"Explain it in the Same Way!" – Model-Agnostic Group Fairness of Counterfactual Explanations[*]

André Artelt[1,2] and Barbara Hammer[1]

[1] Bielefeld University, Bielefeld, Germany
[2] University of Cyprus, Nicosia, Cyprus

Abstract. Counterfactual explanations are a popular type of explanation for making the outcomes of a decision making system transparent to the user. Counterfactual explanations tell the user what to do in order to change the outcome of the system in a desirable way. However, it was recently discovered that the recommendations of what to do can differ significantly in their complexity between protected groups of individuals. Providing more difficult recommendations of actions to one group leads to a disadvantage of this group compared to other groups.
In this work we propose a model-agnostic method for computing counterfactual explanations that do not differ significantly in their complexity between protected groups.

Keywords: XAI · Fairness · Counterfactuals

1 Introduction

Transparency and fairness are essential building blocks of modern trustworthy decision making systems deployed in the real world [33,23]. Transparency is often realized by means of explanations – i.e. explanations of the system's behavior are provided to the user [24]. A lot of different explanation methodologies have been developed by researchers establishing the field of eXplainable Artificial Intelligence (XAI) [24,3]. Developed explanation methodologies include feature relevance/importance methods [15] and examples based methods [1]. Instances of popular example based methods are contrasting explanations like counterfactual explanations [36,34] and prototypes & criticisms [19] – these methods use a set or a single example for explaining the behavior of the system. The aspect of fairness [23] usually refers to the requirement that the output (i.e. predictions) of the systems must not discriminate single individuals or groups of individuals. Individual fairness states that similar individuals must be treated similarly [13] – i.e. the behavior or output of a fair system should not differ significantly between similar individuals. The key problem here is to come up with a suitable definition of similarity. On the other hand, group fairness states that the system

must not behave significantly different between individuals from different protected groups, which are created based on protected attributes such as gender, race, etc., – i.e. the behavior/output of the system must be independent of the protected attribute [26,23]. Here the challenges is to model independence of the protected attribute – different modelings give rise to different fairness criteria which contradict each other [26].

Both aspects, transparency/explanations and fairness, have been extensively studied in the past [2,29,3,16,26]. Only recently, the research community started to investigated the fairness of different explanation methodologies and discovered that fairness issues might exist there as well [9,35]. In particular, it was observed that provided explanations might differ significantly between similar individuals or between protected groups of individuals [17,31,7]. This becomes problematic especially in case of actionable recourse – i.e. explanations that tell the user what to do in order to observe a different outcome. A common way of achieving actionable recourse is to provide recommendations as explanations such as counterfactual explanations – as some recommendations are easier to full fill than others, significant differences in these recommendations across groups or similar individuals can therefore disadvantage certain individuals or groups of individuals.

Our contributions We propose a novel model-agnostic method for computing counterfactual explanations, a very popular instance of contrasting explanations, that satisfy a particular notion of group fairness – i.e. the counterfactual explanations should not differ significantly between different protected groups. We also extensively empirically evaluate the performance of our proposed method in several experiments.

The remainder of this work is structured as follows: In Section 2 we review counterfactual explanations and (re-)introduce the group fairness problem of counterfactual explanations with which we are dealing in this work. Next, in Section 3, we propose our methodology for computing group fair counterfactual explanations, which we evaluate empirically in Section 4. Finally, after discussing important characteristics of our proposed method and its broader impact in Section 5, the work closes with a conclusion and summary in Section 6.

2   Counterfactual Explanations

Counterfactual explanations (often just called *counterfactuals*) are an instance of contrasting explanations, stating changes to the input that lead to a different outcome (i.e. behavior of the system). Counterfactuals can be interpreted as a recommendation of actions that lead to a different specific behavior or output of the system. Since this mimics the way humans explain things [10] – i.e. humans often ask questions like *"What needs to be changed in order to observe a different outcome?"* –, counterfactuals became quite popular in the XAI community [24,34]. For illustrative purposes, consider the example of loan application: *Imagine you applied for a credit at a bank. Unfortunately, the bank rejects your*

application. Now, you would like to know why. In particular, you would like to know what would have to be different so that your application would have been accepted. A possible explanation might be that you would have been accepted if you had earned 500$ more per month and if you had not had a second credit card. In order to keep the explanations "simple" – i.e. easy to understand –, a common approach is to minimize the number of recommended actions by staying as close as possible to the original input. This is formalized in the following definition Definition 1.

Definition 1 ((Closest) Counterfactual Explanation [36]). *Assume a prediction function $h : \mathbb{R}^d \to \mathcal{Y}$ is given. Computing a counterfactual $\vec{x}_{cf} \in \mathbb{R}^d$ for a given input $\vec{x}_{orig} \in \mathbb{R}^d$ is phrased as the following optimization problem:*

$$\underset{\vec{x}_{cf} \in \mathbb{R}^d}{\arg\min} \; \ell\left(h(\vec{x}_{cf}), y_{cf}\right) + C \cdot \theta(\vec{x}_{cf}, \vec{x}_{orig}) \tag{1}$$

where $\ell(\cdot)$ denotes a loss function that penalizes deviation of the prediction $h(\vec{x}_{cf})$ from the requested target prediction y_{cf}, $\theta(\cdot)$ denotes a penalty for dissimilarity of $\vec{x}_{cf}$ and $\vec{x}_{orig}$[‡], and $C > 0$ denotes the regularization strength.

Note that the actual explanation – i.e. recommendation of actions to change some features in a particular way – is given by the difference $\delta = \vec{x}_{cf} - \vec{x}_{orig}$. Furthermore, $\theta(\cdot)$ can also be interpreted as the complexity of the counterfactual $\vec{x}_{cf}$ – i.e. measuring how difficult it is to execute the recommended changes δ, since changing some features might be easier or more difficult than changing others, and also the amount of change might contribute to the complexity & difficultiness of the explanation. For illustrative purposes, we again consider the previous example of a loan application: *While a valid explanation might be to increase the monthly income by 500$, another valid explanation might be to increase monthly income by 400$, canceling a second credit card and having some property as a collateral security. Clearly, the latter one is much more difficult to full fill[§] than the first one – therefore, one would assign a higher score $\theta(\cdot)$ to the latter.*

In the remainder of this work, we assume a binary classification problem: In this case, we denote a (closest) counterfactual $\vec{x}_{cf}$ (according to Definition 1) of a given sample $\vec{x}_{orig}$ under a prediction function $h(\cdot)$ simply as $\vec{x}_{cf} = \mathrm{CF}(\vec{x}_{orig}, h)$ and drop the target label y_{cf} because it is uniquely determined.

The counterfactuals $\vec{x}_{cf}$ from Definition 1 are also called *closest counterfactuals* [18] because they try stay as close as possible to the original input $\vec{x}_{orig}$. However, closest counterfactual are not necessarily plausible & actionable which might hinder they use in practice [37] – there exist numerous extensions [5,22,6,32,28], that try to improve plausibility & actionability of counterfactuals from Definition 1. The most simple approach for generating plausible,

---

[‡]This acts as a regularization of the complexity of the final explanation – i.e. prefer "simple & low complexity" explanations.

[§]Note that the notion of difficultiness and complexity of a given counterfactual highly depends on the domain and use-case – in this work we use the 1-norm as a general proxy of the explanations complexity.

and therefore hopefully actionable, counterfactuals is to limit the feasible set of solutions to Eq. (1) to a set of training samples [28] – i.e. simply picking the most similar sample from the training set as a counterfactual explanation. While this option is easily implementable and guarantees plausible counterfactuals, it comes however at the price of potentially leaking sensitive information and might therefore pose a privacy issue [25]. Nevertheless, this is the method we use whenever we refer to plausible counterfactuals.

2.1 Fairness of Counterfactual Explanations

In this work, we assume that the individuals are grouped into two so called protected groups $\mathcal{D}_1$ and $\mathcal{D}_2$ based on some sensitive attribute like sex, race, etc. While fairness of the predictions of AI systems is a well known challenge that has been studied for quite some time [23,27,11], potential fairness issues of explanations were only recently discovered by the community [35,7,31,17]. In particular, it was observed that closest counterfactual explanations (Definition 1) might differ significantly between individuals from different groups [17] or also between similar individuals [7].

In this work, we follow the common fairness notion of counterfactuals as introduced in [17], that is based on the complexity & difficultiness of the explanation, which is measured by the distance $\theta(\cdot)$ between the original sample $\vec{x}_{\mathrm{orig}}$ and the corresponding counterfactual explanation $\vec{x}_{\mathrm{cf}}$ – we refer to this quantity as the *cost* of a counterfactual:

Definition 2 (Cost of a Counterfactual Explanation). *For a given sample $\vec{x}_{orig} \in \mathbb{R}^d$ and a classifier $h(\cdot)$, we define the cost of the corresponding closest counterfactual Definition 1 $\vec{x}_{cf}$ as the distance between the original sample and the counterfactual:*

$$\theta(\vec{x}_{orig}, \mathit{CF}(\vec{x}_{orig}, h)) \tag{2}$$

where we substituted $\vec{x}_{cf}$ with $\mathit{CF}(\vec{x}_{orig}, h)$ to stress the dependencies on $\vec{x}_{orig}$, $h(\cdot)$ and the explanation generation method $\mathit{CF}(\cdot, \cdot)$ itself.

Note that in Definition 2 we completely rely on $\theta(\cdot)$ for measuring the complexity and costs of the recommended actions by the counterfactual explanation – note that in practice, $\theta(\cdot)$ must be chosen carefully and might require some domain knowledge.

It can be perceived as problematic if the cost of the counterfactuals (i.e. recommended actions) differs significantly between protected groups of individuals (e.g. males vs. females) [17,9,31] or between similar individuals [7], since achieving the algorithmic recourse would be much easier for the individuals getting low-cost explanations than those getting high-cost (i.e. more difficult to execute) explanations – this would disadvantage certain groups of individuals which violates social norms of fairness [14,21]. For illustrative purposes, we again come back to our running example of loan applications: *For instance, it can be considered as unfair if the counterfactual explanations presented to males on average ask only for a minor increase in their monthly income while those presented to*

females ask for a major increase in their monthly income. We formalize this intuition of group fairness of counterfactual explanations by using the definition of a cost (Definition 2) of a counterfactual:

Definition 3 (Group Fair Counterfactual Explanation). *We say that counterfactual explanations are fair iff their costs (Definition 2) does not differ significantly between the different groups [17,31] – i.e. the distributions of $\theta(\cdot)$ are "sufficiently similar":*

$$p_{\vec{x}_{orig} \sim \mathcal{D}_1}\Big(\theta(\vec{x}_{orig}, CF(\vec{x}_{orig}, h))\Big) \approx p_{\vec{x}_{orig} \sim \mathcal{D}_2}\Big(\theta(\vec{x}_{orig}, CF(\vec{x}_{orig}, h))\Big) \quad (3)$$

Since dealing with the entire distribution $p_{\vec{x}_{orig} \sim \mathcal{D}_?}$ might be too complicated, we relax and reduce Eq. (3) to the expected value [31]:

$$\mathop{\mathbb{E}}_{\vec{x}_{orig} \sim \mathcal{D}_1}\big[\theta(\vec{x}_{orig}, \mathrm{CF}(\vec{x}_{orig}, h))\big] \approx \mathop{\mathbb{E}}_{\vec{x}_{orig} \sim \mathcal{D}_2}\big[\theta(\vec{x}_{orig}, \mathrm{CF}(\vec{x}_{orig}, h))\big] \quad (4)$$

which can be easily estimated using given sets of samples.

Existing work from literature In order to achieve group fairness of counterfactual explanations – i.e. satisfying Eq. (3) – in case of a support vector machine (SVM), the authors of [17] propose to change the training of the SVM such that the counterfactual explanations satisfy the group fairness criterion Eq. (4). More specifically, they propose an iterative method where they add recourse constraints to the original SVM optimization problem. In addition, they also propose a model-agnostic method where they iteratively generate samples using LIME [30] that are used to retrain the original classifier $h(\cdot)$ using these samples.

However, to the best of our knowledge, all existing methods either focus on specific models and/or do not consider changing the model-agnostic algorithm $\mathrm{CF}(\cdot, \cdot)$ for computing the final counterfactuals that satisfy Eq. (3). In this work, we aim to address this research gap by proposing a model-agnostic algorithm $\mathrm{CF}(\cdot, \cdot)$ for computing group fair counterfactual explanations (see Section 3) without retraining or changing the given model $h(\cdot)$.

3 Computing Group Fair Counterfactual Explanations

In order to achieve the relaxed group fairness of counterfactual explanations (Definition 3) as stated in Eq. (4), we propose to change the method $\mathrm{CF}(\cdot, \cdot)$ for computing counterfactuals. More specifically, we propose to extend Eq. (1) as follows:

$$\min_{\vec{x}_{\mathrm{cf}} \in \mathbb{R}^d} \theta(\vec{x}_{\mathrm{orig}}, \vec{x}_{\mathrm{cf}}) + C_0 \cdot \ell\left(h(\vec{x}_{\mathrm{cf}}), y_{\mathrm{cf}}\right) + C_1 \cdot \max\left(z - \theta(\vec{x}_{\mathrm{orig}}, \vec{x}_{\mathrm{cf}})\right)$$
$$\text{where } z \sim p(\theta(\vec{X}_{\mathrm{orig}}, \vec{X}_{\mathrm{cf}})) \tag{5}$$

where the only difference to Eq. (1) is the addition of the last regularization term $C_1 \cdot \max\left(z - \theta(\vec{x}_{\mathrm{orig}}, \vec{x}_{\mathrm{cf}})\right)$ in Eq. (5) which adds a penalty if the distance (i.e.

Algorithm 1 Computing Group Fair Counterfactual Explanations

Input: A classifier $h(\cdot)$; Samples grouped into two groups $\mathcal{D}_1$ and $\mathcal{D}_2$ based on their binary sensitive attribute; Regularization strengths $C_0, C_1 > 0$
Output: Group fair counterfactuals $\vec{x}_{\mathrm{cf}}$

1:   ▷ **Initialization step**
2:      ▷ Compute the cost of counterfactual explanation for all samples in both sets
3: $\mathcal{Z}_1 = \{\theta(\vec{x}_{\mathrm{orig}}, \mathrm{CF}(\vec{x}_{\mathrm{orig}}, h)) \quad \forall \vec{x}_{\mathrm{orig}} \in \mathcal{D}_1\}$
4: $\mathcal{Z}_2 = \{\theta(\vec{x}_{\mathrm{orig}}, \mathrm{CF}(\vec{x}_{\mathrm{orig}}, h)) \quad \forall \vec{x}_{\mathrm{orig}} \in \mathcal{D}_2\}$
5:          ▷ Check which of the two groups is disadvantaged compared to the other
6: $\mathcal{Z} = \emptyset$
7: **if** $\frac{1}{|\mathcal{Z}_1|} \sum_{z_i \in \mathcal{Z}_1} z_i > \frac{1}{|\mathcal{Z}_2|} \sum_{z_j \in \mathcal{Z}_2} z_j$ **then**
8:      $\mathcal{Z} = \mathcal{Z}_1$
9: **else**
10:      $\mathcal{Z} = \mathcal{Z}_2$
11: **end if**
12:
13:   ▷ **Main step**
14:      ▷ For all new samples $(\vec{x}_{\mathrm{orig}}, y_{\mathrm{orig}})$ we compute a group fair counterfactual as follows:
15: $z = \mathrm{RANDOM.CHOICE}(\mathcal{Z})$ ▷ Select a random cost from the disadvantaged group
16:                          ▷ Compute the group fair counterfactual by solving Eq. (5)
17: $\vec{x}_{\mathrm{cf}} = \underset{\vec{x}_{\mathrm{cf}} \in \mathbb{R}^d}{\arg\min}\ \theta(\vec{x}_{\mathrm{orig}}, \vec{x}_{\mathrm{cf}}) + C_0 \cdot \ell\left(h(\vec{x}_{\mathrm{cf}}), y_{\mathrm{cf}}\right) + C_1 \cdot \max\left(z - \theta(\vec{x}_{\mathrm{orig}}, \vec{x}_{\mathrm{cf}})\right)$

---

cost/complexity) of the counterfactual $\vec{x}_{\mathrm{cf}}$ is larger than some particular value z sampled from the distribution of distances (i.e. costs) $\theta(\cdot)$ of the *disadvantaged group* – by this we make sure that the difference between the advantaged and disadvantage group disappears on average. Note that for this, we need access to a set of samples (including their counterfactual explanation) from the disadvantaged group, so that we can estimate $\mathbb{E}\left[\theta(\vec{X}_{\mathrm{orig}}, \vec{X}_{\mathrm{cf}})\right]$ since we do not know the true distribution. For practical purposes, we suggest to simply select the distance (i.e. cost) of a random sample from the disadvantaged group as a value of z – by this we obtain a randomized algorithm/method for computing counterfactual explanations and the counterfactuals are no longer deterministic. After we obtained a value of z, we can solve Eq. (5) either by using any black-box solver, like for instance Downhill-Simplex, resulting in a closest counterfactual or we could obtain a plausible counterfactual by simply select a sample from the training set $\mathcal{D}_1 \cup \mathcal{D}_2$ that minimizes Eq. (5). The entire proposed methodology for computing group fair counterfactual explanations is described in full detail in Algorithm 1.

Note that the major advantage of this approach, compared to existing methods [17], is that the model $h(\cdot)$ remains untouched – i.e. we do not assume that we have access to model internals, or more crucially, that we are able to retrain the entire model which might be quite costly in practice – which perfectly aligns with the model-agnostic paradigm of counterfactual explanations [36].

4   Experiments

We empirically evaluate our proposed method (Algorithm 1) for generating group
fair counterfactual explanations on a diverse set of data sets and classifiers. We
also consider and compare our proposed method (Algorithm 1) for two different
types of counterfactuals: Closest counterfactuals (Definition 1) and plausible
counterfactuals (i.e. samples from the training set). The Python implementation,
including all necessary details for reproduction, of the experiments is publicly
available on GitHub[¶].

4.1   Data

Credit card clients data set The Credit card clients data set [39] (*Credit*) contains
30000 data records of customers used for default payment prediction. Each record
is described by 23 attributes (8 categorical, 14 numerical and 1 binary). The
sensitive attribute is "sex" but we also remove "age" and "marriage".

Communities & Crime data set The Communities & Crime data set [12] (*Crime*)
contains 1994 socio-economic data records from 46 states in the USA. Following
the preprocessing as suggested in [20], we are left with 100 encoded attributes
and that are used to predict the crime rate (low crime rate vs. high crime rate).
The sensitive attribute is "race".

Law school data set The Law school data set [38] (*Law*) contains 20798 law
school admission records. Each record (student) is described by 12 attributes
(3 categorical, 3 binary and 6 numerical). The binary target attribute describes
whether the student as admitted or not. The sensitive attribute is "sex" but we
also remove the attribute "race".

4.2   Setup

We run the following procedure, for every data set and classifier – we consider
logistic regression (*Logreg*), decision trees (*Dectree*) and Gaussian naive Bayes
(*GNB*) –, in a 3-fold cross validation:

1. Build the classifier using all available training data – i.e. ignore any group
   assignments based on sensitive attributes – but sensitive attributes are NOT
   included as features.
2. Compute two counterfactual explanations for each sample of all test samples
   that are classified correctly – the closest counterfactual (Definition 1) is
   computed by using [8] and for a plausible counterfactual we simply select
   the sample from the training set that minimizes Eq. (1). We then group
   these samples and their counterfactuals according to the sensitive attribute.

[¶]https://github.com/HammerLabML/ModelAgnosticGroupFairnessCounterfactuals

3. Use Algorithm 1 to compute group fair counterfactuals of all correctly classified samples. Again, we solve Eq. (5) in two different ways: A closest counterfactual by solving Eq. (5) using the Downhill-Simplex method, and a plausible counterfactual by limiting the set of feasible solutions of Eq. (5) to the training set.

For the purpose of evaluating our proposed method Algorithm 1, we compare the costs per group of the closest counterfactuals (Definition 1) with those of the group fair counterfactuals (Definition 3).

4.3   Results

Table 1. Results for closest counterfactuals – we report the median cost of counterfactuals (Definition 2), all number are rounded to two decimal points.

Clf	Data set	Group-0	Group-1	Group-0-Fair	Group-1-Fair	Diff	Diff-Fair
Logreg	Credit	0.96	0.97	1.62	1.65	0.01	0.03
	Crime	1.75	1.2	2.06	1.91	0.55	0.15
	Law	1.8	1.81	2.6	2.63	0.01	0.02
Dectree	Credit	0.23	0.25	0.43	0.43	0.02	0.0
	Crime	1.52	1.06	1.61	1.52	0.46	0.09
	Law	0.46	0.45	0.76	0.75	0.01	0.01
GNB	Credit	6.71	6.39	10.26	10.01	0.32	0.24
	Crime	7.06	5.47	7.21	8.16	1.59	0.95
	Law	5.71	5.73	6.93	6.95	0.02	0.02

Table 2. Results for plausible counterfactuals – we report the median cost of counterfactuals (Definition 2), all number are rounded to two decimal points.

Clf	Data set	Group-0	Group-1	Group-0-Fair	Group-1-Fair	Diff	Diff-Fair
Logreg	Credit	3.81	3.65	10.33	10.41	0.17	0.08
	Crime	62.1	55.34	61.86	61.62	6.76	0.24
	Law	4.25	4.31	5.30	5.33	0.06	0.03
Dectree	Credit	3.03	2.96	10.13	10.15	0.07	0.02
	Crime	63.04	53.92	61.73	62.32	9.12	0.59
	Law	2.89	2.97	6.06	6.41	0.08	0.35
GNB	Credit	4.47	4.18	9.87	10.14	0.28	0.26
	Crime	74.42	60.36	58.88	61.06	14.05	2.18
	Law	5.11	5.16	4.63	4.63	0.05	0.01

A summary of results for the closest counterfactuals are shown in Table 1 and the results for plausible counterfactuals are shown in Table 2. We observe
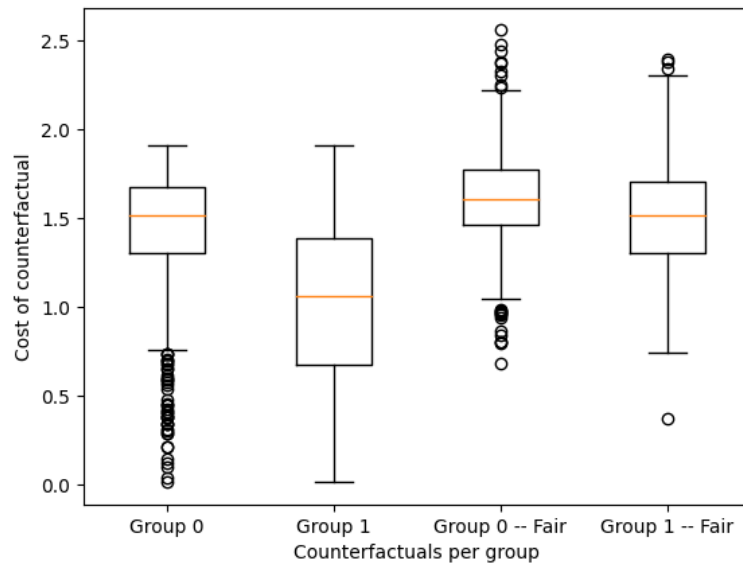
Fig. 1. Normal vs. group fairness constrained counterfactuals: Group wise cost of closest counterfactuals for a decision tree classifier on the Communities & Crime data set.

two things: First, not all scenarios show group fairness issues – i.e. in some cases there is no significant difference in the costs across the different groups. Second, if there is a significant difference, our method almost always improves the group fairness significantly – i.e. the difference in the costs across different groups is much smaller, which we additionally illustrate in Figure 1 for the particular scenario of closest counterfactuals of a decision tree classifier on the CommunitiesCrime data set. We do not observer any difference between closest and plausible counterfactuals. These findings demonstrate the strong performance of our proposed model-agnostic method for computing group fair counterfactual explanations.

5   Discussion & Broader Impact

Advantages. The key advantage of our proposed method over existing work is that it is completely model-agnostic – i.e. we do not assume access to any model internals or that we are able to retrain the model.

   Computational complexity. Before we can compute group fair counterfactuals, we first have to compute counterfactual explanations of all samples from the training set that is used to estimate the distribution of costs. The computational complexity of this step mainly depends on the number of samples and the classifier $h(\cdot)$, since the computational complexity of computing a counterfactual varies significantly between different classifiers and how many assumptions we can make on the classifier [4]. However, this initial step has to be done only once and does not need to be repeated if the data distribution and model do not change. The computation of the group fair counterfactual is then done by solving Eq. (5) using a black-box method like Downhill-Simplex method.

   Limitations. There exist two potential limitations of our proposed method: First, the initialization step of Algorithm 1 assumes access to a sufficiently large set of training samples that can be used to estimate the costs of counterfactuals across the two protected groups. In practice, it might be challenging to obtain enough samples from a minority group in order to reliable estimate the cost of the counterfactuals. Furthermore, the computation of these counterfactuals induces some computational costs as discussed above. Second, the advantage of being model-agnostic, and not considering any special cases where we might have access to model internals, comes at the price that we can not exploit any structure of $h(\cdot)$ for efficiently computing a group fair counterfactual by solving Eq. (5). As a consequence, we have to rely on computationally expensive methods for solving Eq. (5). Note that this is in contrast to the initialization step of Algorithm 1, where we have to compute normal counterfactuals and can exploit any structure or access to model internals if those are available.

   Broader impact. Counterfactual explanations are widely used in practice to provide users with actionable feedback how to change the outcome of a decision making system. Potential fairness issues, as described in this work, are therefore harmful to individuals and must be eliminated as much as possible. We therefore expect that our proposed method have a high impact in practice

as it can significantly improve the fairness of counterfactual explanations. Because it is completely model-agnostic, it can be applied to any models used by practitioners, and our publicly available Python implementation of Algorithm 1 is easy to use and therefore allows a direct and seamless integration into existing XAI pipelines.

6   Summary & Conclusion

In this paper, we looked into the issue of group fairness of counterfactual explanations, and proposed a model-agnostic method for computing group fair counterfactuals. We empirically evaluated our proposed algorithm on a diverse set of experiments and observed a strong performance. The key advantage over existing methods from literature is that our method does not need access to the model or have to retrain the model – i.e. our method is completely model-agnostic and can be applied to any model that is going to be explained by counterfactuals.

Based on this work, there emerge a couple of potential directions for future research:

– In this work we ignored any potential access to model internals or structure of $h(\cdot)$ for computing group fair counterfactuals. However, as it was shown in [4], exploiting model internals can lead to much more efficient algorithms for computing normal closest counterfactuals (Definition 1). In this context it would interesting and relevant to study how such available knowledge could be exploited for computing group fair counterfactuals by solving Eq. (5).
– As discussed, the initialization step of Algorithm 1 has to be recomputed if the data distribution or $h(\cdot)$ changes. Since the initialization step can be computationally expensive, it would be relevant to study if and how an adaptation of this initialization step is possible in case of any changes in the data distribution or the model $h(\cdot)$.

References

1. Aamodt, A., Plaza., E.: Case-based reasoning: Foundational issues, methodological variations, and systemapproaches. AI communications (1994)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
3. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
4. Artelt, A., Hammer, B.: On the computation of counterfactual explanations - A survey. CoRR **abs/1911.07749** (2019)
5. Artelt, A., Hammer, B.: Convex density constraints for computing plausible counterfactual explanations. 29th International Conference on Artificial Neural Networks (ICANN) (2020)
6. Artelt, A., Hammer, B.: Convex optimization for actionable\& plausible counterfactual explanations. arXiv preprint arXiv:2105.07630 (2021)

7. Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., Hammer, B.: Evaluating robustness of counterfactual explanations. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 01–09. IEEE (2021)
8. Artelt, A.: Ceml: Counterfactuals for explaining machine learning models - a python toolbox. https://www.github.com/andreArtelt/ceml (2019 - 2022)
9. Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., Ghassemi, M.: The road to explainability is paved with bias: Measuring the fairness of explanations. arXiv preprint arXiv:2205.03295 (2022)
10. Byrne, R.M.J.: Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In: IJCAI-19 (2019)
11. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018)
12. Dheeru, D., Taniskidou, E.K.: Uci machine learning repository (2017)
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Goldwasser, S. (ed.) Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012. pp. 214–226. ACM (2012). https://doi.org/10.1145/2090236.2090255, https://doi.org/10.1145/2090236.2090255
14. Fehr, E., Schmidt, K.M.: A theory of fairness, competition, and cooperation. The quarterly journal of economics **114**(3), 817–868 (1999)
15. Fisher, A., Rudin, C., Dominici, F.: All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. arXiv e-prints arXiv:1801.01489 (Jan 2018)
16. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. p. 329–338. FAT* '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287589, https://doi.org/10.1145/3287560.3287589
17. Gupta, V., Nokhiz, P., Roy, C.D., Venkatasubramanian, S.: Equalizing recourse across groups. arXiv preprint arXiv:1909.03166 (2019)
18. Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I.: A survey of algorithmic recourse: contrastive explanations and consequential recommendations. ACM Computing Surveys (CSUR) (2021)
19. Kim, B., Koyejo, O., Khanna, R.: Examples are not enough, learn to criticize! criticism for interpretability. In: Advances in Neural Information Processing Systems 29 (2016)
20. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery p. e1452 (2022)
21. Loewenstein, G.F., Thompson, L., Bazerman, M.H.: Social utility and decision making in interpersonal contexts. Journal of Personality and Social psychology **57**(3), 426 (1989)
22. Looveren, A.V., Klaise, J.: Interpretable counterfactual explanations guided by prototypes pp. 650–665 (2021)
23. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
24. Molnar, C.: Interpretable Machine Learning (2019)
25. Pawelczyk, M., Lakkaraju, H., Neel, S.: On the privacy risks of algorithmic recourse. arXiv preprint arXiv:2211.05427 (2022)

26. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Comput. Surv. **55**(3) (feb 2022). https://doi.org/10.1145/3494672, `https://doi.org/10.1145/3494672`
27. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Computing Surveys (CSUR) **55**(3), 1–44 (2022)
28. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P.: Face: feasible and actionable counterfactual explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 344–350 (2020)
29. Rawal, A., Mccoy, J., Rawat, D.B., Sadler, B., Amant, R.: Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives. IEEE Transactions on Artificial Intelligence **1**(01), 1–1 (2021)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386 (2016)
31. Slack, D., Hilgard, A., Lakkaraju, H., Singh, S.: Counterfactual explanations can be manipulated. Advances in Neural Information Processing Systems **34**, 62–75 (2021)
32. Smyth, B., Keane, M.T.: A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In: International Conference on Case-Based Reasoning. pp. 18–32. Springer (2022)
33. Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., Van Moorsel, A.: The relationship between trust in ai and trustworthy machine learning technologies. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 272–283 (2020)
34. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review (2020)
35. Von Kügelgen, J., Karimi, A.H., Bhatt, U., Valera, I., Weller, A., Schölkopf, B.: On the fairness of causal algorithmic recourse. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 9584–9594 (2022)
36. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
37. Warren, G., Smyth, B., Keane, M.T.: "better" counterfactuals, ones people can understand: Psychologically-plausible case-based counterfactuals using categorical features for explainable ai (xai). In: International Conference on Case-Based Reasoning. pp. 63–78. Springer (2022)
38. Wightman, L.F.: Lsac national longitudinal bar passage study. lsac research report series. (1998)
39. Yeh, I.C., Lien, C.h.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert systems with applications **36**(2), 2473–2480 (2009)