Impact of Explanation Technique and Representation on Users' Comprehension and Confidence in Explainable AI

Julien Delaunay, Luis Galárraga, Christine Largouët, Niels Van Berkel

Impact of Explanation Technique and Representation on Users' Comprehension and Confidence in Explainable AI

JULIEN DELAUNAY, Top Health Tech *, Spain
LUIS GALÁRRAGA, Inria/IRISA, France
CHRISTINE LARGOUËT, Institut Agro/IRISA, France
NIELS VAN BERKEL, Aalborg University, Denmark

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. In this paper, we 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

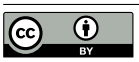
ACM Reference Format:
Julien Delaunay, Luis Galárraga, Christine Largouët, and Niels van Berkel. 2025. Impact of Explanation Technique and Representation on Users' Comprehension and Confidence in Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW113 (April 2025), 28 pages. https://doi.org/10.1145/3711011

1 Introduction

Artificial Intelligence (AI) algorithms have become ubiquitous for decision-making, including in high-stakes domains such as law [5, 71] and healthcare [12, 24]. This has raised numerous critical questions and concerns. One of these concerns arises from the fact that current AI algorithms can be incredibly complex, making algorithmic decision-making opaque—i.e., the algorithms behave like black boxes [63]. 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 comprehension [28, 58] and trust [39, 54] in AI systems, thereby facilitating their adoption.

Over the last five years, the XAI community has focused primarily on developing methods to compute local explanations for AI models. These approaches explain the reasoning of an AI system when applied to a single case, i.e., a **target instance**, and can be categorised into three broad 'explanation families': feature-attribution, rule-based, and counterfactual [8, 25, 27, 33]. There are a large number of explanatory methods, some of which have been widely adopted by data

Authors' Contact Information: Julien Delaunay, jdelaunay@topdoctors.es, Top Health Tech , Barcelona, Spain; Luis Galárraga, luis.galarraga@inria.fr, Inria/IRISA, Rennes, France; Christine Largouët, christine.largouet@irisa.fr, Institut Agro/IRISA, Rennes, France; Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Aalborg, Denmark.



ACM 2573-0142/2025/4-ARTCSCW113
https://doi.org/10.1145/3711011

practitioners [26, 43, 58, 59]. Despite this plethora of XAI methods, much work has pointed to a lack of end-user involvement in the evaluation of such methods [1, 4, 23, 60]. For example, 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 techniques are introduced without a clear understanding of how the intended end-users perceive or interpret these explanations.

In contrast to the XAI and ML communities, user studies on AI explanations are commonplace within the wider Human-Computer Interaction (HCI) and CSCW communities [14, 39, 72]. This line of work underscores the importance of evaluating the impact of explanations on comprehension (i.e., do users understand the AI system better thanks to the explanation?) and confidence (i.e., to which extent explanations increase or decrease users' confidence in AI recommendations?). However, existing studies typically focus on specific use cases, for example in a particular domain, with a single explanation technique, or with a small and very specific cohort (e.g., CS students). Furthermore, these studies tend to rely on human-generated explanations rather than explanations generated by real-world AI systems. This creates a barrier to extrapolating these results to other XAI scenarios, and is also unable to provide comparative evidence on the suitability of different explanation techniques 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 information perception [14, 72], our investigation also includes a comparison of the effect of the visual representation of the explanation on user comprehension and confidence.

Our investigation consists of a user study involving 280 crowd-workers, who were given an AI-assisted prediction task across two use cases: prediction of the risk of obesity and recidivism. The AI agents operate on tabular data and provide explanations for their predictions. We compute these explanations using established explanation techniques, i.e., LIME [58], Anchors [59], and Growing Fields [20, 41]. 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' comprehension and confidence;
(2) A methodological framework for user studies designed to measure the impact of AI explanations on users' comprehension and confidence.

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

2 Related Work

Our work lies at the intersection of XAI, 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.

2.1 XAI Techniques for Local Explanations

An AI model is an agent f that takes an instance x as input and returns an output $f(x)$. The instance x consists of features, e.g., attributes of a person for tabular data, image pixels, or words in a text. The output $f(x)$ can be a class, e.g., low risk vs. high risk, or a number, e.g., a price estimate. An explanation is an expression that describes the relationships between the input and output of an AI model f [42]. Explanations can be computed via a post-hoc explanation module or, in the

case of white box algorithms, extracted directly from f. When the explanation focuses on a single instance, it is called a *local explanation*. Local explanations have recently received more attention from machine learning (ML) researchers [33]. Based on prominent XAI surveys [8, 27], we can categorise these explanations into three main types:

Feature-attribution explanations. These explanations provide the contribution of the input features to the output of a black box on a target instance. Here, the magnitude of a feature's contribution informs us of its importance for a particular prediction outcome, while the sign denotes a positive or negative correlation with that outcome. As well as classic white-box methods such as linear regression, there are a number of methods that can compute such scores from black-box models in a post-hoc fashion. Some of these work for specific models, such as neural networks [65, 69], while others, such as LIME [58] and SHAP [43], are model-agnostic. This has made them popular among researchers and practitioners. We use LIME in our study, but SHAP could have been a viable alternative.

Rule-based explanations. Approaches such as Anchors [59] and LORE [26] compute explanations in the form of decision rules on the input features. Anchors is model- and data- agnostic and relies on bandit exploration to compute a single general and accurate decision rule that mimics the behaviour of a black box on the target instance [59], while LORE operates on tabular data and learns a decision tree trained on artificial instances that resemble the target instance [26]. Explanatory rules can, therefore, be extracted from this decision tree. We chose Anchors for our experiments because it provides a single explanation rule without additional computation.

Counterfactual explanations. These explanations convey the minimum adjustments required in the target instance to alter the model's prediction. They, therefore, identify the most *sensitive* features within the agent's decision process. Counterfactuals are similar to adversarial examples in that 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 [34]. Counterfactual explanations, on the other hand, do not have this constraint because they aim to be actionable and understandable. Methods such as Growing Spheres [41], FACE [56] or DICE [52] perturb the target instance, i.e., they create new instances by increasingly changing various attributes in the target instance until they identify an instance that changes the model's prediction. Our experiments use the Growing Fields algorithm [20], an extension of the Growing Spheres algorithm [41] that supports both continuous and categorical attributes. We chose this algorithm because of its simplicity. Contrary to other approaches [52, 56], it does not impose additional constraints on the counterfactuals (e.g., diversity, likelihood), the evaluation of which is beyond the scope of our study.

2.2 Evaluating Explainable AI Systems

Explainability is an inherently human-centric property. Consequently, Miller argues that the development of effective explanations requires joint efforts of the XAI and HCI research communities [49]. While the HCI community has emphasised the need for human-centred evaluations of XAI systems [23], several surveys have highlighted the scarcity of XAI papers that evaluate novel explanation techniques through user studies [1, 4, 23]. Among these user studies, most evaluated either the validity of their novel explanation technique [38, 44, 58, 59, 61, 81] or the impact of the visual representation of the explanation [14, 53, 55, 80]. A limitation of these works is that they are typically limited to the evaluation of one type of explanation technique [38, 53, 61] and one application domain [55, 81]. Some prominent explanation techniques, such as LIME [58] and Anchors, have evaluated the quality of explanations with a small number of computer science students already familiar with machine learning [59]. In our work, we set out to compare three different explanation techniques on two distinct datasets with lay users.

To study the impact of explanations, prior work has mostly evaluated users' trust and understanding in highly specific settings [14, 35, 39, 66, 74]. For instance, Arora et al. [66] studied the impact of interactive explanations on user comprehension. The results of this study confirmed that explanations help users identify key elements for the prediction. Cheng et al. [14] 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 [14, 72]. Van Berkel et al. compared textual and scatterplot representations and showed that the usage of a scatterplot visualisation led to lower perceived fairness [72]. Other works have compared the effects of different explanation techniques on users [35, 66, 74]. For instance, Van der Waa et al. compare hand-crafted example-based and rule-based explanations for the self-management of diabetes [74]. De Jong et al. explored the use of partial explanations to reduce user's overreliance on explanations [17]. Their results show that partial explanations can reduce overreliance on AI suggestions as compared to 'full' explanations.

In this study, we provide a comprehensive evaluation that compares three established explanatory techniques, generated by LIME, Anchors, and Growing Fields. We compare these explanations across two visual representations, namely graphical and textual. Following the recent guidelines for evaluating XAI applications [74], we experiment with a large cohort, on two different datasets, and collect both perceptual and behavioural metrics of user understanding and confidence.

3 Explanation Techniques and Representations

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

3.1 Datasets & AI models

Datasets. Our evaluation is conducted on two datasets widely used by the XAI community [2, 10, 18, 36, 68, 82], namely COMPAS [11] and Obesity [47]. 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 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 an 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 and healthcare [3, 75]. We chose to include more than one domain following the recommendation that a meaningful application-agnostic XAI evaluation should include more than one domain [72, 74], and strike a balance between simplicity—participants should grasp the domain—and plausibility—the task should be sufficiently challenging 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. We chose this model because it is a true black-box model with a strong prediction power. Its decision boundary is too complex to be easily understood by examining the model parameters. We note that other powerful black-box models, such as random forests or gradient-boosting trees, would also have been suitable for this task. We trained the MLPs on 70% of the instances and

[1]We removed weight as a variable from this 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

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 cases presented to participants from the Obesity (left) and COMPAS (right) datasets.

evaluated them on the remaining 30%. We obtained accuracies of 67% and 78% for the COMPAS and Obesity datasets, respectively. Although these accuracy levels may appear low, they are in line with those observed in the literature [40, 77]. We did not reveal these accuracy scores in the experiments to avoid any bias on the participants' confidence in the model. For 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 for consistency reasons: 'underweight', 'healthy', 'overweight', and 'obese'. We then generated three different explanations for each instance in the test set: a feature-attribution explanation based on LIME [58], a rule-based explanation based on Anchors [59], and a counterfactual explanation using Growing Fields [20]. The methods were used with the default parameters except that (a) Anchors used the discretisation proposed by Delaunay et al. [19], and (b) we computed the attribution of all features in the LIME explanation—contrary to the default configuration that only picks the top six features.

For each dataset, we selected five individuals from the test set to be presented to the participants—one for each of the four predicted classes plus an additional individual used as an example. Figure 1 depicts the information of an individual (one per dataset) as shown to the participants. The grey column contains the various features while the corresponding defendant or patient data are shown 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 techniques do not provide the same exact insights into the AI's prediction process, the explanations are usually conveyed using different representations, which furthermore depend on the type of data (e.g., image, tabular, text, etc). For tabular data, existing XAI toolkits[4] opt for a graphical representation based on bars for feature attribution explanations. Conversely, the most common representation for rule-based and counterfactual explanations is natural language. To control for this visual representation in our experiments, participants are presented with common graphical and textual representations for all three explanation techniques, as illustrated in Figure 2.

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

[3] https://github.com/j2launay/user_eval

[4] AI360, Dalex, H2O, eli5, InterpretML, What-if-Tool, Alibi, Captum.

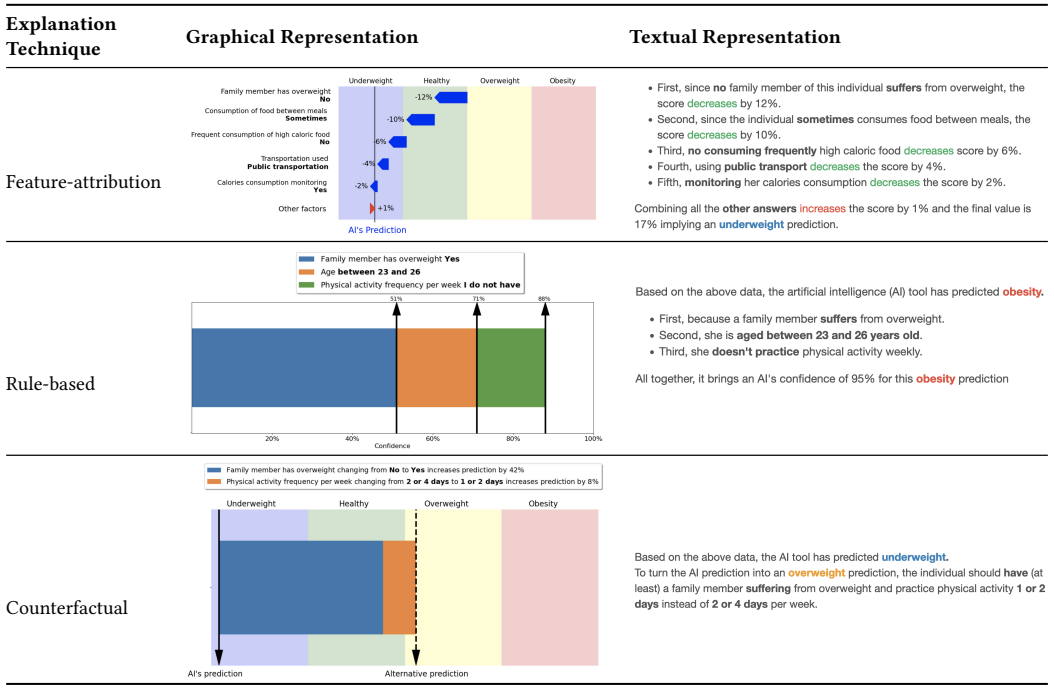


Fig. 2. Different explanations for a random individual in the Obesity dataset.

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 [58] 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 describes towards which side the feature biases the AI model's prediction (underweight vs. obese, low risk vs. high risk). To limit explanation complexity, our representation groups features with a marginal attribution under 'Other features'. Here, the aggregated attribution is the sum of the attribution scores of those features (for more details read Appendix D.1).
- Rule-based explanations are depicted using stacked bars, as per Molnar [51], 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; adding condition (b) increases 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 states 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 techniques we present the explanation as a bulleted list of the relevant attributes. The list is a manual transcription of the contents of the graphical representations, starting from the most impactful feature. This transcription was reviewed and validated by all authors. Each item from the list describes the effect of one feature on the model's answer. This effect can be how much the feature contributes to the model's prediction (feature-attribution), how much it boosts the confidence of the prediction (rule-based), or how sensitive the AI model is to changes in the input features (counterfactual). 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 and colour.

4 Method

While the XAI community has proposed multiple post-hoc explanation techniques based on feature attribution, rules, and counterfactual instances, no prior work has evaluated the impact on users' comprehension and confidence for all these techniques. This motivates our first research question **RQ1: "How do local explanation techniques, i.e. feature-attribution, rule-based, or counterfactuals, affect users' comprehension and confidence of an AI model?"** Existing works have shown that explanations improve users' ability to comprehend a model [58, 66]. Hence, this question underlies our first general hypothesis; (H1) **explanations improve participants' comprehension of and confidence in a model**. In addition, we observe that decision rules have consistently outperformed other techniques in helping users understand the inner mechanisms of a model [59, 66]. 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 [55, 74]. We, therefore, follow a more exploratory approach to study the impact of explanation techniques on user confidence and do not hypothesise on this aspect.

As highlighted in prior work [14, 72], the visual representation of an explanation impacts the users' perception. This leads to our second research question, **RQ2: "How does an explanation's visual representation impact the users' comprehension and confidence?"** As it is common 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' comprehension and confidence** (H3), whereas **a textual representation leads to higher comprehension and confidence for rule-based and counterfactual explanations** (H4).

Our study seeks to clarify the relationships between users' comprehension and confidence in an AI model (dependent variables), based on (i) the explanation technique—feature-attribution, rule, or counterfactual—and (ii) the visual representation—graphical or textual (independent variables).

4.1 Task

Our two user studies (Obesity and Recidivism) follow a between-subject design, in which each participant interacts with one explanation technique and one representation 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 these predictions, participants rely on an AI recommendation, as described in Section 3.1, complemented with an explanation. We created individual surveys for each dataset, explanation technique, and explanation representation. For each dataset, we also defined a *control group* for which participants did not receive any explanation. Figure 3 outlines the process of these surveys. Each survey is composed of three phases:

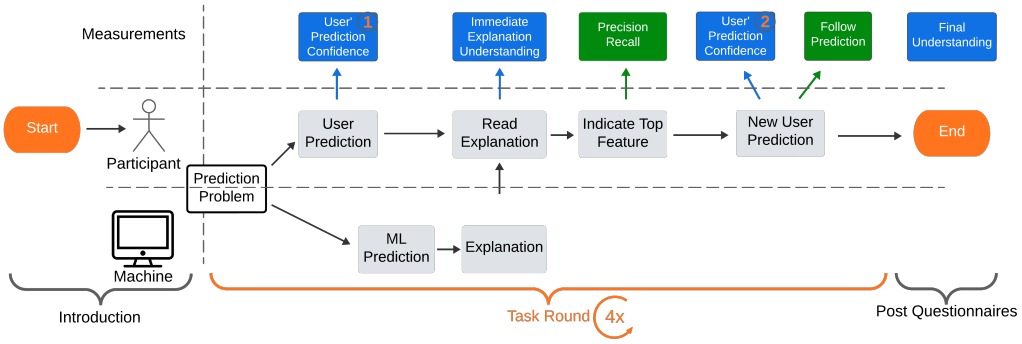


Fig. 3. Experimental workflow used to assess participant perception and behaviour when interacting with a given explanation technique. Behavioural measurements are in green, while self-reported measurements are in blue. The task round is repeated for four different prediction problems.

Introduction. A description of the tasks assigned to the participant and the information used by the AI model to make recommendations (cf. Figure 1). We asked participants two multiple-choice questions to verify whether they understood the task, namely 'How is Body Mass Index calculated?' and 'Why is recidivism risk calculated?'.
Task Round. 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 on a 5-point Likert item. Following this assessment, the participants have access to the AI model's prediction along with an explanation (cf. Figure 2)—excluding participants on the baseline condition. 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. Finally, participants can reconsider their initial prediction and answer two questions to report their understanding of the explanation ('immediate explanation understanding', see Figure 3) and their confidence in their prediction ('participant prediction confidence') on a 5-point Likert item.
Post-Questionnaire. After the prediction tasks, the participants answer an 8-question questionnaire to report their understanding of the AI model, as detailed in Section 4.2.

4.2 Scales & Metrics

To assess the impact of our independent variables—explanation technique and representation—, we employed various metrics to evaluate participants' comprehension and confidence. These elements are frequently identified as crucial measurements in human-centred XAI [31, 50, 62]. Several user studies have shown that perceived comprehension and actual comprehension may differ [14, 15, 30]. Therefore, we distinguish between self-reported and behavioural metrics. Figure 3 shows when these parameters are measured (a detailed example of the measurement process is provided in Appendix E).

Comprehension. A widely accepted definition of a good explanation is its capacity to be understood by a human within a reasonable time frame [42]. We thus gauge participants' comprehension of the model through four aspects divided into two behavioural and two self-reported metrics.

- **Self-Reported Understanding (Immediate and Final)**:
  - **Immediate Understanding.** Self-reported comprehension of the system prediction on a five-point Likert item during the explanation review.

  - **Final Understanding.** This was obtained from an adapted questionnaire by Madsen and Gregor [45] on perceived technical competence and comprehension across eight 5-point Likert items. The questions used and Likert scale are described in Appendix C.1
- **Behavioural comprehension (Precision and Recall):** Building on the methodology proposed by Weld and Bansal [79], we assess participants' behavioural understanding through a simple quantitative task [67]. We ask participants to identify the features that have the most impact on the classifier's prediction according to the explanation. This task evaluates participants' ability to interpret the information provided by the explanations. Understanding is a multifaceted process, with our measures capturing a specific aspect of understanding.
  - **Precision.** Measures the proportion of features correctly identified by the participant among *all the features* they selected.
  - **Recall.** Computes the ratio of features correctly identified by the participant among *all the features* deemed impactful by the explanation.

Confidence. A common measure of user confidence is the agreement rate between the users and the AI model [9, 70, 76]. Therefore, we build upon the methodology of Broon and Holmes [9] to measure users' behavioural confidence.

- **Behavioural Confidence (Following prediction):** Proportion of times the participants modified their prediction in favour of the AI's prediction (when the participant's initial prediction differs from the model's prediction).
- **Self-Reported Confidence (Δ Confidence):** Difference between self-reported confidence before and after seeing the AI-based predictions and explanations.

4.3 Participants

We recruited participants through the Prolific Academic platform. We restricted participation to crowdworkers with at least a high school degree given the complexity of the task. We chose not to limit recruitment to a particular geographical location to promote participant diversity. Finally, we ensured that participants could participate only once. After accepting the task, participants were redirected to the survey. Based on a pilot evaluation with 20 people, we estimated a completion time of 15 minutes for the participants in the control group, and 20 minutes for those with explanations. All participants were paid £9.30 per hour.

To limit Type II errors, we determined the number of respondents on the basis of a power calculation using G*Power [57]. 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 [29]. 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 researchers and companies often rely on them for data labelling tasks [22]. It is therefore vital to investigate their perception and response to AI explanations. We notice, however, that crowdworkers do not capture the particularities of all users, e.g., domain experts. We discuss this limitation in Section 6.4.

Following the task introduction, we assessed whether the participants had read and understood the task through two test questions. Forty participants answered those questions incorrectly and were therefore replaced by 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 participants' comprehension. Then, we assess the influence of these factors on participants' confidence in the AI agent. Finally, we explore the correlation between behavioural and perceived measurements. All the experimental resources of our study are available on Github[5].

	Comprehension							
	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.

To discern the factors that impact participants' comprehension and confidence, we employed a linear model and an ANOVA analysis for each application domain (recidivism and obesity). The linear model includes demographic data (age, gender, education level) along with explanation technique and visual representation as predictive variables. For each statistically significant predictor, we conduct a post hoc analysis using *t*-tests with Bonferroni correction.

5.1 Comprehension

The ANOVA F-scores of each predictor and comprehension metric (both self-report and behavioural) can be found in Table 1. We first observe that the participants' 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 self-reported comprehension right after seeing the explanations (Immediate), we observe a statistically significant effect 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, whereas a significant impact on recall is only observed in the Obesity dataset.

Figure 4 depicts participants' perceived comprehension of the AI system across the explanation methods for both domains. Participants presented with rule-based explanations, , for the obesity dataset report a better understanding of the model in comparison to the control group.

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

[5]https://github.com/j2launay/user_eval

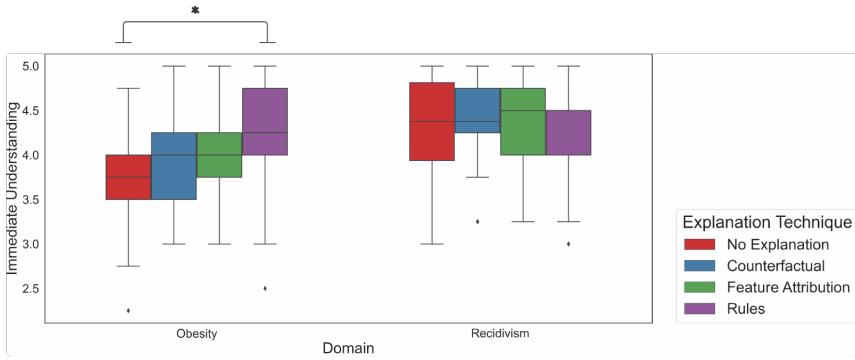


Fig. 4. Perceived understanding of participants **(Immediate)** for both the Obesity and Recidivism domains based on the explanation technique.

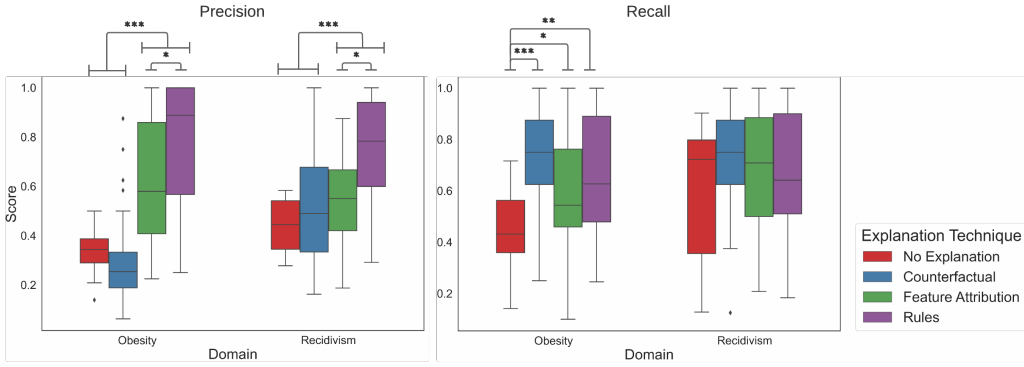


Fig. 5. Precision and recall between the features indicated as important by the participants for the AI's prediction and the features indicated in the explanation. Results are shown for each explanation technique and domain.

to the control group (precision 0.3). Concerning the participants' recall, we observed that in the Obesity domain, participants presented with explanations obtained significantly higher recall than participants without any explanations.

5.2 Confidence

We now assess participants' confidence in the AI system and report the corresponding F-values in Table 2. Our ANOVA analysis shows 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, participants' 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 consider these occurrences as 'initial disagreement'. We find that for the Obesity dataset, the interaction between explanation technique and visual representation significantly impacts the behavioural confidence (Follow Prediction) of initial disagreement cases.

	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).

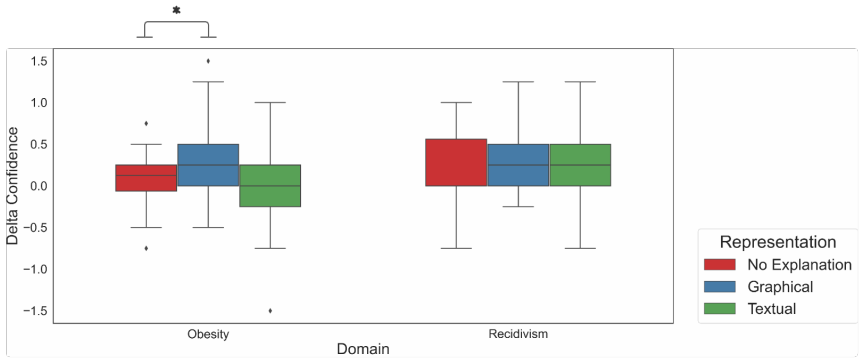


Fig. 6. Difference between the self-reported confidence in the participants' prediction after and before seeing the AI's prediction and explanation (when provided). Results are shown for each domain and representation. Values above zero denote an increase in confidence in the model.

Figure 6 shows that in the Obesity domain, participants exposed to a graphical representation report 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.

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

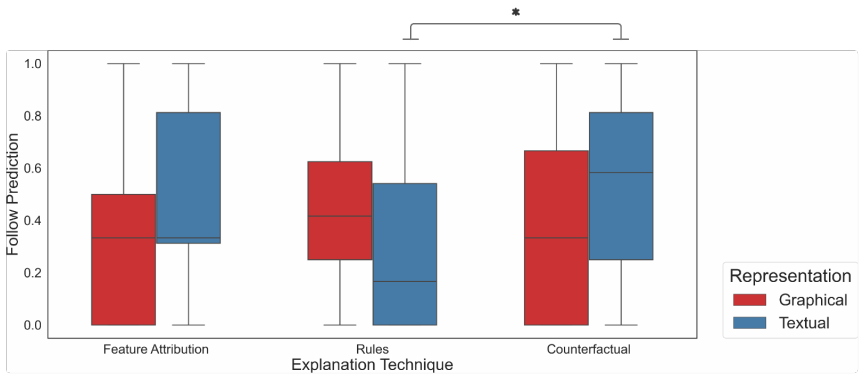


Fig. 7. Proportion of time the participants change their initial prediction to follow the AI's prediction. Results are shown for the Obesity dataset on the combination of explanation technique and representation.

5.3 Perception vs. Behaviour

Finally, we assess the alignment between self-reported and behavioural comprehension and confidence. First, we report the Pearson correlation between perceived and behavioural comprehension. Our results indicate no correlation between participants' perceived comprehension (either immediate or final) and their actual comprehension of the model, as measured by the precision and recall scores. Second, we assess the correlation between self-report and behavioural confidence in the model. We observe correlation scores of 0.43 and 0.49 between the perceived confidence when facing an explanation (Δ Confidence) and the proportion of participants following the AI's prediction (Follow Prediction) for the COMPAS and Obesity datasets, respectively. This suggests a moderate positive correlation between these two measurements.

6 Discussion

We now discuss our key findings, draw design lessons for XAI practitioners, highlight limitations, and outline future research perspectives.

6.1 Impact of Explanation Technique

We assessed the effects of three explanation techniques on participants' comprehension and confidence of two AI models **(RQ1)**. Our findings support our general hypothesis **(H1)**, namely that explanations increase both (a) the participants' comprehension of the AI model and, (b) participants' confidence in the model's predictions. The study also confirms **H2**, i.e., rule-based explanations are most effective in explaining the workings of an AI system. This also aligns with prior work [59, 66]. We hypothesise that this preference for rules is attributable to two factors: (a) its alignment with common educational reasoning principles, and (b) the simplicity of rules in comparison with other explanation techniques. 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 domain. Having more features to grasp makes explanations more beneficial for understanding AI agents. Further, participants' firsthand experience with defendants might be limited, whereas they are more likely to harbour preconceptions about the causes of obesity.

On the other hand, our study shows that participants' precision and self-reported comprehension are comparable to the control group for counterfactual explanations. This stands in 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 (high 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 other explanation techniques.

6.2 Impact of Representation

The impact of representation on users' perception has been well-established [14, 72], and our findings corroborate these prior results **(RQ2)**. In particular, we found that the graphical representations induce higher perceived confidence compared to textual representations (Figure 6). We reckon that these results stem from the elaborate appearance of the graphical presentation, which may give the impression of a greater underlying effort, thereby increasing users' confidence.

Our findings corroborate **H4**, stating that users' confidence in 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, participants' confidence or comprehension of 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, for example by allowing users to interact with the data and narrow down relevant information [73]. Critically, 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 evaluations 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. These needs may include legal requirements or educational purposes [7, 13]. Our experiments show that pre-conceptions and prior knowledge can elicit scepticism towards AI systems. This phenomenon has been also observed in prior work [46], 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 [6]. Prior work highlights that even domain experts display an excess of confidence in AI in the presence of explanation techniques such as feature attribution [32]. 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., $A \Rightarrow Obese$ is not the same as $A \Leftrightarrow Obese$). 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 **system and explanation complexity**. We hypothesise that additional input features in an AI agent may increase the perceived benefit of explanations. It has been also documented that comprehensibility decreases with explanation complexity [16, 48]. 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 identify several limitations to our study. First, our participants consist of crowdworkers, a choice motivated by the increasing role of crowdworkers in the training of and interaction with AI systems. While our participants faced stereotypical decision scenarios, our results may not transfer to domain experts or computer scientists [21, 50, 60]. Contrary to a general audience, computer scientists may be familiar with particular explanation styles and representations, while domain experts may hold stronger preconceptions about their domain of expertise. Furthermore, we did not assess our participants' prior knowledge of the chosen domains, which could have affected participants' performance.

Third, we acknowledge that the impact of explanation techniques on comprehension may also vary with the data modality [30]. 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. Our experimental design required us to control for chart type and, as such, introduce bar types for all explanation techniques. Bar charts, as used in our experiments, are widely employed for feature-attribution explanations on tabular data [55], but are less common for rule-based and counterfactual explanations. Therefore, the effectiveness of various chart styles for representing different explanation types deserves further investigation.

Further, we evaluated participant comprehension through a relatively straightforward task, namely, the identification of the most important features in a decision process. Other tasks could provide additional insights into participant understanding, e.g., use the explanation to reproduce the AI's model behaviour, answer what-if scenarios, or generate explanations [7, 37].

Finally, we highlight that the analysis of our post-questionnaire on understanding yielded unexpectedly non-significant differences across various explanation techniques and representations in contrast to prior work [74, 78]. 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. This limited interaction might have contributed to the absence of statistical significance in our findings, as previously suggested by Van der Waa et al. [74]. 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, as reported in Section 5.3, we found no correlation between users' perceived understanding and their actual comprehension of the model, as measured by the precision and recall scores. These findings are in line with existing research [15, 30, 64, 77]. Understanding why users elicit confidence without corresponding behavioural alignment, or why they perceive comprehension without demonstrating it in practice remains a valuable open research direction.

7 Conclusion

In this article we report on a study of the impact of explanation technique and visual representation on users' comprehension and confidence in XAI systems. 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 results in the highest users' comprehension. Counterfactual explanations presented as text elicited higher levels of confidence, while the opposite was observed for feature-attribution and rule-based explanations. Our results are not entirely consistent across the two 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.

Acknowledgments

This work is supported by the Carlsberg Foundation, grant CF21-0159 and the ANR JCJC FAbLe, grant no. ANR-19-CE23-0019-01.

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. doi: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. arXiv:2106.07483 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 and Decision Making* 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 *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1078–1088. https://www.ifaamas.org/Proceedings/aamas2019/pdfs/p1078.pdf

[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* 30, 4 (2022), 593–610. doi:10.1007/s10506-022-09324-9

[6] 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. *Proceedings ACM Human-Computer Interaction* 7, CSCW1, Article 27 (April 2023), 17 pages. doi:10.1145/3579460

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

[8] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and Survey of Explanation Methods for Black Box Models. arXiv:2102.13076 https://arxiv.org/abs/2102.13076

[9] 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. https://www.scienceopen.com/document?vid=f92a0587-18fc-4e40-b561-5ee691c53fc9

[10] Lorella Bottino and Mario Cannataro. 2023. Explanation of machine learning models for predicting obesity level using Shapley values. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, Los Alamitos, CA, USA, 3288–3291. doi:10.1109/BIBM58861.2023.10385994

[11] 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 (2009), 21–40. doi:10.1177/0093854808326545

[12] 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 (2018), 143–144. doi:10.3399/bjgp18X695213

[13] 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, Cham, 48–71. https://link.springer.com/chapter/10.1007/978-3-031-44067-0_3

[14] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300789

[15] 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 *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. doi:10.1145/3397481.3450644

[16] Nelson Cowan. 2010. The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science* 19, 1 (2010), 51–57. doi:10.1177/0963721409359277 arXiv:https://doi.org/10.1177/0963721409359277 PMID: 20445769.

[17] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proceedings of the ACM on Human-Computer Interaction - CSCW* (2025), 1–30.

[18] Luca Deck, Astrid Schomäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. 2024. Mapping the Potential of Explainable AI for Fairness Along the AI Lifecycle. arXiv:2404.18736 https://arxiv.org/abs/2404.18736

[19] Julien Delaunay, Luis Galárraga, and Christine Largouët. 2020. Improving Anchor-based Explanations. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3269–3272. doi:10.1145/3340531.3417461

[20] Julien Delaunay, Luis Galárraga, and Christine Largouët. 2022. When Should We Use Linear Explanations?. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 355–364. doi:10.1145/3511808.3557489

[21] Julien Delaunay, Luis Galárraga, Christine Largouët, and Niels van Berkel. 2023. Adaptation of AI Explanations to Users' Roles. 7 pages. https://vbn.aau.dk/en/publications/adaptation-of-ai-explanations-to-users-roles 2023 ACM CHI Conference on Human Factors in Computing Systems, CHI 23 ; Conference date: 23-04-2023 Through 28-04-2023.

[22] 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. doi:10.1145/3531146.3534647

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

[24] 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. doi:10.1038/nature21056

[25] Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 38, 5 (01 Sep 2024), 2770–2824. doi:10.1007/s10618-022-00831-6

[26] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. https://api.semanticscholar.org/CorpusID:44063479

[27] 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. doi:10.1145/3236009

[28] David Gunning. 2019. DARPA's explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, ii. doi:10.1145/3301275.3308446

[29] Joseph Hair, William Black, Barry Babin, and Rolph Anderson. 2013. *Multivariate Data Analysis*. Pearson Education Limited, Bagsvaerd, Denmark. https://books.google.es/books?id=VvXZnQEACAAJ

[30] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5540–5552. doi:10.18653/v1/2020.acl-main.491

[31] Guy Hoffman. 2019. Evaluating Fluency in Human-Robot Collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218. doi:10.1109/THMS.2019.2904558

[32] 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.

[33] Alon Jacovi. 2023. Trends in Explainable AI (XAI) Literature. arXiv:2301.05433 https://arxiv.org/abs/2301.05433

[34] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. 2023. Adversarial Counterfactual Visual Explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 16425–16435. doi:10.1109/CVPR52729.2023.01576

[35] Todd Kulesza, Simone Stumpf, Margaret 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 *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE Computer Society, Los Alamitos, CA, USA, 3–10. doi:10.1109/VLHCC.2013.6645235

[36] Francesca Lagioia, Riccardo Rovatti, and Giovanni Sartor. 2023. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. *AI & SOCIETY* 38, 2 (01 Apr 2023), 459–478. doi:10.1007/s00146-022-01441-y

[37] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *Proceedings ACM Human-Computer Interaction* 7, CSCW2, Article 357 (Oct. 2023), 35 pages. doi:10.1145/3610206

[38] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco California USA). ACM, New York, NY, USA, 1675–1684. https://pubmed.ncbi.nlm.nih.gov/27853627/

[39] 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, Aachen, Germany. http://ceur-ws.org/Vol-2582/paper6.pdf

[40] 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

[41] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2018. Comparison-Based Inverse Classification for Interpretability in Machine Learning. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*. Springer International Publishing, Cham, 100–111. doi:10.1007/978-3-319-91473-2_9

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

[43] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777. doi:10.5555/3295222.3295230

[44] 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 *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 1139–1149. doi:10.1145/3447548.3467265

[45] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. *American Conference for Irish Studies* 1 (01 2000).

[46] 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 *Proceedings CHI (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 442, 18 pages. doi:10.1145/3544548.3581575

[47] 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. doi:10.1016/j.dib.2019.104344

[48] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 81–97. doi:10.1037/h0043158

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

[50] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11 (2021), 24:1–24:45. doi:10.1145/3387166

[51] Christoph Molnar. 2018. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published. https://christophm.github.io/interpretable-ml-book/

[52] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. doi:10.1145/3351095.3372850

[53] Jeroen Ooge and Katrien Verbert. 2022. Explaining Artificial Intelligence with Tailored Interactive Visualisations. In *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 120–123. doi:10.1145/3490100.3516481

[54] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 383 (Nov. 2024), 31 pages. doi:10.1145/3686922

[55] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. doi:10.1145/3411764.3445315

[56] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 344–350. doi:10.1145/3375627.3375850

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

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Washington, D.C., USA, Article 187, 9 pages. https://aaai.org/papers/11491-anchors-high-precision-model-agnostic-explanations/

[60] Mireia Ribera and Àgata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *Proceedings of the International Conference on Intelligent User Interfaces (CEUR Workshop Proceedings)*. CEUR-WS.org, Aachen, Germany, 7. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

[61] Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating Realistic Natural Language Counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 3611–3625. doi:10.18653/v1/2021.findings-emnlp.306

[62] 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

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

[64] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. doi:10.1145/3301275.3302308

[65] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, Brookline, Massachusetts, USA, 3145–3153.

[66] Arora Siddhant, Pruthi Danish, Sadeh Norman, Cohen William W., Lipton Zachary C., and Neubig Graham. 2022. Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 5 (2022), 5277–5285. doi:10.1609/aaai.v36i5.20464

[67] 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 *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 23. https://openreview.net/forum?id=v6s3HVjPerv

[68] Eduardo A. Soares and Plamen Angelov. 2019. Fair-by-design explainable models for prediction of recidivism. *CoRR* abs/1910.02043 (2019), 5. http://arxiv.org/abs/1910.02043

[69] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, Brookline, Massachusetts, USA, 3319–3328. doi:10.5555/3305890.3306024

[70] 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 *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 767–781. doi:10.1145/3490099.3511160

[71] 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 (2022), 417–473. doi:10.1007/s10506-021-09301-8

[72] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York,

NY, USA, Article 245, 13 pages. doi:10.1145/3411764.3445365

[73] Niels van Berkel, Benjamin Tag, Rune Møberg Jacobsen, Daniel Russo, Helen C. Purchase, and Daniel Buschek. 2024. Impact of interaction technique in interactive data visualisations: A study on lookup, comparison, and relation-seeking tasks. *International Journal of Human-Computer Studies* 192 (2024), 103359. doi:10.1016/j.ijhcs.2024.103359

[74] 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. *Artificial Intelligence* 291 (2021), 103404. doi:10.1016/j.artint.2020.103404

[75] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 1–47. doi:10.2139/ssrn.3063289

[76] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. doi:10.1145/3397481.3450650

[77] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (2022), 27:1–27:36. doi:10.1145/3519266

[78] 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 *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 7–9. doi:10.1145/3308532.3329441

[79] Daniel S. Weld and Gagan Bansal. 2018. Intelligible Artificial Intelligence. *CoRR* (2018), 8. http://arxiv.org/abs/1803.04263

[80] 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 Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65. doi:10.1109/TVCG.2019.2934619

[81] Linyi Yang, Eoin 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 *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6150–6160. doi:10.18653/v1/2020.coling-main.541

[82] Jun Yuan and Enrico Bertini. 2022. Context sight: model understanding and debugging via interpretable context. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (Philadelphia, Pennsylvania) *(HILDA '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 7 pages. doi: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 provide the details of the code, classifier, and datasets utilised in our experimental evaluation. Appendix B presents a comprehensive table with 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 the choices 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 [6].

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[7].

Obesity: This dataset is originally composed of 16 features and a target obtained from questions detailed in [47]. 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. There are five binary 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 [47][8].

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.

[6] https://github.com/j2launay/user_eval
[7] https://github.com/propublica/compas-analysis/
[8] 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 participants to complete two various questionnaires, each one evaluating a given criteria. In this section we present 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 [45]. This questionnaire is originally :

(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.

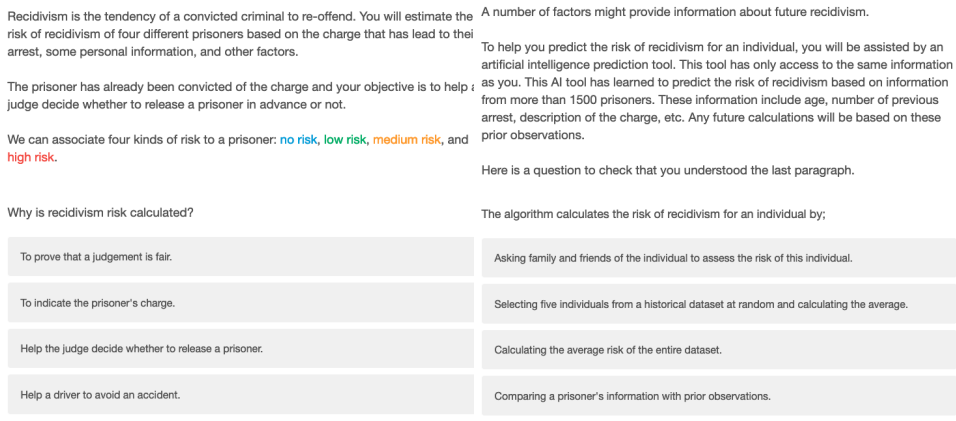
For each of these questions, Madsen and Gregor [45] recommended this 5 Likert scale:

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

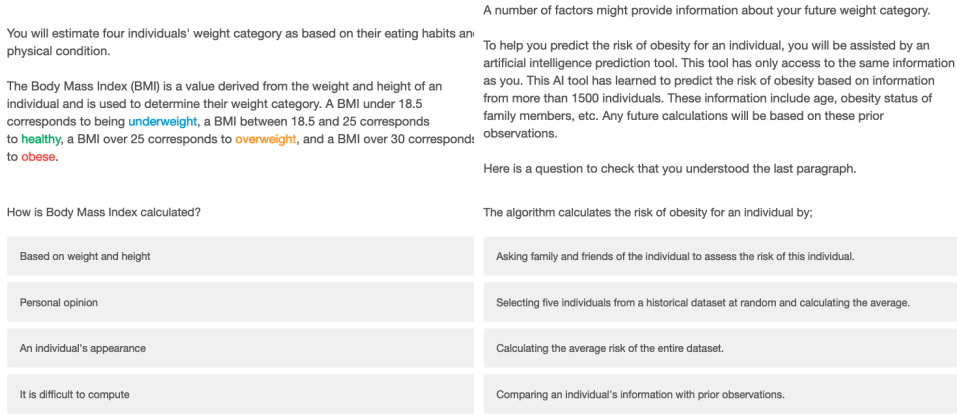
We ask the user two questions in order to verify that they understand and will try efficiently to complete the questionnaire.

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.

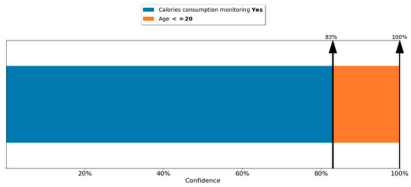
In the representation of **rule-based explanations**, we utilised stacked bars, starting with the rule's condition that induced the highest initial confidence in the model's prediction. Subsequently, we iteratively added conditions that improved the most the model's confidence, given that existing conditions 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 in 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.

Based only on the above information, the artificial intelligence (AI) tool has predicted **healthy**.
The following graph shows the criteria that impacted the AI's prediction. Each of the colored bars represent the importance of one particular user's answer to the final prediction.
The numerical values at the top correspond to the increasing confidence that the AI tool predicts **healthy** for this user.

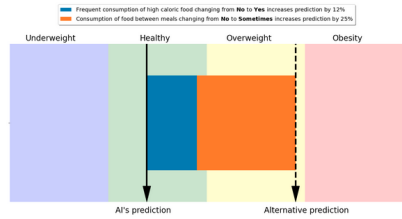


You now know everything required to proceed to the tasks!

Rule-based.

As highlighted in the graph below and based only on the above information, the AI tool has predicted **healthy**.

The following graph shows the criteria that impacted the AI's prediction. The AI computes a value between 0% and 100% to classify the individual. This value corresponds to the "AI's prediction" vertical black bar and falls into one of the four categories: **underweight** (below 25%), **healthy** (between 25% and 50%), **overweight** (between 50% and 75%), and **obesity** (above 75%).



The colored bars indicate what the individual must do in order to modify the AI's prediction the most effectively. The length of the bars correspond to the importance of changing one answer's value to another.

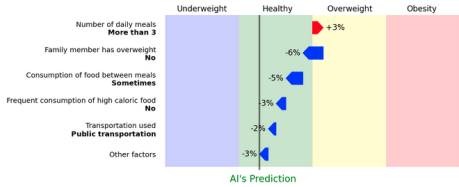
You now know everything required to proceed to the tasks!

Counterfactual.

Based only on the above information, the AI tool has predicted **healthy**.

The following graph shows the criteria that impacted the AI's prediction. The red bars indicate an increased chance of being overweight and obese. The blue bars indicate an increased chance of being underweight or healthy.

The values on the side of the bars correspond to the impact of the specific factor on the prediction. The "Other parameters" bar indicates the impact of all other factors not presented in the graph.



By summing the values associated with each response by the AI, we obtain a value between 0% and 100%. This value corresponds to the vertical black bar and falls in one of the four categories: **underweight** (below 25%), **healthy** (between 25% and 50%), **overweight** (between 50% and 75%), **obesity** (above 75%).

You now know everything required to proceed to the tasks!

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 participants' 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 behavioural 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 behavioural 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 [45] 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.

Received January 2024; revised July 2024; accepted October 2024