Evaluating the Practicality of Counterfactual Explanations

Nina Spreitzer[1,*], Hinda Haned[1,2] and Ilse van der Linden[1,2,*]

[1]University of Amsterdam, Amsterdam, The Netherlands
[2]Civic AI Lab, Amsterdam, The Netherlands

Abstract

Machine learning models are increasingly used for decisions that directly affect people's lives. These models are often opaque, meaning that the people affected cannot understand how or why the decision was made. However, according to the General Data Protection Regulation, decision subjects have the right to an explanation. Counterfactual explanations are a way to make machine learning models more transparent by showing how attributes need to be changed to get a different outcome. This type of explanation is considered easy to understand and human-friendly. To be used in real life, explanations must be practical, which means they must go beyond a purely theoretical framework. Research has focused on defining several objective functions to compute practical counterfactuals. However, it has not yet been tested whether people perceive the explanations as such in practice. To address this, we contribute by identifying properties that explanations must satisfy to be practical for human subjects. The properties are then used to evaluate the practicality of two counterfactual explanation methods (CARE and WachterCF) by conducting a user study. The results show that human subjects consider the explanations by CARE (a multi-objective approach) to be more practical than the WachterCF (baseline) explanations. We also show that the perception of explanations differs depending on the classification task by exploring multiple datasets.

Keywords

explainable AI, counterfactual explanations, practicality, human-friendly explanations, user study

1. Introduction

Machine learning (ML) models are increasingly used for automated decision-making impacting people's lives [1]. Some typical applications for ML model decisions are approving a requested loan [2], hiring an applicant [3], or setting the price rates for insurance contracts [4]. ML models are often opaque, meaning users cannot trace back how the decision is made [5]. In light of this automated decision-making, the European Union put forward a General Data Protection Regulation (GDPR) [6]. The GDPR includes a "right to explanation" [7], meaning affected people are entitled to request an explanation for a decision that has been made about them. To serve this right the research field of explainability for ML models is continuously growing [8]. Explainability aims to make the functioning of a model clear and easy to understand for

XAI.it 2022 - Italian Workshop on Explainable Artificial Intelligence
*Corresponding authors.
nina.c.spreitzer@gmail.com (N. Spreitzer); h.haned@uva.nl (H. Haned); i.w.c.vanderlinden@uva.nl (I. v. d. Linden)
https://github.com/ninaspreitzer (N. Spreitzer)

a given audience [9]. However, an ongoing debate in legal and ML communities discusses what this right should entail and what specific requirements must be met [10]. One of the challenges is that the audience to whom the explanation is directed may vary. Arrieta et al. [9] define several groups of people requiring explainability including system developers, users with domain expertise, and users affected by the model decisions. Since the audience does not necessarily have technical skills or domain knowledge, explainability methods must also be suitable for non-technical users without expertise. Researchers in this area [11][9][12] commonly distinguish between two families of explainable approaches, namely ante-hoc and post-hoc approaches. Ante-hoc approaches focus on making models inherently interpretable [13]. Whereas, post-hoc methods use comprehensible representations to produce useful approximations of the model's decision-making process [14]. This is done while treating the model as a black box without trying to reveal any knowledge about the model's functioning [15].

We focus on a particular method of post-doc explanations, called counterfactual explanations [16]. A counterfactual explanation proposes minimal changes to the input data that lead to a different model outcome. It answers the natural question: "Would changing a certain factor have changed the outcome?" [17]. They can be seen as recommendations for what to change to achieve a desired model outcome [18]. We focus on two methods for computing counterfactual explanations, the original approach proposed by Wachter et al. [16], which we refer to as *WachterCF*, and a framework proposed by Rasouli et al. [19], called CARE. Section 4 elaborates on how the methods compute a counterfactual instance and on the differences between them. Counterfactual explanations are viewed to be easy to understand [16] and human-friendly [8]. Wachter et al. [16] state that counterfactual explanations are "practically useful for understanding the reasons for a decision". The Oxford Dictionary[1] defines practical as "concerned with the actual doing or use of something rather than with theory and ideas". Consequently, explanations must go beyond a purely theoretical concept to serve as practical explanations. However, there is limited work that attempts to evaluate the perception of counterfactual explanations in practice, and previous work has offered criticism on their applicability in real-life settings (see Section 2). Our contribution lies in first defining a set of properties that counterfactual explanations should satisfy in order to be considered practical. These properties are then used to define questions for a user study. The user study tests how users perceive counterfactual explanations computed by CARE and WachterCF in two different contexts. The research questions we explore in this work are as follows.

RQ: How practical are counterfactual explanations for human subjects?

- Q1 What properties must counterfactual explanations serve to be practical for human subjects?
- Q2 How do human subjects perceive counterfactual instances proposed by CARE compared to WachterCF?
- Q3 How does the perception of counterfactual explanations differ depending on the classification task?

[1]https://languages.oup.com/google-dictionary-en/

The remainder of this paper is structured as follows. In Section 2, we present related work, including an overview of proposed methods to compute counterfactual explanations and conducted user studies. Section 3 outlines the methodology we follow to explore our research questions. In Section 4 we specify desired properties for practical explanations for human subjects. Section 5 details the experimental setup of our user study, and Section 6 highlights the most important findings. Finally, we conclude the paper with a discussion by reflecting on the results and limitations, as well as outlining future work.

2. Related Work

In this section, we first outline proposed methods for computing counterfactual explanations. Secondly, we elaborate on limitations of the use of counterfactual explanation methods in practice. This is followed by a discussion of related work addressing these limitations by introducing new objectives to achieve practicality. Finally, we discuss related user studies conducted to evaluate counterfactual explanations.

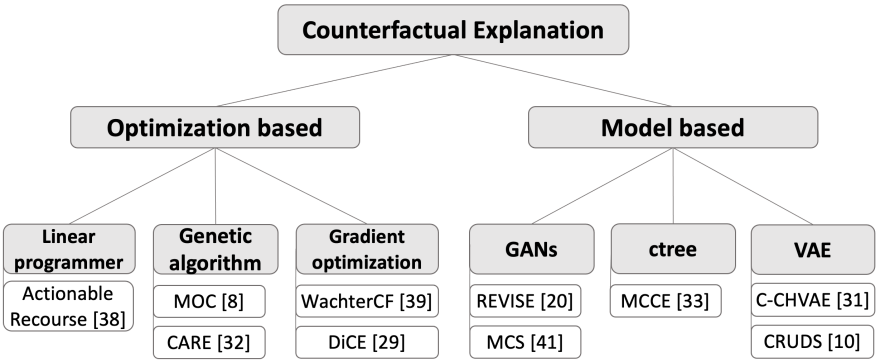

Figure 1: Overview of counterfactual explanation methods, categorized into optimization-based and model-based approaches.

2.1. Methods for generating counterfactual explanations

In recent years, many methods for generating counterfactual explanations have been proposed [20]. A complete literature review of the proposed methods exceeds the scope of this paper; Figure 1 provides an overview. We distinguish between approaches that are optimization-based [21][22][19][1][16] and ones that are model-based [23][24][25][20][26]. Since we compare two optimization-based methods, we further elaborate on the different approaches of those methods. The first counterfactual explanation method, WachterCF [16], defines a counterfactual instance as the closest point to the input data that results in a different prediction. The closest point is found by minimizing a distance function between the original and the counterfactual feature vector using stochastic gradient descent. The method is restricted in the way that the black-box model needs to be differentiable, and the proposed distance function only works with continuous features. Ustun et al. [1] solve the optimization problem using a linear program,

focusing on instances that suggest altering features only in a way that is actionable for the end-user. Similar to WachterCF, Diverse set of Counterfactual Explanations (DiCE) [22] is based on gradient optimization but provides the user with multiple diverse counterfactual instances. DiCE also introduces additional terms to the objective function to include further constraints. The Multi-Objective Counterfactual explanations (MOC) [21] is the first framework that formalizes counterfactual explanations as a multi-objective optimization problem. Additionally, Rasouli et al. [19] propose a framework to generate Coherent Actionable Recourse based on sound counterfactual Explanations (CARE), solving a multi-objective problem based on a hierarchical objective set. The overview in Figure 1 is based on a categorization of methods following Redelmeier et al. [20].

2.2. Limitations on counterfactual explanations in practice

In their work, Barocas et al. [27] show that the computation of counterfactual explanations often relies on widely overlooked assumptions that are necessary for counterfactual explanations to be accepted in real life. It is assumed that the highlighted feature changes always correspond to actual actions. The main problem with this is that the features are not independent, as the actions are likely to affect other features simultaneously. The authors also state that most computed counterfactual explanations only rely on the distribution of training data and overlook that not everything that can be considered equivalent in the training data is automatically equivalent for individual data subjects. In other words, specific suggestions can be practical for some people but may not be practical for others. Laugel et al. [15] claim that assumptions make counterfactual explanations unreliable in many contextual uses. They argue that post-hoc explanations may not be faithful to original data because they are prone to robustness and complexity problems, such as overfitting or immoderate generalization. This would lead to unsatisfying interpretability. The authors outline that a counterfactual instance needs to provide a set of plausible changes a human can act upon in practice and also needs to result from given data points.

2.3. Taxonomy of desiderata for counterfactual explanations

Researchers have responded to this criticism by developing counterfactual frameworks with different objective functions to satisfy specific properties [1][19][22][21][28][29]. Redelmeier et al. [20] outline some of these attributes that have been proposed in previous work. These include counterfactuals that aim to be *close* (decrease distance to original input data), *sparse* (reduce number of feature changes), *feasible* (increase density of area around counterfactual instance), *actionable* (restrict the features that can change with constraints), and *diverse* (increase distance between suggested counterfactual instances for same input). WachterCF focuses on being close by minimizing the distance between a counterfactual instance and the original input data. Thus, it does not explicitly aim to satisfy any other objectives. In contrast, CARE solves a multi-objective problem including four desirable properties, such as *proximity* (being a neighbor of the ground-truth data [15]), *connectedness* (relationship between counterfactual instance and training data), *coherency* (keeping the consistency of (un)changed features) and *actionability* (include preference, e.g. restrictions or immutability of features). CARE's objectives are set

up as a hierarchy with four modules. *Validity* is on the bottom as a foundation, followed by *Soundness*, *Coherency*, and at the top *Actionability*. The modules are independent of each other, and all relate to an individual objective function. The low-level modules, validity and soundness, deal with fundamental and statistical properties of counterfactual explanations. A valid counterfactual is the foundation, meaning the counterfactual instance needs to alter the outcome of the ML model. Soundness includes proximity and connectedness. Hence, a sound counterfactual instance originates from similar observed data and connects to existing knowledge. The high-level modules manage coherency between features and user preferences. The coherency module constructs a correlation model and uses this model to guarantee that all connected features alter in accordance with a particular feature change. Furthermore, CARE ensures actionability by enabling users to set preferences. Those preferences can include defining (im)mutable features and also setting value ranges of specific attributes.

2.4. User studies evaluating counterfactual explanations

Warren et al. [30] have examined how well users are able to predict model outcomes after looking at counterfactual explanations, and self-reported satisfaction and trust in the explanations. Our user study adds to their findings by assessing user perceptions of practicality properties grounded in social sciences [31]. Förster et al. [28] examined the coherency of counterfactual explanations by testing if users consider an instance as real (taken from the training data) or fake (computed counterfactual instance). Additionally, the users were asked if they thought the instance was suitable to explain the outcome. The classification task explored in their study was housing price prediction. The user study design that we present examines classification tasks that affect human decision subjects. Our work can contribute to better understand user perception of other desiderata and to what extent it depends on the classification task.

3. Methodology

To answer the research questions, we first formalize our evaluation of practicality by defining a set of properties. In the user study, we evaluate the practicality properties by assessing how human subjects perceive counterfactual explanations generated by CARE compared to explanations generated by WachterCF. We explore this for two classification tasks using different datasets.

3.1. Formalizing Practicality

To determine what properties counterfactual explanations must satisfy to be practical, we review existing literature and examine the requirements of ML explanations. Based on that, we define a set of properties that make counterfactual explanations practical for humans. We focus on properties that affect how human subjects perceive counterfactual explanations. Thus, we examine the relation between explanation and user perception rather than the relation between explanation and model behavior (i.e. robustness). The defined properties serve as a benchmark for the practicality of generated counterfactuals.

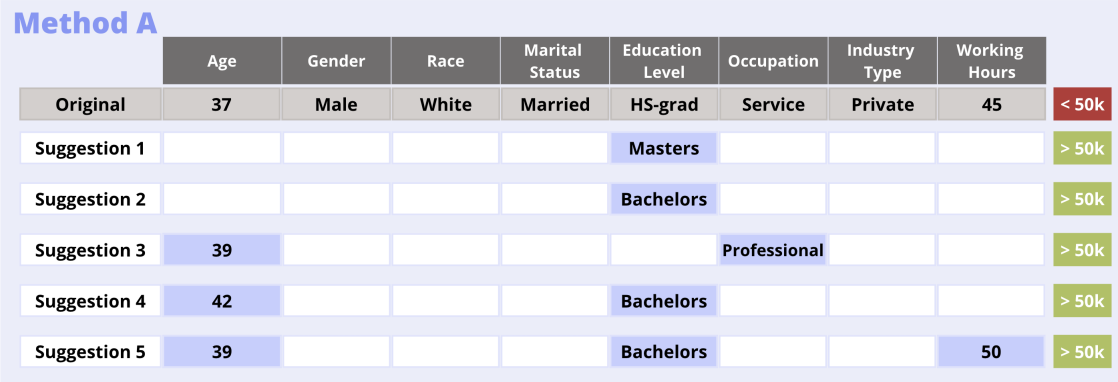

Figure 2: Example screen of user study: Counterfactual explanations computed with CARE presented in a visually appealing manner. The counterfactual instance computed with WachterCF followed the same design.

3.2. User Study

The remaining sub-questions are answered by performing a user study with human subjects. The study aims to compare the perception of explanations provided by CARE with explanations provided by WachterCF. We use the *Wilcoxon Sign Ranked* test [32] to determine if the results are significantly different. The test is non-parametric, meaning the answers do not need to follow a normal distribution and tests the difference between paired samples.

3.2.1. Define Practicality Evaluation

To measure the practicality of counterfactual instances, we formulate questions mapped to practicality properties. We compare the methods indirectly, meaning we first evaluate explanations computed by CARE and then ask the same questions for explanations following WachterCF. The answers are given without direct comparison. Only one question at the end asks directly which method is preferred. This question helps to see if the personal opinion matches the findings.

3.2.2. Define Target Group & Sample Size

Before conducting a user study, the target group to be studied must be defined. Since any individual could be affected by automated decision-making, we use convenience sampling [33] by collecting information from a conveniently available pool of respondents. This type of sampling is prone to bias, which will be discussed in Section 7. One distinction we make in the analysis is whether participants are familiar with machine learning methods because we believe there may be a potential difference in the perception of explanations depending on the prior knowledge of participants. Our objective was to gather enough responses to test for statistical significance. For this reason, we aimed for 100 respondents, preferably evenly distributed in terms of machine learning literacy.

3.2.3. Classification Tasks and Data

To examine different classification tasks, we use two separate datasets. In the user study, the two scenarios were evenly distributed among the participants. Both datasets used can be accessed through the UCI machine learning repository [34]. The Adult Income dataset [35] is used to classify whether an individual is likely to have an income of more than 50k per year. The Student Performance dataset [36] is used for predicting what final grade an individual is likely to have for a school subject. For the user study, we simplify the classification task to predict if a student will pass or fail the subject. Both decisions are based on personal information about the individuals.

3.2.4. Preprocessing & Classification Model

The preparation of the Adult dataset follows Zhu [37] by dropping columns that are not needed and grouping categorical values to end up with a more concise set of possible values. For the student dataset, we dropped columns to make the number of possible feature changes similar to the adult dataset to ensure a fair comparison. An overview of the final datasets can be found in the Appendix. For the remaining steps until the computation of the counterfactual explanation, we follow Rasouli et al. [19]. The numerical features of both datasets are standardized and categorical features are converted to ordinal encoding and are further one-hot encoded. The data sets are split into 80% training and 20% test set, and only the training set is used for creating the classification model. We follow Rasouli et al. [19] in the choice of model: a multi-layer neural network with two hidden layers, each including 50 neurons. The resulting model for the Adult Income classification task has an *F1-score* of 0.81. The model for the Student Performance dataset has an *F1-score* of 0.71.

3.2.5. Generating Counterfactual Explanation

With this setup, we create counterfactual instances using WachterCF and CARE. CARE provides the user with the possibility of defining constraints for actionability. We pre-define these constraints following Rasouli et al. [19] by setting gender and race as *fix* (immutable value) and age as *ge* (can only be greater or equal to the current value). Additionally, the CARE framework also allows the end-user to set the number of how many counterfactual explanations should be given for a single instance. We chose to set this number for five as we believe that five different suggestions is both comprehensible and noticeable different from providing only one counterfactual suggestion.

4. Practicality

Considering that the recipient of the explanation is a human subject, it is crucial to make the explanations human-friendly. In addition to other objective functions, the explanations must be presented in a way such that people understand them [38], and the goodness of an explanation depends on its audience [9]. Thus, human perception is an essential factor for explanation methods [13]. We recognize that there is a possible trade-off between making explanations

human-friendly while retaining high accuracy. Nevertheless, we highlight that this trade-off is not relevant for our current work since the aim is not to improve the human-friendliness of explanations. This work assesses how humans perceive the explanation methods that claim to be practical for human subjects.

Miller [31] summarizes human-friendly characteristics. Depending on the context, some features may be more important than others [38]. Concerning counterfactual explanations, we have defined the following set of properties that lead to practical explanations. This set is then used as a benchmark to evaluate how humans perceive counterfactual explanations in practice.

- **Contrastiveness**: Humans are not interested in why an event happened but rather in why that event happened instead of another. In the context of counterfactual explanations, we measure contrastiveness by how well the user understands what needs to change to get the opposite outcome.
- **Selectivity**: Generally, humans do not expect a complete cause of an event. Humans are used to selecting a smaller set of causes and treating it as a complete explanation. Therefore, counterfactual explanations can provide selectivity by changing only a subset of features as well as providing different suggestions.
- **Social**: The explaining method is part of an interaction between the end-user and the system explaining. As a result, the social environment, the target audience, and the use case need to be considered. For counterfactual explanations, this means that the proposed changes should be made realistically in the given use case of the affected person.
- **Truthful**: A human-friendly explanation needs to make sense. In other words, the user must perceive the counterfactual suggestions to reach the other result as plausible.
- **Consistent with prior beliefs**: As described by the confirmation bias [39], people tend to ignore information that is inconsistent with their prior beliefs. Applied to counterfactual explanations, this means that end-users are more likely to consider explanations that suggest changes that are expected in advance.

5. Experimental Setup

The following section describes how the defined practicality properties are measured and how the user study is designed.

5.1. Practicality Measurements

Table 1 provides an overview of our user study questions and the possible responses. We specify questions to measure how well the counterfactual instances satisfy the practicality properties. We map the questions shown to the properties as we believe they represent expectations of counterfactual explanations with respect to those properties. Question 1 is asked at the beginning, followed by Questions 2-8 are asked once for each counterfactual method. Finally, Question 9 which explanation is preferred. To make the scenario and questions more understandable, we call the person represented by the input data *Charlie*.

Table 1
User Study Questions: The following questions are formulated based on the property set to evaluate practicality.

Question	Measurement	Property
1 What attribute(s) would you expect to change for Charlie to instead get the outcome of "earning above 50k" / "passing the course"?	Multiple Choice: List of features	Consistency prior beliefs
2 How surprised are you with the suggested changes in attributes to get the outcome of "earning above 50k" / "passing the course"?	Likert Scale: 1. Not at all 7. Very surprised	Consistency prior beliefs
3 How well does the method explain to you what Charlie needs to change to get "earning above 50k" / "passing the course"?	Likert Scale: 1. Not at all 7. Very well	Contrastive-ness
4 Based on the explanation, what attribute(s) would you consider as most important to change the model outcome?	Multiple Choice: List of features	Consistency prior beliefs
5 In your opinion, the amount of five different suggestions / one suggestion is ___ to explain the model outcome.	Single Choice: Too little/ Enough/ Too many	Selectivity
6 In your opinion, the variation of attributes in the suggestion(s) is ___ to explain the model outcome.	Single Choice: Too little/ Enough/ Too much	Selectivity
7 Do you think Charlie could realistically act upon the suggestions to change the model outcome to "earning above 50k" / "passing the course"?	Likert Scale: 1. Not at all 7. Fully	Social
8 Do you think the suggestions make sense in order to retrieve the model outcome to "earning above 50k" / "passing the course"?	Likert Scale: 1. Not at all 7. A lot	Truthful
9 Which method would you prefer as an explanation for the outcome of the ML model?	Single Choice: Method A Method B	

5.2. User Study Setup

The questionnaire is designed using the survey tool Qualtrics[2] and send out digitally. The study starts by introducing ML models, how they are used for automated decision-making, and how counterfactual instances help to provide explanations. Next, the scenario and the datasets are explained, including an illustrative counterfactual instance, similar to Figure **??**. Then, the participants are asked what properties they expect to change to retrieve the other outcome. This is followed by showing the first counterfactual explanation, which is computed by CARE. The participants are asked to answer questions 2-8 concerning the CARE explanation. After answering the questions, they are shown the second explanation computed by WachterCF and are asked the same questions. Finally, the last question asks which method is preferred and gives an opportunity to leave a comment. Figure 2 illustrates how the first explanation calculated by CARE is shown to the participants. We considered several options for the order of showing

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

explanations from the two different methods. The final user study shows each participant the CARE explanations first followed by the WachterCF explanation. Another option would be to randomize the order of explanation methods shown between participants. The second option has the advantage of reducing bias introduced by the order. We chose the fixed order (CARE followed by WachterCF) with the reasoning that our focus is on evaluating CARE (a method intended for practicality) against the baseline WachterCF method. By showing CARE first we ensure that CARE is evaluated a priori without the influence of another counterfactual explanation. If CARE is not perceived to be more practical than the WachterCF explanation shown second should obtain similar responses.

5.3. Randomized Instances

The instance shown to participants refers to individuals who initially received a negative prediction (less than 50k per year or a negative final grade) and are given a counterfactual instance as an explanation. We randomly select twenty instances from the Adult Income test data and ten instances from the Student Performance test data. The number differs because the Adult Income dataset contains more instances than the Student Performance dataset. The instances are evenly distributed among users and randomly assigned. Furthermore, counterfactual explanations are presented in a visually appealing manner, since we want participants to focus on the content of the explanation.

6. Results

We have collected 135 responses. 69 participants received the Adult Income dataset, and 66 received the Student Performance dataset. Additionally, 70 responses indicate that they are familiar with ML models, while the remaining 65 imply not being that familiar with ML models. We analyze the results the following way. First, we compare the responses of explanations computed by CARE compared to the ones computed by WachterCF. These results are used primarily to answer the second sub-question. Secondly, we examine the responses for both datasets separately to answer our last sub-question if the perception differs between different classification tasks. Finally, we investigate whether there is a difference in responses among the respondents that are or are not familiar with machine learning models.

6.1. Overall Comparison: CARE vs. WachterCF

We start by analyzing all questions that include Likert scales (Questions 2, 3, 7 & 8) and single-choice answers (Questions 5 & 6). Then we examine questions that ask for selecting features (Questions 1 & 4), followed by looking into the participants' preference of the two methods (Question 9).

6.1.1. Results to Likert Scale & Single Choice Questions

Figure 3 shows a stacked bar chart for Questions 2,3,5,6, and 7, comparing the responses of WachterCF and CARE. Red colors represent negative responses, yellow colors represent

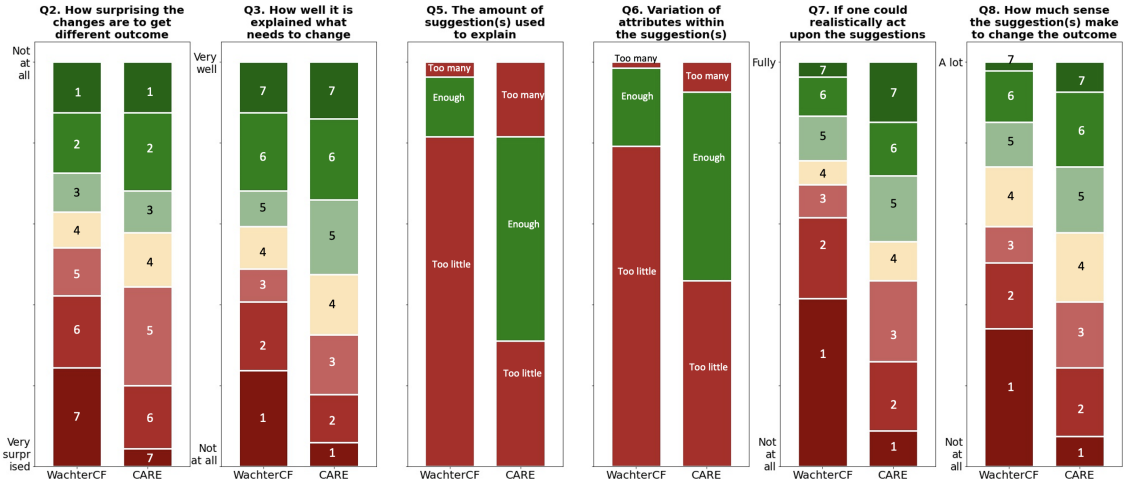

Figure 3: Responses to questions with quantitative responses (Likert scale and Single Choice). Red-colored answers (bottom) indicate negative responses, yellow (middle) neutral and green positive (top). For each question the responses for WachterCF is compared with CARE. Based on the P values of a Wilcoxon Sign Ranked test the responses differ significantly between the methods. The P value for each question shown is accordingly: 0.001, 0.002, 9.3*e−−09, 1.6*e−−06, 1.6*e−07, 3.1*e−06

neutral responses, and green colors represent positive responses. Looking at the graphs, it is noticeable that the counterfactual explanations calculated with WachterCF received more negative responses than those calculated with CARE. We use a *Wilcoxon Sign Ranked* test to evaluate the significance of this result. This test is chosen as we assess the difference in answers of matched samples. The null hypothesis states that the mean is the same. We aim to reject this hypothesis by showing that the means of the responses differ significantly. The resulting P values are all smaller than 0.01 and thus confirm that the responses for WachterCF are significantly different from the responses for CARE. From this, we can conclude that the counterfactual instances computed by CARE are considered more practical. Taking those results into account, we map back to the defined set of practicality properties.

- **Consistent with prior beliefs**: Question 2 assesses if the explanations are consistent with the participant prior beliefs by asking if the explanation is surprising to them. The results show that CARE provides less surprising explanations than WachterCF.
- **Contrastiveness**: Question 3 shows that suggestions by CARE provide a better explanation to the user of what needs to be changed to get the different model outcome, which serves as an estimation for being contrasting.
- **Selectivity**: Questions 5 & 6 aim to evaluate the selectivity by asking participants to assess the number of suggestions and variation of features. CARE is rated to select a better subset of features than WachterCF.
- **Social**: Question 7 evaluates a social perspective, showing that CARE serves more realistic suggestions considering the specific use case.
- **Truthful**: Question 8 shows that suggestions by CARE are perceived to make more sense than suggestions by WachterCF, which maps to being truthful.

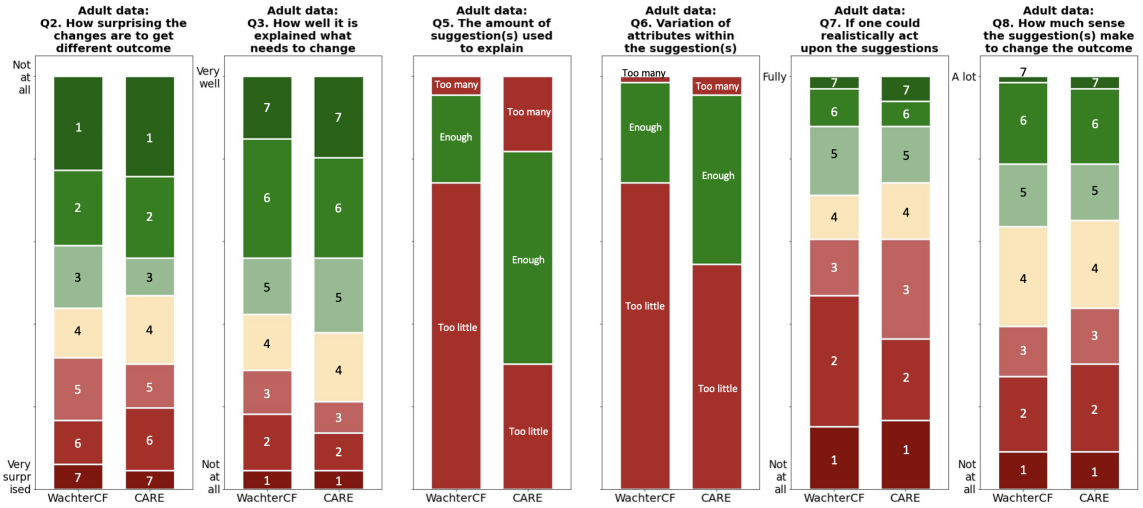

Figure 4: Responses to Questions with quantitative responses (Likert scale and Single Choice) exclusively for the classification task of the Adult Income dataset. Red-colored answers (bottom) indicate negative responses, yellow (middle) neutral and green positive (top). For each question the responses for WachterCF are compared with CARE. Based on the P values of a Wilcoxon Sign Ranked only Questions 5 & 6 are significantly different. The P value for each question shown is accordingly: 0.95, 0.2, 0.001, 0.01, 0.5, 0.8

6.1.2. Results of Questions with Feature Selection

Question 1 indicates the prior beliefs of the participants by asking what features they expect to change to retrieve the desired outcome before seeing any explanation. Question 4 elaborates on what features are considered important after seeing counterfactual explanations. We analyze the responses by computing a *percentage of agreement*, which shows how many of the features considered as most important after seeing the explanation were also selected in Question 1. For example, if Age and Working Hours are considered as the most important features, but only Age is chosen to be expected in Question 1, the *percentage of agreement* is 50%. By comparing the overall *percentage of agreement* of WachterCF and CARE, we can see that explanations provided by CARE seem to show a higher *percentage of agreements* than ones by WachterCF. However, according to the *Wilcoxon Sign Ranked* test we do not have enough evidence to show a significant difference for this comparison.

6.1.3. Subjective Comparison of the two methods according to participants

Question 9 directly asks the participants what method is preferred as an explanation. Out of the 135 answers, 113 selected CARE over WachterCF, and only 22 choose WachterCF as the preferred method. Therefore, we can conclude that humans subjectively prefer CARE over WachterCF.

6.2. Difference in Classification Task

The responses show that the perception of WachterCF and CARE differs depending on the classification task. To further assess whether this difference influences the obtained results, we perform the same analysis as in section 6.1 but split the responses according to the classification task assessed by the users. To test for significance between the responses to the different methods we use again a *Wilcoxon Sign Ranked* test. The Student Performance data shows a similar pattern as the overall results shown in Figure 3. CARE is seen as more practical compared to WachterCF and all differences are statistically significant. On the contrary, the responses to the Adult Income dataset do not show the same results. Figure 4 shows the different responses to the two counterfactual explanations for instances from the Adult Income dataset. Only Questions 5 & 6 show a significant difference in answers, which ask to indicate the perception regarding the practicality property *Selectivity*. All other questions do not differ significantly. Question 7 even shows a slight tendency in favor to WachterCF.

We propose three reasons for this. First, when looking at the actual explanations, WachterCF often only changes the attribute *Age* in both datasets. If so, *Age* decreases for the Student Performance dataset, but increases for the Adult Income dataset. This shows that a higher age leads to the outcome of earning more than 50k per year. Decreasing the age is not an actionable explanation, but getting older is happening without any active changes. Therefore, seeing this as an explanation gives satisfaction to the user. On the other hand, the explanations for the Student Performance dataset show that WachterCF also includes decreasing age as it does not include any preference settings, which leads to impracticality. Another reason might be that the student performance dataset contains more attributes to be modified than the adult income dataset. More attributes may lead to more complexity in computing practical counterfactual explanations. Another reason is that the decision of the Student Performance dataset is about whether a student is likely to pass or fail a class, which is a high-impact decision. In contrast, the purpose of the Adult Income is to decide whether a person is likely to earn more or less than 50k per year, which is not a critical decision. Therefore, participants may be more demanding regarding the practicality of the explanations when students' lives are directly affected.

6.3. Diversity in users' technical literacy

Lastly, we examine whether there is a difference in responses when comparing participants who have indicated to be familiar with machine learning models to those who are not. To test this, we divide the responses for each question into two groups indicating whether the user is familiar with machine learning models or not. To test if the differences are significant we use the *Mann-Whitney U* because the answers come from independent samples. Of the 15 questions (seven quantitative questions for each method and one comparison question), only Question 7 shown for the WachterCF explanation differs significantly ($p = 0.002$), which asks whether the suggestion can be realistically implemented. Participants with a technical background responded more positively than those with a non-technical background. One possible reason is that Question 7 offers a relatively large leeway for interpretation compared to other questions.

Participants with a technical background might better understand what we are trying to address with this question. In contrast, people with a non-technical background might interpret the question differently.

7. Discussion and Future Work

In this section, we discuss the limitations and weaknesses of the methodological design and reflect on the results of the user study transitioning to possible future work.

7.1. Limitations and Weaknesses

One limitation of the user study is that the user study participants were selected through convenience sampling [33]. This type of sampling is prone to bias: One possible bias in the target group is that participants are most likely to be highly educated. This could lead to a judgment that may not represent other social groups. Furthermore, the user study design does not represent a realistic scenario. Participants judge cases of people they do not know and to whom they have no emotional attachment. If the participants or people close to them would be affected by model outcomes, the expectation of explanations may be higher. Another aspect of this is that the datasets have been preprocessed before conducting the user study. This may not be the case in a real-life scenario; therefore, the quality of counterfactual explanations could suffer.

7.2. Impact of the scenario of an explanation

This paper shows that the responses to the majority of the questions differ depending on the scenario the counterfactual instance is used for. This could indicate that the perception of explanations differs according to the context of the ML decision. An explanation might be considered practical for a specific use case but impractical for a different one by the same user. This may indicate that the practicality of counterfactual explanation depends not only on the counterfactual method, but also on the type of ML task (e.g. the decision), the data used to decide, and the complexity of this decision (e.g. how many features are used to decide). Further research is necessary to determine what conclusions can be drawn about this.

7.3. Future Work

We gain insights that counterfactual explanations with the more recent method are overall perceived as more practical than with the original method, but also that responses differ depending on the use case of the decision. Future research should examine how the perception differs depending on different types of decisions and the number of features used. In addition, it is important to see how the results would have changed when users themselves are affected by the decision rather than evaluating an explanation for a stranger. Another area to explore further is how user perceptions of different explanation methods compare, such as feature attribution methods (like LIME [40]) or causal explanations [30].

8. Conclusions

People have a right to explanation when affected by a ML decision. This helps them to understand better why and how the model made this particular decision. Counterfactual explanations are a way to provide transparency by showing which attributes must change in what way to achieve a different outcome. Research has focused on developing various frameworks for computing counterfactuals. However, counterfactual explanations must be practical to be used by human subjects in practice. To answer our research question about the practicality of two counterfactual methods, we first define the following properties to measure practicality: Contrastiveness, Selectivity, Social, Truthful, and Consistent with prior beliefs of the user. To test how people perceive the explanations, we conduct a user study to compare CARE (a more recent multi-objective approach) against the original method WachterCF (as a baseline) for two different scenarios. The overall responses show that people perceive explanations computed with CARE as significantly more practical than those computed with WachterCF. Furthermore, we find our results differed between the Adult Income dataset and the Student Performance dataset, which indicates that the perception might differ depending on the use case of the explanation.

Acknowledgments

We are grateful to the participants of the user study. And we thank Emma Beauxis-Aussalet (Vrije Universiteit Amsterdam) for the thoughtful feedback which improved the findings of this paper.

Reproducability

We provide a repository containing the code and data that was used for this research. It also includes further details on the user study design. The repository is available at https://github.com/ninaspreitzer/practicality-counterfactual-explanations.

References

[1] B. Ustun, A. Spangher, Y. Liu, Actionable recourse in linear classification, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 10–19.

[2] A. E. Khandani, A. J. Kim, A. W. Lo, Consumer credit-risk models via machine-learning algorithms, Journal of Banking & Finance 34 (2010) 2767–2787.

[3] I. Ajunwa, S. Friedler, C. E. Scheidegger, S. Venkatasubramanian, Hiring by algorithm: predicting and preventing disparate impact, Available at SSRN (2016).

[4] L. Scism, New york insurers can evaluate your social media use-if they can prove why it's needed, The Wall Street Journal (2019).

[5] S. Herse, J. Vitale, M. Tonkin, D. Ebrahimian, S. Ojha, B. Johnston, W. Judge, M.-A. Williams, Do you trust me, blindly? factors influencing trust towards a robot recommender system, in: 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN), IEEE, 2018, pp. 7–14.

[6] 2018 reform of eu data protection rules, ???? URL: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

[7] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI magazine 38 (2017) 50–57.

[8] C. Molnar, Interpretable machine learning, Lulu. com, 2020.

[9] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[10] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the general data protection regulation, International Data Privacy Law 7 (2017) 76–99.

[11] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai). ieee access, 6: 52138–52160, 2018, 2018.

[12] K. Sokol, P. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 56–67.

[13] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, Artificial Intelligence 296 (2021) 103473.

[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[15] T. Laugel, M.-J. Lesot, C. Marsala, M. Detyniecki, Issues with post-hoc counterfactual explanations: a discussion, arXiv preprint arXiv:1906.04774 (2019).

[16] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[17] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, et al., Accountability of ai under the law: The role of explanation, arXiv preprint arXiv:1711.01134 (2017).

[18] A. Artelt, B. Hammer, On the computation of counterfactual explanations–a survey, arXiv preprint arXiv:1911.07749 (2019).
[19] P. Rasouli, I. C. Yu, Care: Coherent actionable recourse based on sound counterfactual explanations, arXiv preprint arXiv:2108.08197 (2021).
[20] A. Redelmeier, M. Jullum, K. Aas, A. Løland, Mcce: Monte carlo sampling of realistic counterfactual explanations, arXiv preprint arXiv:2111.09790 (2021).
[21] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, in: International Conference on Parallel Problem Solving from Nature, Springer, 2020, pp. 448–469.
[22] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.
[23] M. Downs, J. L. Chu, Y. Yacoby, F. Doshi-Velez, W. Pan, Cruds: Counterfactual recourse using disentangled subspaces, ICML WHI 2020 (2020) 1–23.
[24] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, J. Ghosh, Towards realistic individual recourse and actionable explanations in black-box decision making systems, arXiv preprint arXiv:1907.09615 (2019).
[25] M. Pawelczyk, K. Broelemann, G. Kasneci, Learning model-agnostic counterfactual explanations for tabular data, in: Proceedings of The Web Conference 2020, 2020, pp. 3126–3132.
[26] F. Yang, S. S. Alva, J. Chen, X. Hu, Model-based counterfactual synthesizer for interpretation, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1964–1974.
[27] S. Barocas, A. D. Selbst, M. Raghavan, The hidden assumptions behind counterfactual explanations and principal reasons, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 80–89.
[28] M. Förster, P. Hühn, M. Klier, K. Kluge, Capturing users' reality: A novel approach to generate coherent counterfactual explanations, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, p. 1274.
[29] K. Kanamori, T. Takagi, K. Kobayashi, H. Arimura, Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization., in: IJCAI, 2020, pp. 2855–2862.
[30] G. Warren, M. T. Keane, R. M. Byrne, Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in xai, arXiv preprint arXiv:2204.10152 (2022).
[31] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.
[32] R. F. Woolson, Wilcoxon signed-rank test, Wiley encyclopedia of clinical trials (2007) 1–3.
[33] I. Etikan, S. A. Musa, R. S. Alkassim, et al., Comparison of convenience sampling and purposive sampling, American journal of theoretical and applied statistics 5 (2016) 1–4.
[34] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.
[35] A. Frank, A. Asuncion, Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california, School of information and computer science 213 (2010).
[36] P. Cortez, A. M. G. Silva, Using data mining to predict secondary school student performance (2008).
[37] H. Zhu, Predicting earning potential using the adult dataset, Retrieved December 5 (2016)

2016.
[38] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.
[39] R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, Review of general psychology 2 (1998) 175–220.
[40] M. T. Ribeiro, S. Singh, C. Guestrin, " 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, 2016, pp. 1135–1144.

A. Overview of preprocessed Adult Income dataset

Column	Type	Values
Age	Continuous	17 - 90
Working Hours	Continuous	2 - 99
Gender	Discrete	Female Male
Race	Discrete	White Other[3]
Education Level	Discrete	Less than High School High School Graduate Some College Associate's Degree Bachelor's Degree Master's Degree Doctoral Degree Professional Degree
Marital Status	Discrete	Single Married Separated Divorced Widowed
Occupation	Discrete	Blue-Collar White-Collar Professional Sales Service Other/Unknown
Industry Type	Discrete	Government Private Self-Employed Other/Unknown

3. Please note that this racial distinction is taken from previous research and was chosen for simplicity. We are aware that it does not properly reflect all ethnicities.

B. Overview of preprocessed Student Performance dataset

Column	Type	Values
Age	Continuous	15 - 19
Absences	Continuous	0 - 30
Gender	Discrete	Female Male
Extra educational support	Discrete	Yes No
Family educational support	Discrete	Yes No
Paid tutor classes	Discrete	Yes No
Study Time	Discrete	Very low Low Medium High Very high
Freetime	Discrete	Very low Low Medium High Very high
Going out	Discrete	Very low Low Medium High Very high