of the XAI and HCI research communities [55]. While the HCI community has emphasised the need for human-centred evaluations for XAI systems [24, 46], several surveys have highlighted the scarcity of XAI papers that evaluate their novel explanation methods through user studies [1, 4, 24]. Among the works that carried out user studies, most assessed either the validity of their novel explanations method [41, 49, 63, 64, 66, 86] or the impact of the explanation's visual representation [16, 60, 61, 85]. A limitation of these works is that they are typically limited to the evaluation of one kind of explanation technique [41, 60, 66] and one application domain [61, 86]. Some prominent explanation methods, such as LIME [63] and Anchors, evaluate the quality of the explanations with a small number of computer science students who are already familiar with machine learning [64]. In our work, we set out to compare three different explanation techniques on two distinct datasets with lay users.

To study the impact, and thus the benefit, of explanations, prior works have mostly evaluated users' trust and understanding in highly specific settings [6, 16, 38, 43, 78]. For instance, Arora et al. [6] studied the impact of interactive explanations on users' understanding. The results of this study confirmed that explanations help users identify key elements for the prediction. Cheng et al. [16] compared the effect of interactive versus static explanations, as well as black-box versus white-box models, on users' trust and understanding. They observe that both white boxes and interactive explanations are beneficial to users' comprehension.

Other researchers have studied the influence of the explanation's representation on users' perceptions [16, 77]. Van Berkel et al. compared textual and scatterplot representations and showed that the usage of a scatterplot visualisation led to lower perceived fairness [77]. Other works have compared the effects of different explanation methods on users [6, 38, 78]. For instance, Van der Waa et al. compare hand-crafted example-based and rule-based explanations for the self-management of diabetes [78].

In this study, we contribute to this research body by providing a comprehensive evaluation that compares three real explanations, generated by the well-established methods LIME, Anchors, and Growing Fields, rather than hand-crafted explanations by domain experts. We compare these methods to two novel visual representations, namely graphical and textual. Following the recent guidelines for evaluating XAI applications [78], we experiment with a large cohort (280 participants), on two diverse datasets, and we collect both perceptual and behavioural metrics for users' confidence and comprehension. To our knowledge, this combination of factors has not been previously investigated.

2.3 Guidelines and Metrics to Conduct User Studies

The evaluation of trust and understanding for XAI systems has been inspired by research in psychology and cognitive sciences, which have produced numerous guidelines for measuring such cognition aspects [33, 37, 79]. Cahour and Forzy [14] formulated a trust scale based on three factors: reliability, predictability, and efficiency. This scale, comprising four questions, directly asks participants about their confidence in the XAI system. Madsen and Gregor [50] proposed an eight-question scale to measure perceived technical competence and comprehension.

Ribeira and Lapedriza [65], as well as Doshi-Velez and Kim [24], classify users into three distinct groups: (a) machine learning practitioners, (b) domain experts, and (c) laypeople. Building upon these three categories, Doshi-Velez and Kim propose to distinguish between application-grounded and human-grounded evaluations. The former involves real-world tasks conducted by computer scientists or domain experts, while the latter includes simplified (and synthetic) tasks, such as providing individuals with input and an explanation and asking them to simulate the model's prediction. Doshi-Velez and Kim also indicated that running evaluations with laypeople offers the advantage of (a) evaluating the impact of the explanations more broadly, and (b) simplifying the execution of the experiments since factors are easier to control.

Our work evaluates the impact of the explanation technique and visual representation with lay users on a human-grounded task. Following the advice from Van der Waa et al. [78], we evaluated the impact of the explanations on two complementary aspects, including confidence and comprehension.

3 EXPLANATION TECHNIQUES AND REPRESENTATIONS

We first present the two datasets, the ML models and the explanation methods used for the experiments. The explanation representations are then introduced.

Gender	Female
Age	23
Height	166
Family member has overweight	No
Frequent consumption of high caloric food	No
Frequency of consumption of vegetables	Sometimes
Number of daily meals	More than 3
Consumption of food between meals	Sometimes
Smoke	No
Consumption of water daily	More than 2L
Calories consumption monitoring	Yes
Physical activity frequency per week	2 or 4 days
Time using technology devices daily	0-2 hours
Consumption of alcohol	Sometimes
Transportation used	Public transportation

Gender	Male
Age	26
Race	Other
Number of juvenile major offences	0
Number of juvenile minor offences	4
Number of previous arrest	3 or more
The degree of the charge	major offences
Description of the charge	Aggravated assault with a deadly weapon

Fig. 1. Example of two individuals presented to the participants for the Obesity (left) and COMPAS (right) datasets.

3.1 Datasets & AI models

Datasets. Our evaluation is conducted on two datasets widely used by the XAI community [2, 11, 19, 39, 73, 87], namely COMPAS [12] and Obesity [52]. COMPAS is a tabular dataset collected in the USA and used to train a model that predicts a criminal defendant's likelihood of re-offending. The Obesity dataset [52] is used to predict the risk of developing obesity based on an individual's body mass index (BMI) and answers to various questions, with data originating from Colombia, Peru, and Mexico[1]. Figure 1 displays a snapshot featuring one individual from each dataset.We selected these datasets as they represent two high-stakes domains that concern everyone and for which explainability and user confidence are deemed important: justice (recidivism) and healthcare (obesity) [3, 80]. We decided to focus on more than one domain following the recommendations from the literature [77, 78] that suggest that a meaningful application-agnostic XAI evaluation should preferably include more than one domain, and strike a balance between **simplicity**—participants should understand the domain of the AI—, and **plausibility**—the task should be difficult enough to justify the need for AI assistance. Detailed information about the datasets is available in Appendix A.

AI Model and Explanations. We trained a multi-layer perceptron (MLP) classifier[2] on each dataset for our experiments. We selected this model due to its predictive power, and its status as a true black-box model. Its decision boundary is too complex to be easily understood by simply examining model parameters. We remark that other powerful black-box models, such as random forests or gradient-boosting trees, would have also been suitable for this task. We trained the MLPs on 70% of the instances and evaluated them on the remaining 30%. An accuracy of 67% and 78% was obtained on COMPAS and Obesity, respectively. Although these accuracy levels might appear low, they are consistent with those observed in the literature [44, 82]—and remained concealed from the participants to avoid any influence on their confidence in the model. On COMPAS, the AI agent was trained to predict the risk of recidivism among four classes: 'very low risk', 'low risk', 'high risk', and 'very high risk'. The original Obesity dataset considers seven weight categories which we simplified into four ordinal classes (to stay consistent with COMPAS): 'underweight', 'healthy', 'overweight', and 'obese'. Then, for each instance in the test set, we generated three different explanations: a feature-attribution explanation based on LIME [63], a rule-based explanation based on Anchors [64],

[1]We removed the weight from the obesity dataset, which otherwise would have oversimplified the prediction task. The task, therefore, becomes to predict the risk of obesity given a patient's eating and activity habits.
[2]https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

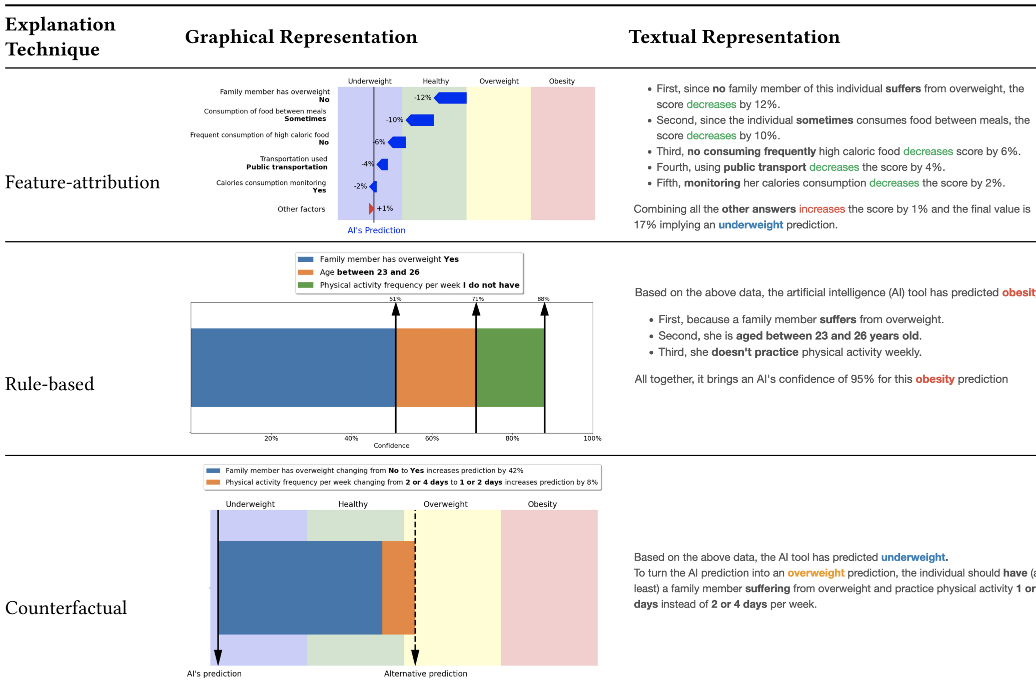


Fig. 2. Different explanations for a given individual on the Obesity dataset.

and a counterfactual explanation using Growing Fields [21]. The methods were used with the default parameters except that (a) Anchors used the discretisation proposed by Delaunay et al. [20], and (b) we computed the attribution of all features in the LIME explanation—contrary to the default configuration that only picks the top 6.

For each dataset, we selected five target individuals in the test set to be presented to participants—one for each of the four predicted classes plus an additional individual used as an example. Figure 1 depicts how information about an individual is shown to the participant for both datasets. The grey column represents the various features while the corresponding prisoner or patient data are in the second column. The code, the datasets, and the experimental results are available on GitHub[3].

3.2 Common Representation for Explanations

Since the studied explanation types do not offer the same exact insights into the AI's prediction process, the explanations are usually conveyed using different representations, which also depend on the nature of the data (*e.g.*, image, tabular, text, etc). When it comes to tabular data, existing XAI toolkits[4] opt for a graphical representation based on bars for feature-attribution explanations – as illustrated in Figure 2. Conversely, for rule-based and counterfactual explanations, the most common representation is natural language (see Figure 2). To control for this visual representation in our experiments, participants are confronted with common graphical and textual representations for all the explanation types.

Graphical Representation. For each explanation method, we depict the graphical representation through diagrams. As our AI models predict four ordinal target outcomes, we choose a

[3]https://anonymous.4open.science/r/user_eval-1776
[4]AI360, Dalex, H2O, eli5, InterpretML, What-if-Tool, Alibi, Captum.

common graphical representation that depicts the spectrum of classes on the x-axis and adds a different background colour to the region covered by each of the classes.

- As proposed by LIME [63] for feature-attribution explanations, the x-axis depicts the contribution of each feature to the predicted class in the form of a directed bar. The length of the bar denotes the magnitude of the attribution, whereas its direction tells us towards which side of the spectrum the feature biases the AI model's prediction (underweight vs. obese, low risk vs. high risk). To keep the explanation's complexity under control, our representation groups features with a marginal attribution under an artificial feature labelled 'Other features'. The aggregated attribution of this label is the sum of the attribution scores of those features (for more details read Appendix D.1).
- For rule-based explanations, we took inspiration from Molnar [58]. Our representation uses stacked bars as well, where each condition of the rule is assigned to a bar with a length proportional to the increase in confidence provided by the condition. Consider the explanation rule in Figure 2, stating that "(a) having family antecedents of obesity, (b) an age between 23 and 26, (c) and practising no physical activity" incurs an "obese" prediction with 90% confidence. The blue bar shows that condition (a) on its own predicts obesity with 50% confidence; conditions (a) and (b) increase the confidence to 71%, and all three conditions increase the confidence to 90%.
- For counterfactual explanations we also employ stacked bars. Each feature in the explanation incurs a hypothetical change of value and is associated with a bar. The length of the bar is proportional to the change incurred in the model's prediction when the value of the input feature is changed. For instance, the counterfactual explanation from Figure 2 says that if the patient: "(a) had family antecedents of obesity, and (b) practised less often physical activity" then the AI model would have predicted "overweight" (the counterfactual class) instead of "underweight".

Text Representation. For all explanation types we present the explanation as a bulleted list. The list is a manual transcription of the contents of the charts starting from the most impactful feature—as reported by the explanations. This transcription was reviewed and validated by all authors. Each item from the list describes the effect of each feature on the model's answer. This effect can be an increase in the confidence of the prediction (for rule-based explanations), how much the feature contributes to the model's prediction (feature-attribution explanation), or how sensitive the AI model is in regards to the changes in the input features (counterfactual explanation). For feature-attribution explanations, we used colours to highlight the direction of the impact of each feature. Finally, the AI model's outcome (*e.g.*, obesity, high-risk) is highlighted in bold.

4 METHOD

While the XAI community has proposed multiple post-hoc explanation methods based on feature attribution, rules, and counterfactual instances, no user studies have compared the impact on users' confidence and comprehension for all these explanation styles. This motivates our first research question **RQ1: "How do local explanation techniques, *i.e.,* feature-attribution, rule-based, or counterfactuals, affect users' confidence and comprehension of an AI model?".** Existing works have shown that explanations improve users' ability to comprehend a model [6, 63]. Hence, this question underlies our first general hypothesis; (H1) **explanations improve the participants' confidence and comprehension of a model**. In addition, we observe that unlike other explanation types [61], decision rules have consistently demonstrated high efficiency in helping users understand the inner mechanisms of a model [6, 64]. This leads to our second hypothesis; (H2) **rule-based explanations contribute the most to participants' comprehension of a model**. In regards to confidence, existing works have failed to show significant improvements in the presence of explanations [61, 78]. We, therefore, follow a more exploratory approach to study the impact of the explanation technique on confidence and do not hypothesise on this aspect.

As suggested in [16, 77], the visual representation of an explanation impacts the users' perception. This leads to our second research question **RQ2: "Does the explanation's visual representation impact the users' confidence and comprehension?".** As the general tendency is to represent feature-attribution explanations graphically and both counterfactual and rule-based explanations textually, our hypotheses are as follows; **for feature-attribution explanations, graphical representations improve users' confidence and comprehension** (H3), whereas **a textual representation elicits higher confidence and comprehension for rule-based and counterfactual explanations** (H4).

Our study seeks to elucidate the relationships between users' comprehension and confidence in the AI model (dependent variables), based on two (i) the explanation style—feature-attribution, rule, or counterfactual—and (ii) the visual representation—graphical or textual (independent variables). Since this requires an elaborated experimental protocol, the paper also contributes with a general workflow (Section 4.1) and a set of scales and metrics (Section 4.2) to conduct such kinds of experiments. These resources are intended to guide other researchers in XAI interested in measuring the impact of explanations on users' confidence and comprehension.

4.1 Task

Our user study follows a between-subject design, in which each participant interacts with one representation and one explanation style across a total of four prediction tasks. These tasks aim to predict either the risk of recidivism of a defendant given their profile or the risk of obesity of a person given some information about their habits. To perform those predictions, participants count on the recommendations of the AI models described in Section 3.1, in addition to an explanation of that recommendation. We created these surveys on Qualtrics[5], for each dataset (COMPAS and Obesity), explanation technique (feature-attribution, rule-based, counterfactual) and representation (graphical vs. textual). For each dataset, we also defined a control group for which participants did not get any explanation. Figure 3 outlines the process followed by each of these surveys. Given a dataset, the only difference among the seven surveys is the explanation provided to the participant. Each survey is composed of three phases:

Introduction. The experiment starts with an introductory description of the tasks assigned to the participant and the information used by the AI model to make recommendations (cf. Figure 1). We subsequently asked participants two questions to verify whether they understood the task.

Task Round. After explaining the experiment, participants are presented with four prediction tasks, each comprising two steps. First, participants assess the risk of either obesity or recidivism based on the provided information and indicate their confidence in their prediction on a 5-point Likert scale. Following this assessment, the participants have access to the AI model's prediction along with its associated explanation (cf. Figure 2). Based on this explanation, we then asked participants to select the features, among all possible features, that were used by the AI model to make its recommendation. Participants can reconsider their prediction and answer two questions to report their understanding of the explanation ('Immediate Explanation Understanding', see Figure 3) and their confidence in their prediction ('User Prediction Confidence') on a 5-point Likert scale.

Post-Questionnaire. After the prediction tasks, the participants answer a 8-question questionnaire where they can report their understanding regarding the AI model.

4.2 Scales & Metrics

To assess the impact of our independent variables—visualisation and explanation technique—, we employed a range of scales and metrics to evaluate users' confidence and comprehension. These

[5] https://www.qualtrics.com/

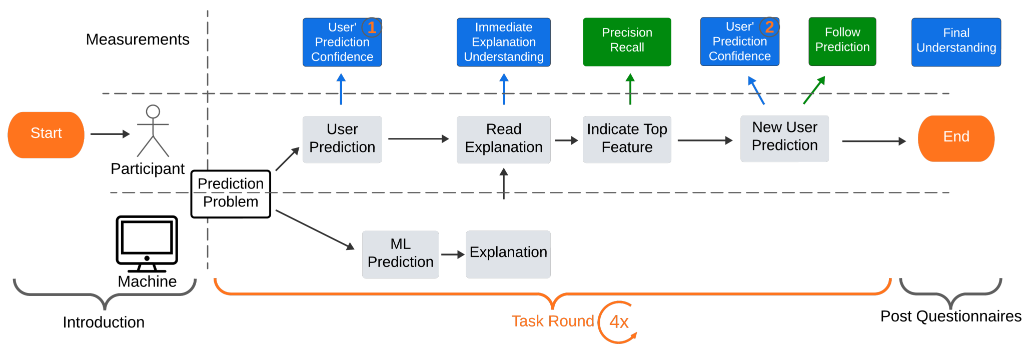


Fig. 3. Experimental workflow used to assess participant perception and behaviour when interacting with a given explanation technique. Behavioural measurements are in green, while self-reported measurements are in blue. The task round is repeated for four different prediction problems.

elements are frequently identified as crucial measurements in human-centred XAI [32, 56, 67]. It has been shown through several user studies that users' perception of comprehension and actual comprehension may disagree [16, 17, 31]. Therefore, we distinguish between the subjective nature of self-reported and behavioural (actual comprehension/confidence) metrics. Figure 3 shows when these parameters are measured (a detailed example of the measurement process is provided in Appendix E).

Confidence. A common measure of user confidence is the agreement rate between the users and the AI model [10, 75, 81]. Therefore, we build upon the methodology of Broon and Holmes [10] to measure users' behavioural confidence.

- **Behavioural Confidence (Follow Prediction):** Proportion of times the participants modified their prediction in favour of the AI's prediction (only when the initial participant's prediction differs from the model's).
- **Self-Reported Confidence (Δ Confidence):** This is the difference between the self-reported confidence before and after accessing the AI-based predictions and explanations ('User Prediction Confidence 2' - 'User Prediction Confidence 1' in Figure 3).

Comprehension. A widely accepted definition of a good explanation is its capacity to be understood by a human within a reasonable time frame [47]. We thus gauge the users' comprehension of the model through four aspects divided into two behavioural and two self-reported metrics.

- **Behavioural comprehension (Precision and Recall):** Building upon the methodology proposed by Weld and Bansal [84], we assess users' behavioural understanding through a simple quantitative task [72]. We ask participants to identify the features that have the most impact on the classifier's prediction according to the explanation. This task evaluates their ability to interpret the information provided by the explanations. Since understanding is a multifaceted process, we acknowledge that these measures capture a specific, still meaningful, aspect of it.
  - **Precision.** It measures the proportion of features correctly identified by the participant among *all the features* selected by the XAI method. It is computed as the number of properly identified features divided by the number of selected features.
  - **Recall.** It computes the ratio of features correctly identified by the participant among *all the correct features* (the features in the explanation).

Surprisingly, participants do not always indicate the features present in the explanation and may select features they think the model should consider (see Section 5.1).
- **Self-Reported Understanding (Immediate and Final Understanding)**:
  - **Immediate Understanding.** This is the self-reported comprehension of the system prediction on a five-point Likert scale during the explanation review.
  - **Final Understanding.** This was obtained from an adapted questionnaire by Madsen and Gregor [50] on perceived technical competence and comprehension on a 5-point Likert scale.

4.3 Participants

We recruited participants through the Prolific Academic platform. We restricted participation to crowdworkers with at least a high school degree to guarantee a reasonable response quality. We also decided not to limit ourselves to a particular geographical location to promote diversity in our sample. Finally, we ensured that participants could participate only once in our study. After accepting the task, participants were redirected to the corresponding Qualtrics survey. Based on a pilot evaluation with 20 people, we estimated a completion time of 20 minutes for the non-control group and 15 minutes for the control group. The control group received the AI prediction without an explanation and was thus expected to fill out the survey faster. Participants were paid £9.30 per hour, which translated into £2.25 for the control-group participants, and £3.10 for the non-control group.

To limit Type II errors, we determined the number of respondents on the basis of a power calculation using G*Power [69]. Given the exploratory nature of our research, we used medium-to-large effect sizes ($f^2 = 0.2$), an alpha level of 0.05, and a power of 0.8, in line with established methodological recommendations [30]. For an *a priori* multiple linear regression model with two predictors, the required minimum group size is 107 participants. We finally recruited 280 participants—140 participants per dataset, or 20 participants per combination of explanation technique and visual representation. Table 4 in Appendix B presents the demographic information of our participants. We recruited crowdworkers as they are commonly relied on by researchers and companies for data labelling tasks [23]. With the growing interaction and collaboration between crowdworkers and (explainable) AI systems, for example to assist in data labelling, it is vital to investigate their perception and response to the provided explanations. We stress that crowdworkers do not capture the particularities of all user types, *e.g.*, domain experts. We discuss this limitation in Section 6.4.

Following the task introduction, we assessed whether the participants had actually read and understood the task through two questions: 'How is Body Mass Index calculated?' for the Obesity dataset and 'Why is recidivism risk calculated?' for COMPAS. We found a total of 40 incorrect answers and replaced these participants from our study with new participants.

5 RESULTS

We present our findings in three sections. We begin by studying the impact of the domain (*i.e.*, dataset), explanation technique, and representation on users' comprehension in Section 5.1. Then, we assess the influence of these factors on users' confidence in the AI agent in Section 5.2. We explore the correlation between behavioural and perceived measurements in Section 5.3. All the experimental resources of our study are available on Github[6].

To discern the factors that impact users' confidence and comprehension of our AI agents, we employed a linear model and an ANOVA analysis for each application domain (recidivism and obesity). The linear model uses demographic data (age, gender, education level) along with

[6] https://anonymous.4open.science/r/user_eval-1776

	Understanding							
	Recidivism				Obesity			
	Self-Reported		Behavioural		Self-Reported		Behavioural	
	Immediate	Final	Precision	Recall	Immediate	Final	Precision	Recall
Technique	0.87	1.20	**16.24*****	1.58	**3.75***	1.35	**31.42*****	**6.37*****
Representation	0.96	0.36	0.13	3.00	0.14	0.55	0.05	2.85
Age	1.07	0.01	1.88	0.10	0.16	0.06	6.41*	0.02
Education	1.63	0.93	0.94	0.43	0.50	0.34	0.25	1.31
Gender	0.54	1.07	0.35	0.30	0.14	0.03	0.18	0.36
Technique:Representation	0.28	0.87	1.12	0.74	0.48	0.16	0.35	4.99**

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1. F value of the ANOVA Table with understanding measurements grouped for each domain by self-reported and behavioural metrics. 'Technique:Representation' denotes the interaction between explanation technique and visual representation.

explanation technique and visual representation as predictive variables. These predictors were categorised to fit the linear model. For each statistically significant predictor, we conduct a post hoc analysis using t-tests with Bonferroni correction to discern statistical differences for each pair of categories of the predictive variables—which we depict with box plots.

5.1 Comprehension

The ANOVA F-scores of each predictor and target comprehension metric can be found in Table 1. We first observe that the users' self-reported understanding of the AI system—based on a post questionnaire **(Final)**—does not vary across the different explanation techniques, visual representations, and demographic categories. These observations hold for both domains. Conversely, when we focus on the self-reported comprehension right after seeing the explanations **(Immediate)**, we observe a statistically significant effect *(p<0.05)* for the explanation technique in the Obesity dataset. Concerning behavioural comprehension, Table 1 highlights that **precision** is significantly affected by the explanation method in both domains *(p<0.001)*, whereas a significant impact on **recall** is only observed in the Obesity dataset.

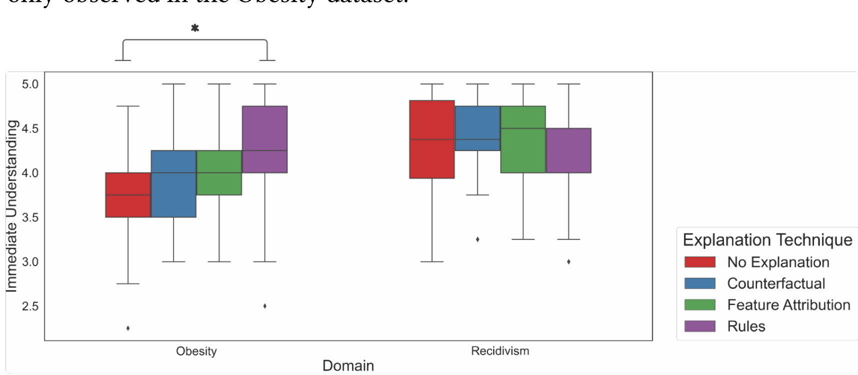


Fig. 4. Perceived understanding of the users **(Immediate)** for both the Obesity and Recidivism domains based on the explanation technique.

Figure 4 depicts the users' perceived comprehension of the AI system across the explanation methods for both domains. Users confronted with rule-based explanations in the obesity domain report a better understanding of the model w.r.t. the control group.

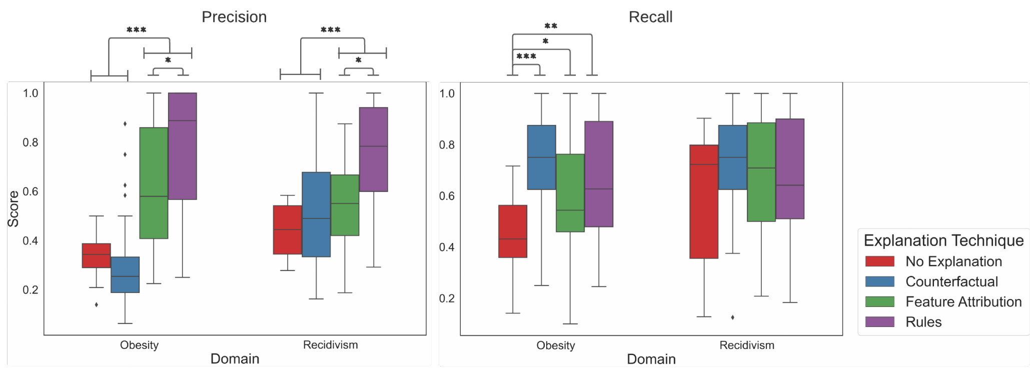


Fig. 5. Precision and recall between the features indicated as important by the users for the AI's prediction and the features indicated in the explanation. Results are shown for each explanation technique and domain.

Figure 5 depicts the precision and recall across domains and explanation methods, revealing that rule-based explanations yield the highest precision score in the obesity domain (median precision of 0.9). On the contrary, counterfactual explanations resulted in poor performances comparable to the control group (precision 0.3). Concerning the participants' recall, we observed that in the Obesity domain, participants who faced explanations obtained significantly higher recall than participants without explanations.

	confidence			
	Recidivism		Obesity	
	Self-Reported	Behavioural	Self-Reported	Behavioural
	Δ Confidence	Follow Prediction†	Δ Confidence	Follow Prediction†
Technique	1.40	0.78	0.12	0.38
Representation	0.04	0.00	**8.22****	0.12
Age	0.46	2.76	0.06	0.00
Education	0.13	0.34	2.14	0.63
Gender	2.16	0.31	0.12	1.11
Technique:Representation	0.35	0.75	0.26	**3.55***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2. F value of the ANOVA Table with confidence measurements grouped by domain and by self-reported and behavioural metrics. 'Technique:Representation' refers to the interaction between the explanation technique and representation († = the metric was computed only on the initial disagreement participants).

5.2 Confidence

We now assess users' confidence in the AI system and report the corresponding F-values in Table 2. Our ANOVA analysis suggests that changes in self-reported confidence before and after seeing the explanation **(Δ Confidence)** are significantly impacted by the explanation visual representation in the Obesity dataset. It is noteworthy that, on average, users' predictions aligned with the AI's in 56% of the cases in the COMPAS dataset, and in 39% of the cases in the Obesity dataset. Thus, we limit our evaluation of behavioural confidence to scenarios where participants are prompted to reconsider their own predictions. We call those participants initial disagreement participants. We find that for the Obesity dataset, the interaction between explanation technique and visual

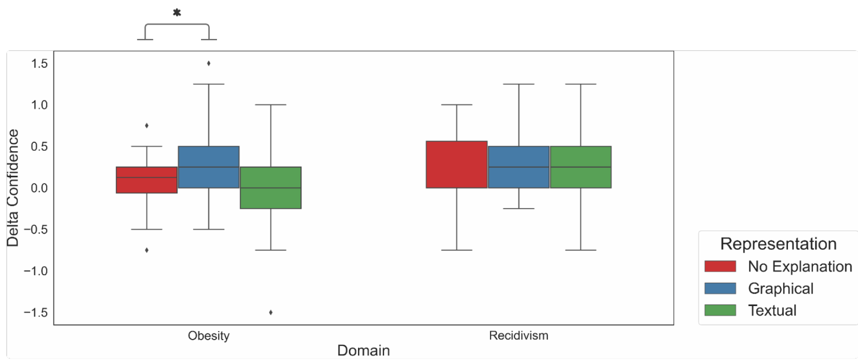


Fig. 6. Difference between the self-reported confidence in the users' prediction after and before seeing the AI's prediction and explanation (when provided). Results are shown for each domain and representation. Values above zero denote an increase in confidence in the model.

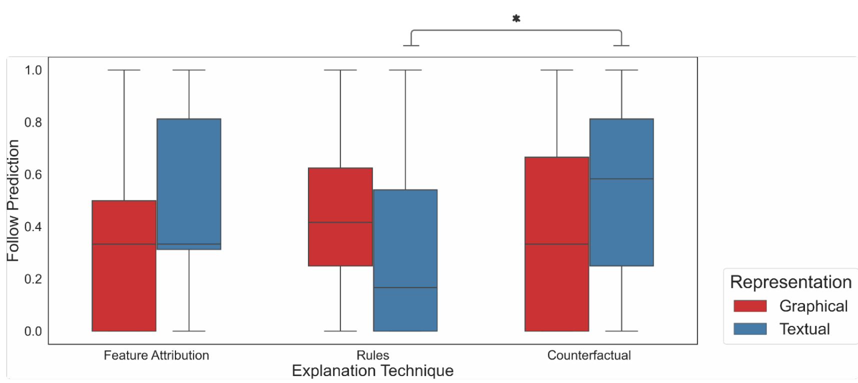


Fig. 7. Proportion of time the participants change their initial prediction to follow the AI's prediction. Results are shown for the Obesity dataset on the combination of explanation technique and representation.

representation significantly impacts the behavioural confidence **(Follow Prediction)** of initial disagreement participants.

Figure 6 shows that in the Obesity domain, participants exposed to a graphical representation report an increased confidence in their predictions after facing the explanation. Further examination reveals that in the Obesity domain, participants with higher educational attainment, who **initially disagreed**, experienced a decrease in confidence. Conversely, in the Recidivism domain, we observed that the confidence of female participants increased less compared to male participants when the AI **confirmed** their initial prediction.

Finally, Figure 7 showcases the average users' behavioural confidence for different explanation methods and representations in the Obesity dataset. We observe that for textual representations, users with counterfactual explanations are more prone to follow the AI system's prediction than participants with rule-based explanations. This suggests that users with rule-based explanations have lower confidence in the model's prediction.

5.3 Perception vs. Behaviour

We study the agreement between the self-reported and behavioural measurements defined in Section 4.2. We thus report the Pearson correlation between perceived confidence (resp. comprehension) and behavioural confidence (comprehension). We observe correlation scores of 0.43 and 0.49 between the perceived confidence when facing an explanation **(Δ Confidence)** and the proportion of users following the AI's prediction **(Follow Prediction)** for the COMPAS and Obesity datasets, respectively. This suggests a moderate positive correlation between these two measurements. In contrast, our results indicate no correlation between users' perceived understanding **(Immediate)** or **(Final)** and their actual comprehension of the model, as measured by the precision and recall scores.

6 DISCUSSION

In the discussion, we address key findings, draw design lessons for XAI practitioners, highlight limitations, and outline future perspectives.

6.1 Impact of Explanation Technique

We assessed the effects of three explanation techniques on participants' confidence and comprehension of two AI models **(RQ1)**. Our findings are in line with existing work and support our general hypothesis **(H1)**, namely that explanations increase both (a) the users' comprehension of the AI model and, (b) the confidence in the model's predictions. The study also confirms our **second hypothesis**, i.e., rule-based explanations are the most effective way to explain the inner workings of an AI system. This also stands in line with existing results [6, 64]. We surmise that this preference for rules is attributable to two factors: (a) its alignment with common educational reasoning principles, and (b) the simplicity of rules. This is supported by our results for both self-reported comprehension (Fig. 4) and precision (Fig. 5). We observe that the effects of explanations on AI-assisted tasks are more pronounced for the Obesity dataset than for COMPAS. We hypothesise that this is the result of (a) the number of features in the datasets (8 for COMPAS and 15 for Obesity), and (b) participants' prior knowledge of the field. Having more features to grasp makes explanations more beneficial. Also participants are unlikely to have firsthand experience with prisoners, but they are more likely to harbour preconceptions about the causes of obesity.

On the other hand, our study reveals a precision and self-reported comprehension comparable to the control group for counterfactual explanations. This outcome stands in stark contrast to the high scores observed for both recall (as illustrated in Figure 5) and behavioural confidence (as shown in Figure 7). This means that our participants tended to follow the AI model's prediction and could accurately identify the features mentioned in the explanation (good recall), but sometimes marked other features as important (low precision). This means that the counterfactual explanations may have been perceived as less complete than the others.

6.2 Impact of Representation

The influence of representation on users' perception has been well-established [16, 77], and our findings corroborate it **(RQ2)**. In particular, we found that the graphical representations induce a higher perceived confidence compared to textual representations (Figure 6). We suspect these results stem from a cognitive bias explained by the apparent complexity of a graphical presentation. This complexity may give the impression of a greater underlying effort, thereby increasing users' confidence in the system.

Our findings corroborate **H4** given that users' confidence for counterfactual explanation is higher with textual representations (Figure 7). Similarly, the post-hoc analysis on the interaction between explanation technique and representation on participants' recall (Table 1) suggests that textual representation appears to ease users' understanding of rule explanations. Our results, though, do not support **H3**, that is, users' confidence or comprehension for feature-attribution explanations is not significantly increased with graphical representations. These results do not intend to discourage the use of visual representations for such explanations. Rather, they underscore the need for improved representation techniques. This is vital to highlight since our experiment studied only one possible visual representation, *i.e.*, bars, which are widely used for feature-attribution explanations.

6.3 Recommendations for XAI Practitioners & Researchers

Our findings underscore the importance of user evaluation in the responsible deployment of XAI tools. We draw a set of recommendations for XAI practitioners and researchers conducting user studies within XAI.

We found that the mere **presence of explanations** has a positive impact on participants' self-reported and behavioural comprehension and confidence. This could be interpreted as support for consistently augmenting AI-based systems with explanations. However, we argue that this only holds when the explanations respond to a concrete user need, particularly in high-stakes domains such as healthcare and law. These needs may include legal requirements or educational purposes [8, 15]. Our experiments show that pre-conceptions and prior knowledge can elicit scepticism towards AI systems. This phenomenon has been also observed in prior work [51], where domain experts seem more prone to challenge AI-based recommendations than non-expert users. Critically, our results suggest that graphical explanations can induce automation complacency, resulting in confidence towards an AI explanation for the wrong reasons [7]. Prior work highlights that even domain experts display an excess of confidence in AI in the presence of explanation techniques such as feature attribution [34]. Consequently, we recommend that system designers inform users upfront about the extent and limitations of the system's explanations. This could mitigate the potential impact that preconceptions, cognitive biases, and the limitations of the AI system itself have on users' comprehension and confidence.

Regarding the **selection of an explanation paradigm**, our results suggest the use of rule-based explanations as a first proposal to describe an AI system's reasoning. Rule-based explanations provide a clear and concise summary of the necessary conditions for a given outcome. Nevertheless, rule-based explanations also pose some limitations. They respond to the question of what are *some* of the necessary conditions for the system to provide a given outcome and are, therefore, not a guarantee of functional causality (i.e., $A \Rightarrow \mathit{Obese}$ is not the same as $A \Leftrightarrow \mathit{Obese}$). This suggests that the choice of an explanation paradigm is better determined by the user's task. For example, 'what-if' tasks may suit counterfactual explanations better. Future work may investigate the effect of presenting users with a combination of multiple explanation paradigms.

Finally, we argue that system designers should bear in mind both the **system and explanation complexity**. We hypothesise that more input features in an AI agent may increase the perceived benefit of explanations. It has been also documented that comprehensibility decreases with explanation complexity as humans can handle at most 7±2 cognitive entities at once [18, 54]. Similarly, we argue for initially compact explanations that can be further detailed or extended upon user request. For example, a feature-attribution explanation could start by highlighting the top three most influential features, grouping the remaining features in a single bucket and allowing users to explore the full feature list if desired.

6.4 Limitations & Future Work

We identified several limitations related to the studied application domain and our participant sample. We resorted to crowdworkers as participants, given their increasing role in the training of and interaction with AI systems. While our participants faced stereotypical decision scenarios, our results may not be directly transferable to domain experts or computer scientists [22, 57, 65]. Indeed, contrary to a general audience, computer scientists may be familiar with particular explanation styles and representations, while domain experts may hold stronger pre-conceptions about their domain of expertise. Furthermore, we did not assess our participants' prior knowledge of the chosen domains, which could affect participants' performance. Future studies could, for example, evaluate the effect of explanation technique and representation on different user groups with different levels of expertise in a particular domain.

Prior research has employed questionnaires to assess how explanation techniques impact users' comprehension [78] and how different explanation representations can influence users' confidence [83]. However, the results from the analysis of our post-questionnaire on understanding yielded unexpectedly non-significant differences across various explanation techniques and representations. This outcome could be explained by the fact that users only engaged with the model a limited number of times and encountered instances that were classified differently. It is conceivable that this limited interaction might have contributed to the absence of statistical significance in our findings, as previously suggested by Van der Waa et al. [78]. To gain a more comprehensive perspective on the model's performance, a larger number of instances or instances with more similar classifications could be included in future evaluations.

Moreover, we observed in Section 5.3, no correlation between users' perceived understanding (**Immediate Understanding** or **Final Understanding**) and their actual comprehension of the model, as measured by the precision and recall scores. These findings are in line with existing research [17, 31, 70, 82]. Understanding why users elicit confidence without the corresponding behavioural alignment, or why they report comprehension without demonstrating it in practice remains an interesting open research question.

We evaluated participant comprehension through a simple task, namely, the identification of the most important features in the decision process – via the explanation. Other validation tasks could provide additional insights into participant understanding, *e.g.*, use the explanation to reproduce the AI's model behaviour on other examples, answer what-if scenarios, generating explanations [8, 40]. While such experiments could rely on our proposed framework (Figure 3), they are more complex and demand a fully-fledged new study. We do not expect our observations about the studied explanation paradigms to be completely portable to other tasks, *e.g.*, what-if scenarios. This remains an open research avenue.

The impact of graphical representation for rule-based and counterfactual explanations should be taken with caution, as it responds to an experimental requirement: the need to control for chart type. Bar charts, as used in our experiments, are widely employed for feature-attribution explanations on tabular data [61]. Therefore, the effectiveness of various chart styles for representing different explanation types deserves further investigation. This also raises the question of whether certain explanation paradigms are best suited to specific visual representations. Finally, and considering the insights from Hase and Bansal [31], we acknowledge that the impact of explanation techniques on comprehension may also vary with the data modality. In our study, the AI models were trained on tabular data. While the studied explanation techniques also apply to other data types such as text and images, the visual representations covered in this study may not suit those data types. Hence, further studies on other data modalities are necessary.

7 CONCLUSION

This study aims to fill the gap between the XAI and HCI communities by studying the impact of explanations and visual representations on users' comprehension and confidence. Our study covered three types of explanations; feature-attribution, rule-based, and counterfactual, each presented either graphically or as textual statements. We evaluated these in two domains: the prediction of recidivism and the risk of obesity. Our results indicate that rule-based explanations with textual representation are most effective for users' comprehension. Counterfactual explanations presented as text elicited higher levels of confidence, while the opposite was observed for feature-attribution and rule-based explanations. Importantly, our results are not entirely consistent across the evaluated domains. This accentuates the opportunities and demands for future studies on the effect of user profiles, data types, and domains on user's perceptions when interacting with AI systems.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E. Gilbert. 2021. Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI. *CoRR* abs/2106.07483 (2021). https://arxiv.org/abs/2106.07483

[3] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics Decis. Mak.* 20, 1 (2020), 310. https://doi.org/10.1186/s12911-020-01332-6

[4] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proc. AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088. https://doi.org/doi/10.5555/3306127.3331806

[5] Michał Araszkiewicz, Trevor Bench-Capon, Enrico Francesconi, Marc Lauritsen, and Antonino Rotolo. 2022. Thirty years of Artificial Intelligence and Law: overviews. *Artificial Intelligence and Law* (Aug 2022). https://doi.org/10.1007/s10506-022-09324-9

[6] Siddhant Arora, Danish Pruthi, Norman M. Sadeh, William W. Cohen, Zachary C. Lipton, and Graham Neubig. 2022. Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations. In *Proc. AAAI*. AAAI Press. https://doi.org/index.php/AAAI/article/view/20464

[7] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum. Comput. Interact. CSCW* 7 (2023), 1–17. https://doi.org/10.1145/3579460

[8] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29, 2 (01 Jun 2021), 149–169. https://doi.org/10.1007/s10506-020-09270-4

[9] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and Survey of Explanation Methods for Black Box Models. *CoRR* (2021). https://doi.org/abs/2102.13076

[10] S Boon and J Holmes. 1991. The Dynamics of Interpersonal Trust: Resolving Uncertainity in the Face of Risk. In *Cooperation and Prosocial Behaviour*. Cambridge University Press, Cambridge, 190–211.

[11] Lorella Bottino and Mario Cannataro. 2023. Explanation of machine learning models for predicting obesity level using Shapley values. In *Proc. in International Conference on Bioinformatics and Biomedicine, BIBM*. IEEE. https://doi.org/10.1109/BIBM58861.2023.10385994

[12] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the compas risk and needs assessment system. *Crim. Justice Behav.* 36, 1 (Jan. 2009), 21–40. https://doi.org/10.1177/0093854808326545

[13] Varun H Buch, Irfan Ahmed, and Mahiben Maruthappu. 2018. Artificial intelligence in medicine: current trends and future possibilities. *The British journal of general practice* 68, 668 (March 2018), 143–144. https://doi.org/10.3399/bjgp18X695213

[14] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science* 47, 9 (2009), 1260–1270. https://doi.org/science/article/pii/S0925753509000587 Research in Ergonomic Psychology in the Transportation Field in France.

[15] Blerta Abazi Chaushi, Besnik Selimi, Agron Chaushi, and Marika Apostolova. 2023. Explainable Artificial Intelligence in Education: A Comprehensive Review. In *Explainable Artificial Intelligence*. Springer Nature Switzerland.

[16] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proc. CHI, Glasgow, Scotland, UK*. ACM. https://doi.org/10.1145/3290605.3300789

[17] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proc. IUI*. ACM, 307–317. https://doi.org/10.1145/3397481.3450644

[18] Nelson Cowan. 2010. The magical mystery four: How is working memory capacity limited, and why? *Curr. Dir. Psychol. Sci.* 19, 1 (2010).

[19] Luca Deck, Astrid Schomäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. 2024. Mapping the Potential of Explainable Artificial Intelligence (XAI) for Fairness Along the AI Lifecycle. *CoRR* abs/2404.18736 (2024). https://doi.org/10.48550/ARXIV.2404.18736

[20] Julien Delaunay, Luis Galárraga, and Christine Largouët. 2020. Improving Anchor-based Explanations. In *Proc. CIKM*. ACM. https://doi.org/10.1145/3340531.3417461

[21] Julien Delaunay, Luis Galárraga, and Christine Largouët. 2022. When Should We Use Linear Explanations?. In *Proc. CIKM*. ACM. https://doi.org/10.1145/3511808.3557489

[22] Julien Delaunay, Luis Galárraga, Christine Largouët, and Niels van Berkel. 2023. Adaptation of AI Explanations to Users' Roles. In *Proc. ACM CHI - Workshop on Human-Centered Explainable AI*. https://vbn.aau.dk/en/publications/adaptation-of-ai-explanations-to-users-roles

[23] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2342–2351. https://doi.org/10.1145/3531146.3534647

[24] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. https://arxiv.org/abs/1702.08608

[25] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118. https://doi.org/10.1038/nature21056

[26] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (Apr 2022). https://doi.org/10.1007/s10618-022-00831-6

[27] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *CoRR* (2018). https://doi.org/abs/1805.10820

[28] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (2019), 93:1–93:42. https://doi.org/10.1145/3236009

[29] David Gunning. 2019. DARPA's explainable artificial intelligence (XAI) program. In *Proc. IUI*. ACM. https://doi.org/10.1145/3301275.3308446

[30] J F Hair, W C Black, B J Babin, and R E Anderson. 2014. *Multivariate data analysis (Seventh edition Pearson new international)*. Pearson Education Limited.

[31] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proc. ACL*. Association for Computational Linguistics. https://doi.org/10.18653/V1/2020.ACL-MAIN.491

[32] Guy Hoffman. 2019. Evaluating Fluency in Human-Robot Collaboration. *IEEE Trans. Hum. Mach. Syst.* 49, 3 (2019), 209–218. https://doi.org/10.1109/THMS.2019.2904558

[33] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR* (2018). http://arxiv.org/abs/1812.04608

[34] Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S Valley, Ella A Kazerooni, Nikola Banovic, Jenna Wiens, and Michael W Sjoding. 2023. Measuring the impact of AI in the diagnosis of hospitalized patients: A randomized clinical vignette survey study. *JAMA* 330, 23 (2023), 2275–2284.

[35] Alon Jacovi. 2023. Trends in Explainable AI (XAI) Literature. *CoRR* (2023). https://doi.org/10.48550/arXiv.2301.05433 arXiv:2301.05433

[36] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. 2023. Adversarial Counterfactual Visual Explanations. In *Proc. CVPR*. IEEE. https://doi.org/10.1109/CVPR52729.2023.01576

[37] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[38] Todd Kulesza, Simone Stumpf, Margaret M. Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE Computer Society. https://doi.org/10.1109/VLHCC.2013.6645235

[39] Francesca Lagioia, Riccardo Rovatti, and Giovanni Sartor. 2023. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. *AI Soc.* 38, 2 (2023). https://doi.org/10.1007/S00146-022-01441-Y

[40] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *CoRR* abs/2301.09656 (2023). https://doi.org/10.48550/arXiv.2301.09656

[41] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proc. SIGKDD*. ACM. https://doi.org/10.1145/2939672.2939874

[42] Retno Larasati, Anna De Liddo, and Enrico Motta. 2020. The Effect of Explanation Styles on User's Trust. In *Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with 25th International Conference on Intelligent User Interfaces (IUI 2020), Cagliari, Italy, March 17, 2020 (CEUR Workshop*

Proceedings, Vol. 2582). CEUR-WS.org. http://ceur-ws.org/Vol-2582/paper6.pdf

[43] Retno Larasati, Anna De Liddo, and Enrico Motta. 2020. The Effect of Explanation Styles on User's Trust. In *Proc. IUI (CEUR Workshop Proceedings)*. CEUR-WS.org. https://doi.org/Vol-2582/paper6.pdf

[44] Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. 2016. How We Analyzed the COMPAS Recidivism Algorithm — propublica.org. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[45] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2018. Comparison-Based Inverse Classification for Interpretability in Machine Learning. In *Proc. IPMU*. Springer. https://doi.org/10.1007/978-3-319-91473-2_9

[46] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. (2021). https://arxiv.org/abs/2110.10790

[47] Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 61, 10 (2018), 36–43. https://doi.org/10.1145/3233231

[48] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc. NIPS*. https://doi.org/doi/10.5555/3295222.3295230

[49] Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. 2021. Leveraging Latent Features for Local Explanations. In *Proc. KDD*. ACM. https://doi.org/10.1145/3447548.3467265

[50] Maria Madsen and Shirley D Gregor. 2000. Measuring Human-Computer Trust.

[51] Gonzalo Gabriel Méndez, Luis Galárraga, Katherine Chiluiza, and Patricio Mendoza. 2023. Impressions and Strategies of Academic Advisors When Using a Grade Prediction Tool During Term Planning. In *Proc. CHI (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 442, 18 pages. https://doi.org/10.1145/3544548.3581575

[52] Fabio Mendoza and Alexis de la hoz Manotas. 2019. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief* 25 (2019), 104344. https://doi.org/10.1016/j.dib.2019.104344

[53] Fabio Mendoza and Alexis de la hoz Manotas. 2019. Estimation of obesity levels based on eating habits and physical condition . UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5H31Z.

[54] George A Miller. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 2 (1956).

[55] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artif. Intell.* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[56] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3-4 (2021), 24:1–24:45. https://doi.org/10.1145/3387166

[57] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11 (2021), 24:1–24:45. https://doi.org/10.1145/3387166

[58] Christoph Molnar. 2018. Interprtable machine learning: A guide for making black box models explainable.

[59] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proc. Conference on Fairness, Accountability, and Transparency*. 607–617.

[60] Jeroen Ooge and Katrien Verbert. 2022. Explaining Artificial Intelligence with Tailored Interactive Visualisations. In *Proc. IUI*. ACM. https://doi.org/10.1145/3490100.3516481

[61] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proc. CHI*. ACM. https://doi.org/10.1145/3411764.3445315

[62] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proc. AIES*. ACM. https://doi.org/10.1145/3375627.3375850

[63] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. SIGKDD*. ACM. https://doi.org/10.1145/2939672.2939778

[64] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proc. AAAI*. AAAI Press. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982

[65] Mireia Ribera and Àgata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *Proc. IUI (CEUR Workshop Proceedings)*. CEUR-WS.org. https://doi.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

[66] Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating Realistic Natural Language Counterfactuals. In *Findings EMNLP*. ACL. https://doi.org/10.18653/v1/2021.findings-emnlp.306

[67] Yao Rong, Tobias Leemann, Thai trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards Human-centered Explainable AI: A Survey of User Studies for Model Explanations. arXiv:2210.11584 [cs.AI]

[68] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (01 May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[69] Rosemarie S. Punla, Candida C. Farro. 2022. Are we there yet?: An analysis of the competencies of BEED graduates of BPSU-DC. *International Multidisciplinary Research Journal* 4, 3 (Sept. 2022), 50–59.

[70] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proc. IUI*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[71] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proc. ICML*. PMLR. https://doi.org/v70/shrikumar17a.html

[72] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. 2022. Do Users Benefit From Interpretable Vision? A User Study, Baseline, And Dataset. In *Proc. ICLR*. OpenReview.net. https://openreview.net/forum?id=v6s3HVjPerv

[73] Eduardo A. Soares and Plamen Angelov. 2019. Fair-by-design explainable models for prediction of recidivism. *CoRR* abs/1910.02043 (2019). http://arxiv.org/abs/1910.02043

[74] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. *CoRR* abs/1703.01365 (2017).

[75] Harini Suresh, Kathleen M Lewis, John Guttag, and Arvind Satyanarayan. 2022. Intuitively Assessing ML Model Reliability through Example-Based Explanations and Editing Model Inputs. In *Proc. IUI*. Association for Computing Machinery, 767–781. https://doi.org/10.1145/3490099.3511160

[76] Andrea Tagarelli and Andrea Simeri. 2022. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artificial Intelligence and Law* 30, 3 (01 Sep 2022), 417–473. https://doi.org/10.1007/s10506-021-09301-8

[77] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proc. CHI*. ACM. https://doi.org/10.1145/3411764.3445365

[78] Jasper van der Waa, Elisabeth Nieuwburg, Anita H. M. Cremers, and Mark A. Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artif. Intell.* 291 (2021), 103404. https://doi.org/10.1016/j.artint.2020.103404

[79] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum. Comput. Interact.* 5, CSCW2 (2021), 1–39. https://doi.org/10.1145/3476068

[80] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* (2017). https://doi.org/abs/1711.00399

[81] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proc. IUI*. Association for Computing Machinery. https://doi.org/10.1145/3397481.3450650

[82] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4 (2022), 27:1–27:36. https://doi.org/10.1145/3519266

[83] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proc. International Conference on Intelligent Virtual Agents, IVA*. ACM. https://doi.org/10.1145/3308532.3329441

[84] Daniel S. Weld and Gagan Bansal. 2018. Intelligible Artificial Intelligence. *CoRR* (2018). https://doi.org/abs/1803.04263

[85] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda B. Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Trans. Vis. Comput. Graph.* 26, 1 (2020), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

[86] Linyi Yang, Eoin M. Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In *Proc. COLING*. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.541

[87] Jun Yuan and Enrico Bertini. 2022. Context sight: model understanding and debugging via interpretable context. In *Proc. of the Workshop on Human-In-the-Loop Data Analytics*. ACM. https://doi.org/10.1145/3546930.3547502

SUPPLEMENTARY MATERIAL

This appendix consists of five sections aimed at providing a comprehensive overview of various aspects related to our experimental evaluation. In Appendix A, we delve into the details of the code, classifier, and datasets utilised in our experimental evaluation. Moving forward, Appendix B presents a comprehensive table detailing the demographic information of our participants. Subsequently, in Appendix C, we provide an overview of the diverse set of questions and surveys used throughout the entire experimental process. To shed light on our approach to representing explanations and communicating them to participants, we offer insights in Appendix D. Following that, we justify in Appendix D some choice we made to represent explanations and how they are described to the participants. Finally, in Appendix E, we illustrate the practical application of our various scales and metrics using a specific participant as an example.

A CODE AND DATA PROCESSING

This section provides useful information to reproduce the presented experimental results. The source code is available in an anonymous repository on GitHub [7].

Compas: In order to generate explanations meaningful to the users, we removed some features and kept this subset of features {Gender, Age, Race, Juvenile felony count, Juvenile misdemeanour count, Priors count, Charge degree, Charge description}. We also removed 508 individuals having a charge description that occurred less than 5 times in the whole dataset. The dataset can be downloaded online[8].

Obesity: This dataset is originally composed of 16 features and a target obtained from questions detailed in [52]. However, we removed the weight since it would be too easy for the model and the user to predict the BMI with both the height and weight. We binaries five features: Gender, family history with overweight, does the user smokes, calorie consumption monitoring, and does the user frequently consumes high-caloric food. The other features were one hot encoded, the original data can be downloaded on this link [53][9].

Table 3 contains the final number of features and instances for both datasets as used in our experiments.

Dataset	Features		Instances
	Numerical	Categorical	
Compas	1	7	5364
Obesity	2	13	2111

Table 3. Description of the datasets.

[7] https://anonymous.4open.science/r/user_eval-1776/README.md
[8] https://github.com/propublica/compas-analysis/
[9] https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

B DEMOGRAPHIC INFORMATION

Table 4 outlines the demographic details of our participants, categorised by domain (Obesity or Recidivism). It is noteworthy that the consent for information from 11 participants in the Obesity group has been revoked.

Domain	**Obesity**		**Recidivism**	
Factor	***N***	% sample	***N***	% sample
Gender				
Female	66	47.14	66	47.14
Male	62	44.29	74	52.86
Prefer not to say	1	0.71	0	0.0
Consent revoked	11	7.86	0	0.0
Age				
< 20	10	7.14	11	7.86
20 < 30	81	57.86	88	62.86
30 < 40	24	17.14	27	19.29
40 >	14	10.0	14	10.0
Nationality				
Africa	45	32.14	37	26.43
Asia	2	1.43	2	1.43
Australia	0	0.0	1	0.71
Europe	77	55.0	82	58.57
North America	5	3.57	15	10.71
South America	0	0.0	3	2.14
Ethnicity (simplified)				
Asian	2	1.43	2	1.43
Black	37	26.43	30	21.43
Mixed	10	7.14	9	6.43
Other	3	2.14	8	5.71
White	77	55.0	91	65.0
Highest education				
Doctorate degree	3	2.14	1	0.71
Graduate degree	27	19.29	24	17.14
High school diploma	47	33.57	37	26.43
Technical college	3	2.14	14	10.0
Undergraduate degree	49	35.0	64	45.71

Table 4. Overview of participants' demographic factors.

C QUESTIONNAIRE

In our survey, we ask the online users to complete two various questionnaires, each one evaluating a given criteria. We present in this section the question and where each questionnaire comes from.

C.1 Understanding Scale

We now present the questions to evaluate the users' perceived understanding of the system from Madsen and Gregor [50]. This questionnaire is composed of 8 questions:

(1) The system uses appropriate methods to reach decisions.
(2) The system has sound knowledge about this type of problem built into it.
(3) The advice the system produces is as good as that which a highly competent person could produce.
(4) The system makes use of all the knowledge and information available to it to produce its solution to the problem.
(5) I know what will happen the next time I use the system because I understand how it behaves.
(6) I understand how the system will assist me with decisions I have to make.
(7) Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
(8) It is easy to follow what the system does.

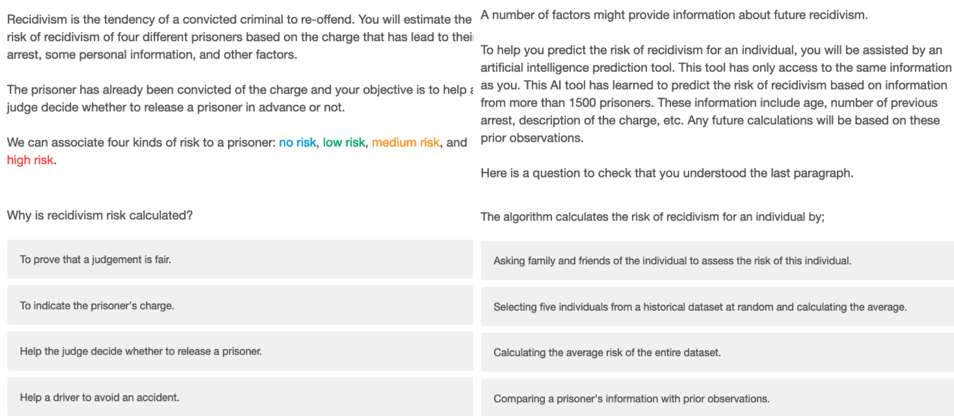
For each of these questions, Madsen and Gregor [50] recommended this 5 Likert scale:

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

C.2 Question to verify user's validity

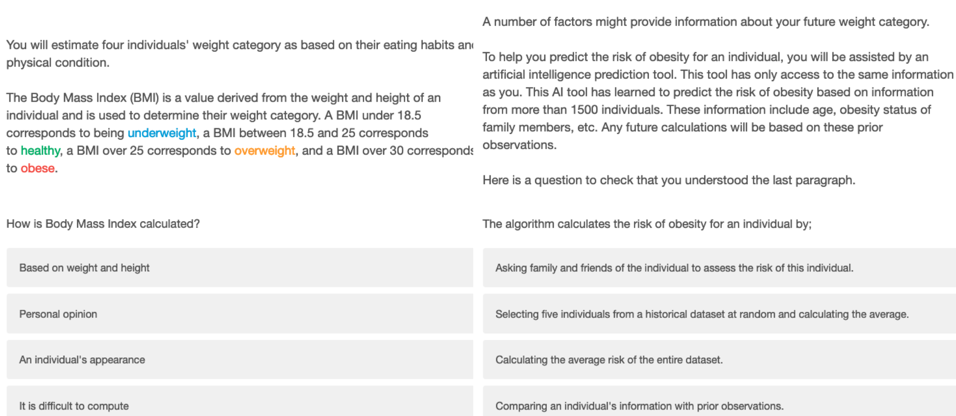
We ask the user two questions in order to verify that they understand and will try efficiently to complete the questionnaire.

Following the task introduction, we assessed whether the participants had actually read and understood the task through two questions: *'How is Body Mass Index calculated?'* for the Obesity dataset and *'Why is recidivism risk calculated?'* for COMPAS. We found 10 and 30 incorrect answers for the first and second questions, respectively. This question had the form *'The algorithm calculates the risk of obesity* (resp. recidivism) *for an individual by;'*. We asked additional users to participate in our study until we had 20 responses for each group that validated our two understanding questions resulting in a final set of 280 participants.



(a) (b)

Fig. 8. Detailed presentation of the two verifying questions at the end of the Compas dataset survey.



(a) (b)

Fig. 9. Detailed presentation of the two verifying questions at the end of the obesity dataset survey.

D EXPLANATION TECHNIQUES AND REPRESENTATIONS

In this section, we first elaborate on the representation of each explanation technique and then the manner in which these explanations were conveyed to the participants.

D.1 Explanation Techniques

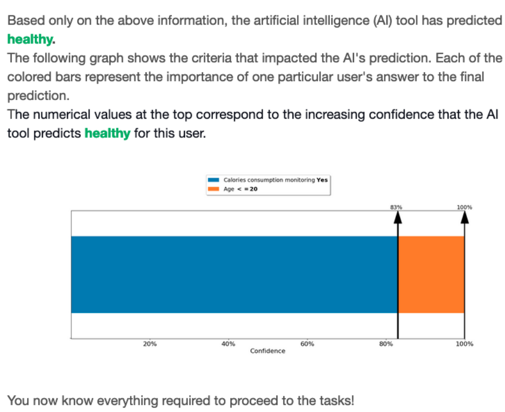
For the graphical representation of **feature-attribution explanations**, we made specific choices to enhance clarity and manage complexity. Unlike standard methods that focus on a limited number of features, we sorted features in decreasing order based on the absolute value of their attribution. Features with attributions less than half the absolute value of the preceding feature were considered marginal and grouped together. For example, in Appendix D.2, features impacting less than 2% are grouped into the last bar, and their cumulative attribution score equals 1% toward the obesity class.

In the representation of **rule-based explanations**, we utilised stacked bars, starting with the rule's condition that induced the highest initial confidence in the model's prediction. Subsequently, we iteratively added conditions that improved the most the model's confidence, given that existing conditions were validated. Additionally, we omitted the background colour representing ordinal classes due to the nature of rule-based explanations. Decision rules signify the minimum requirement for the model's prediction toward one class, offering no information on the model's behaviour on other classes.

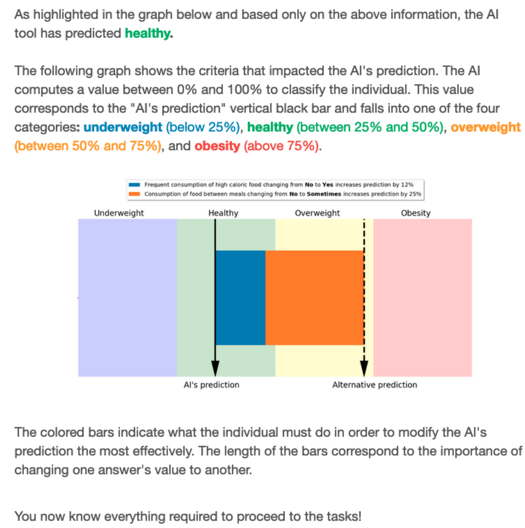
Consistency in representation was maintained for **counterfactual explanations**, employing stacked bars. The length of each bar indicates the extent to which changing a feature's value is necessary to shift the model's answer from one predicted class to another (the counterfactual class). We begin by displaying the feature that most impacts the prediction, then, with this feature changed, we identify the second most impactful feature, continuing until the prediction shifts between classes.

D.2 Explanation Paragraph in Example Round

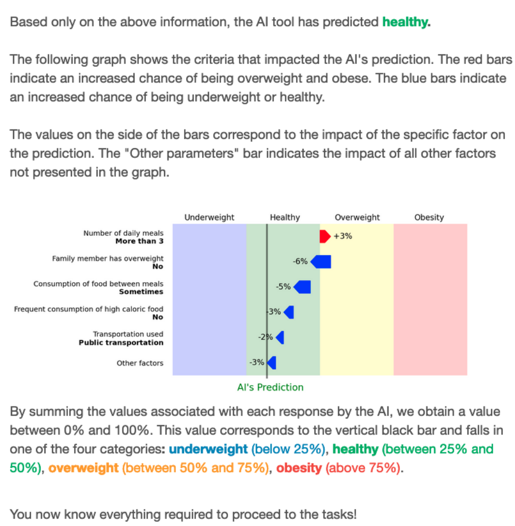
During the introduction step, specifically when participants were exposed to an explanation for the first time, a detailed description of the visual representations was provided. This paragraph underwent a thorough review by 20 individuals, including 9 computer scientists and 11 laypeople, to ensure comprehensiveness and effectiveness in conveying the explanation. The resulting explanation paragraphs are detailed below.



Rule-based.



Counterfactual.



Linear.

Fig. 11. Detailed presentation of the three graphs presentation in the introduction and more precisely the first time the participant had access to an explanation in the survey.

E SCALES & METRICS (ILLUSTRATION FOR ONE PARTICIPANT)

In this section, we provide a detailed example of how we employed the scales and metrics introduced in Section 4.2 for one participant from the rule-based explanation group. This example is designed to provide the reader with a detailed explanation of how we assessed various facets of participant' behaviour and perception. We recall that Figure 3 shows the times at which these parameters are measured. For this illustration, let us refer to this participant as "User J." User J participated in predicting the risk of obesity in response to four distinct scenarios, and their responses are reported in Figure 12.

	1st User's **Prediction**	1st User's **Confidence**		AI's Prediction	Top Features According to the Rule-based Explanation	Top Features According to the User	2nd User's **Prediction**	2nd User's **Confidence**	Perceived **Understanding**
Q1: What is the risk of obesity? (Scénario 1)	No Risk	2/5		Low Risk	• Monitoring Calory • Consumption of High-Caloric Food	• Monitoring Calory • Age • Gender	Low Risk	3/5	3/5
Q2: What is the risk of obesity? (Scénario 2)	Low Risk	3/5		Medium Risk	• Family Member has Overweight • Physical Activity Frequency	• Family Member has Overweight • Physical Activity Frequency	Medium Risk	4/5	4/5
Q3: What is the risk of obesity? (Scénario 3)	Medium Risk	1/5		No Risk	• Monitoring Calory • Physical Activity Frequency • Age	• Monitoring Calory • Age	Low Risk	3/5	5/5
Q4: What is the risk of obesity? (Scénario 4)	High Risk	4/5		High Risk	• Consumption of High-Caloric Food • Family Member has Overweight • Transportation Used	• Physical Activity Frequency • Consumption of High-Caloric Food • Smoke	High Risk	3/5	1/5

Fig. 12. Example of answers from participant "User J" from the rule-based explanation group. The values within the columns "1st User's Confidence", "2nd User's Confidence", and "Perceived Understanding" are on a 5-Likert scale.

E.1 User's Initial Prediction and Confidence

In Figure 12, User J's initial predictions, scaled from 1 (no risk) to 4 (high risk), are accompanied by their initial confidence levels, measured on a 5-point Likert scale. The Likert scale spans from "strongly disagree" to "strongly agree." User J's initial predictions are shown in the "1st User's Prediction" column, and their initial confidence is recorded in the "1st User's Confidence" column.

E.2 AI Model Predictions and Explanations

User J's predictions are followed by the AI model's predictions and associated explanations, presented as depicted in Figure 2. These explanations comprise lists of the most influential features considered by the AI model for each prediction scenario. For example, in Figure 2, the most important features for the feature attribution are *Family member has overweight*, *Consumption of food between meals*, *Consumption of high caloric food*, *Transportation used*, and *Calories consumption monitoring*. In contrast, for counterfactual, this is only the *Family member has overweight* and *Physical activity frequency* while rule-based also includes the *Age* feature.

E.3 User's Final Prediction and Confidence

During the task round, User J was asked to select, from the list of features, which features they considered most important for the AI model's prediction. Subsequently, User J was given the opportunity to reevaluate their prediction in the "2nd User's Prediction" column and provide their final confidence in their prediction in the "2nd User's Confidence" column.

E.4 User's Perceived Understanding

User J was also asked to rate their "Perceived Understanding" on a 5-point Likert scale to indicate their understanding of how the model made the prediction.

E.5 Metrics Calculation

The metrics for User J's responses were calculated as follows:

- **Δ-Confidence:** The Δ-Confidence was computed by subtracting the initial confidence from the final confidence for each scenario. User J's Δ-Confidence values are 1, 1, 2, and -1 for the four scenarios. The average Δ-Confidence for User J is thus 3/4.
- **Behavioral Trust (Follow Pred.):** We assessed behavioral trust by tracking instances where the user modified their initial prediction to match the AI model's prediction. It is important to note that we only considered scenarios where the user's initial prediction differed from the AI model's prediction. Thus, User J modified their initial prediction to align with the AI model's prediction in 2 out of 3 such scenarios, resulting in a behavioral trust score of 2/3.
- **Immediate Understanding:** User J's immediate understanding is the average value of their Likert-scale ratings for understanding across all four scenarios. In this case, it is (3 + 4 + 5 + 1) / 4, which equals 13/4.
- **Behavioral Understanding (Precision and Recall.):** To measure User J's precision and recall, we compared the list of features they identified as important to those highlighted in the explanation for each scenario. The precision and recall values for each scenario were calculated as follows:

Scenario Q1:
- Precision = 1/3 (User identified three features, one matched AI explanation),
- Recall = 1/2.

Scenario Q2:
- Precision = 1 (User and AI explanation lists are identical),
- Recall = 1.

Scenario Q3:
- Precision = 1 (User identified 2 features, both matched AI explanation),
- Recall = 2/3.

Scenario Q4:
- Precision = 1/3 (User identified 1 feature, which matched AI explanation),
- Recall = 1/3.

Please note that these are simplified examples, and in practice, the lists of important features in explanations are typically longer.

Confidence	The system uses appropriate methods to reach decisions	The system has sound knowledge about this type of problem built into it.	...	I understand how the system will assist me with decisions I have to make.	It is easy to follow what the system does.		Average
User J's Answers	3/5	4/5	...	3/5	4/5		3.5/5

Fig. 13. Example of answers from one participant to the Understanding survey. We measure the users' perceived comprehension of the AI system on a scale from 1 to 5.

E.6 Post-Questionnaires

In Figure 13, we present an example of a survey measuring User J's perceived comprehension of the AI system. This survey was adapted from Madsen and Gregor [50] and employed a Likert scale ranging from 1 to 5. The average of User J's responses to the eight survey questions provides a representation of their perceived understanding, which, in this case, is 3.5 out of 5.