

# Discrimination Induced by Algorithmic Recourse Objectives

Nicholas Perello

University of Massachusetts Amherst  
Amherst, Massachusetts, USA  
nperello@umass.edu

Yair Zick\*

University of Massachusetts Amherst  
Amherst, Massachusetts, USA  
yzick@umass.edu

Cyrus Cousins

Duke University  
Durham, North Carolina, USA  
originalcyruscousins@gmail.com

Przemyslaw Grabowicz\*

University College Dublin  
Dublin, Ireland  
przemek.grabowicz@ucd.ie

## Abstract

We study fairness in algorithmic recourse, which provides users who receive undesirable outcomes from decision-making systems with explanations on how to achieve their desired results. To produce explanations that emulate real-world settings and preferences, the literature proposes various constraints, e.g., sparsity, on the objective of recourse methods. We argue that these constrained objectives can induce disparate recourse costs between protected groups, especially when groups' recourse objectives differ from the objectives of the methods providing them recourse. We formalize discrimination in algorithmic recourse with respect to legal notions and the cost of recourse. Namely, we introduce a novel notion of induced discrimination that captures the excess cost of recourse induced by having misaligned objectives between groups and methods. Using this definition, we show that recourse objectives can induce discriminatory recourse, independent from the bias of the model being explained. To inhibit discrimination, we extend prior work on welfare-based group-fair supervised learning to measure group welfare under multiple recourse objectives. We then propose a recourse method that maximizes group welfare and empirically show its robust performance over multiple recourse objectives.

## CCS Concepts

• **Computing methodologies** → *Machine learning*; • **General and reference** → *Evaluation*.

## Keywords

algorithmic recourse, counterfactual explanations, algorithmic fairness, explainable AI

## ACM Reference Format:

Nicholas Perello, Cyrus Cousins, Yair Zick, and Przemyslaw Grabowicz. 2025. Discrimination Induced by Algorithmic Recourse Objectives. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, June 23–26, 2025, Athens, Greece. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3715275.3732110>

\*Equal Advising



This work is licensed under a Creative Commons Attribution 4.0 International License. *FAccT '25, Athens, Greece*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1482-5/25/06

<https://doi.org/10.1145/3715275.3732110>

## 1 Introduction

The rise of complex automated decision-making systems has led to an increasing need to ensure that these often black-box systems offer fair and explainable outcomes. That is, automated decision makers should not bias their decisions based on extraneous features, such as race or gender. Recent years have seen a significant research focus on the design of non-discriminatory ML decision-making systems. Algorithmic explainability constitutes a significant part of this effort. However, this raises an important issue: since explanation frameworks are products of the underlying model and data, they may carry over potential biases into their own internal design. This work examines whether state-of-the-art recourse methods treat protected groups fairly.

Recent work suggests that popular feature highlighting explainability methods, such as SHAP [19], are biased with respect to an individual's group membership [1]. Unlike feature highlighting methods, algorithmic recourse offers suggestions — typically via counterfactual explanations — on how a data point (say, the profile of a job applicant) should change its feature values to attain a desired outcome (a job offer). Counterfactual explanations typically search for the minimal changes required to attain the desired outcome, i.e., a *proximal counterfactual*, and the effort needed for such changes is the cost of following the counterfactual explanation. Recent works argue that proximity is insufficient for providing actionable and interpretable recourse [14, 18, 21, 23]. Therefore, state-of-the-art recourse methods constrain their objectives to account for other recourse desiderata [15]. For instance, *sparsity* requires that recourse suggest a minimal number of changes to features and *plausibility* requires recourse to offer realistic interventions. Enforcing additional objectives results in algorithmic recourse that does not necessarily minimize distance. We show that the additional cost of recourse induced by these objectives is disparate between protected groups, especially when the underlying data is biased. In other words, we show that the design of effective recourse methods presents significant challenges to algorithmic fairness.

**Example 1.1.** Consider the senior STEM job application scenario in Fig. 1. A decision-making system determines whether to hire someone based on their education level and years of work experience. Due to discrimination in postsecondary education and jobs in STEM, accepted female applicants are a relatively homogeneous highly-educated and experienced group, whereas accepted male applicants of various education and work experience levels are well-represented in the dataset. These characteristics lead to the

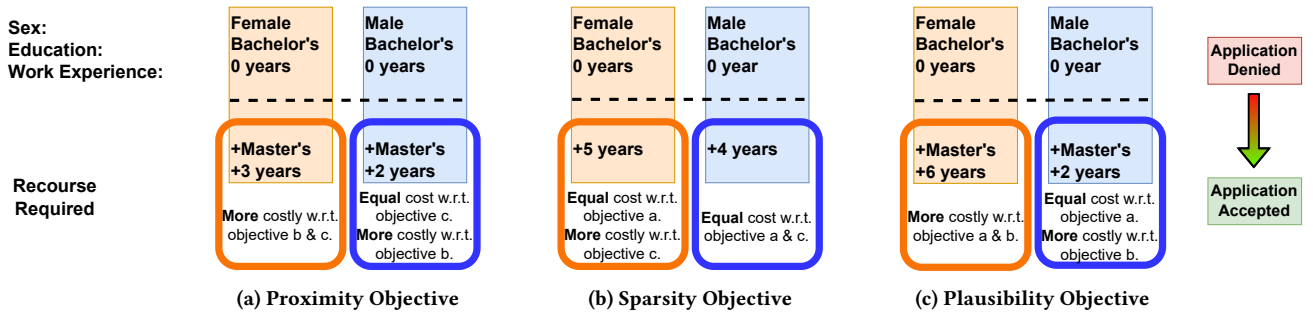


**Figure 1: A hypothetical job application scenario with men and women. Two applicants receive recourse from three recourse methods with different objectives. The larger sample of men and historical bias against women leads to more counterfactuals for men which can satisfy multiple recourse objectives. We assume a Master's is equivalent to 2 years of work.**

male applicant having more recourse options that potentially share a common counterfactual between all three recourse objectives. However, the female applicant's recourse options differ across all three objectives, putting them at risk to receive costlier recourse when compared across objectives. For instance, say the female applicant was given recourse through a method that aims to provide *plausible* counterfactuals, i.e., ones that are likely to be followed. Due to historical bias, such counterfactuals are sparse, thus this method rejects the closest  $\ell_2$  recourse and requires the applicant to be more similar to other female applicants by making them work longer (Fig. 1c). However, the applicant could reasonably have preferred the recourse under the proximity objective, which requires three fewer years (Fig. 1a). Conversely, the male applicant is offered a much closer point, since the closest recourse is already plausible.

While our example is clearly contrived, it is based on a realistic use-case: automated decision makers — and consequently algorithmic recourse — are used for job applications [2, 15]. With this in mind, practitioners deploying decision-making systems should ensure not only that their learning methods are fair, but that any recourse they provide induces minimal discrimination. Moreover, since groups may have varying preferences for algorithmic recourse objectives, recourse methods should account for differences in group *objectives* and *preferences*. State-of-the-art recourse methods implicitly assume that group preferences and societal objectives are aligned. However, user studies challenge this assumption; for instance, Kuhl et al. [16] show that users benefit the most from (and are likely to prefer) proximity objectives over plausibility objectives. Additionally, these objectives are not universally agreed upon as recourse desiderata among researchers [18]. Therefore, it is reasonable to expect that some objectives will disadvantage sensitive groups, resulting in recourse that is worse than what they would have gotten under their preferred objective, as seen in Example 1.1. Ideally, we know each groups' recourse objective. However, often this is not realistic, so we instead assume that each objective comes from a known set of objectives that groups can choose from. Then, a fair recourse method should perform well over each group's costs across this set of objectives.

Prior works on fairness and recourse are scarce; most of the works that do acknowledge this intersection focus on creating

supervised learning methods that result in fair recourse costs according to proposed definitions of fairness [11, 17] or observe that their introduce methods produced differing cost w.r.t. Euclidean distance between groups [14, 24]. While these aspects are important, these works do not investigate the different recourse methods themselves and their objectives.

## 1.1 Our Contributions

We analyze recourse methods' objectives with respect to discrimination. We define notions of direct and indirect discrimination in the context of recourse.

We introduce a new measure of induced discrimination in recourse. This measure captures the disparity in “extra cost” across recourse recipients caused by misaligned recourse objectives between groups and methods (§3). We analyze Equalized Recourse, a fair recourse metric that aims to equalize the cost of recourse between groups [11]; in particular, recourse can induce discrimination, even when it performs well according to Equalized Recourse (§3.1).

As a final fairness metric, we evaluate the welfare (or disutility/malfare) of different groups under several recourse methods (§4). We propose a objective-robust recourse method that minimize malfare objectives and, consequently, best jointly satisfies the set of possible group objectives. These notions also allow for unequal improvements in recourse cost across groups, which is not allowed by inequality measures. Analysis of real-world datasets shows how malfare minimizing recourse methods select counterfactuals that balances disparity and overall recourse costs (§5).

## 1.2 Related Work

Fairness in ML literature typically seeks to ensure that ML systems do not treat someone unfavorably because of a protected feature, such as race or gender. To create achievable fairness objectives, definitions of discrimination are often based on statistical independence or on causal relations from the protected feature(s) [2]. As noted by Grabowicz et al. [10], the aforementioned definitions of fairness constraints can conflict with legal systems where decision-making systems may be exempt from these constraints by providing legally-admissible justifications of disparate impact. To better tie fairness in ML and legal texts together the authors propose influence-based

definitions of fairness, where influence can be causal influence, model input influence, or a feature relevance measure. We can tie these influence-based notions to our definitions of discriminatory recourse.

We focus on the recourse provided by counterfactual explanations. Wachter et al. [28] introduce counterfactual explanations as a tool for building accountability and transparency between data controllers and subjects beyond their claimed limitations of the “EU General Data Protection Regulation.” They propose a objective that minimizes the distance between the original point and a *counterfactual point* while pushing the counterfactual towards to be the desired outcome.

To induce their introduced desiderata of sparsity, Wachter et al. [28] propose as their distance function the Manhattan distance normalized by the median absolute deviation instead of Euclidean distance. The authors argue that counterfactuals that need fewer features to change are more human interpretable. Proceeding works introduce new desiderata and counterfactual explanation methods that encourage them, e.g., plausibility via direct presence in the dataset and the FACE method [15, 23]. In §3.1, we detail proximity and plausibility objectives’ relation to fair counterfactual explanations.

Gupta et al. [11] introduce “Equalizing Recourse,” the first notion of recourse fairness. They propose to measure fairness by the group-wise cost of recourse w.r.t. Euclidean distance to the decision boundary for negatively classified samples, and use this to calibrate a fair classifier. Slack et al. [25] uses this as a fair recourse measure, which we modify for indirect discrimination in Def. 3.1.

Taking this initial definition, Kügelgen et al. [17] introduce fairness for causal recourse. Causal recourse literature calls for recourse that adheres to the causal relationships between features [15]. However, this will result in discrimination if there are unfair relationships between protected and non-protected features. Confirming this, Kügelgen et al. [17] observed group cost disparities in causal recourse w.r.t. causal fairness metrics and proximity. Additionally, the authors found that there was no clear relationship between fairness w.r.t. causality and proximity. The discrimination and malfare definitions we introduce intuitively support causal recourse methods, given that all the objectives we evaluate them with are non-causal, e.g., proximity. If we want to compare recourse methods across causal recourse objectives, the evaluated recourse methods would have to be causal to guarantee compatibility. This is because some causal recourse objectives are only compatible with causal recourse methods, e.g., the percentage of recourse solutions that are successful under the true structural causal model [20].

Yetukuri et al. [30] acknowledges bias via plausibility objectives and proposes a plausibility-based fairness measure and objective for optimizing a fair classifier. These works on fairness and recourse, however, do not investigate the differences between recourse methods and how sensitive groups can be disadvantaged when these methods’ objectives do not align with group objectives. Other works have acknowledged the cost disparities given by their recourse methods, and suggest that their methods can be used to reveal discrimination [14, 18]. Therefore, we introduce three definitions of recourse discrimination such that these analyses can be tied to prior notions of fairness.

Toni et al. [27] presents a recourse method that accounts for user preferences. This human-in-the-loop method uses preference elicitation to iteratively ask users to pick from a set of interventions to estimate and generate interventions closer to their preferences. Here, an intervention is a set of actions where each action is a change to a single feature. Their method would allow users to choose interventions that align with their preferences. However, we differ by focusing on one-shot recourse scenarios where there is no iterative human feedback. We aim to find recourse robust to user preferences obtained from at worst a preliminary iteration of human feedback.

## 2 Preliminaries

We use a dataset  $\mathcal{D}$  with  $n$  samples, non-protected features  $\vec{x} \in \mathcal{X}$ , a categorical protected feature  $z \in \mathcal{Z}$ , and binary outcomes  $y \in Y = \{0, 1\}$ , i.e., the dataset  $\mathcal{D}$  is a list of  $n$  triples  $\mathcal{D} = \{(\vec{x}_1, z_1, y_1), \dots, (\vec{x}_n, z_n, y_n)\} \in (\mathcal{X} \times \mathcal{Z} \times Y)^n$ , where the  $i$ -th user has a set of unprotected features  $\vec{x}_i$ , a protected feature value of  $z_i$ , and a label  $y_i$ . We denote  $U(\mathcal{D})$  as the uniform distribution over  $\mathcal{D}$ . In Example 1.1,  $\vec{x}$  is education and work experience,  $z$  is a binary gender identity, and  $y$  indicates whether an application is accepted. Let  $f : (\mathcal{X}, \mathcal{Z}) \rightarrow Y$  be a binary classifier trained over the dataset  $\mathcal{D}$ . We use  $\mathcal{D}_g$  to denote the samples in  $\mathcal{D}$  whose protected feature is  $g$ , i.e.,  $\mathcal{D}_g = \{(\vec{x}, z, y) \in \mathcal{D} : z = g\}$ . We introduce definitions for a binary protected feature, but extensions to a categorical protected feature are straightforward. Following prior literature [11, 17, 25], our introduced definitions and recourse method focus on negatively classified users who want to be positively classified. We refer to the positively classified points users want as *counterfactuals*. We let  $\mathcal{D}^-$  be the samples in  $\mathcal{D}$  whose outcome  $y$  is 0, and  $\mathcal{D}^+$  be the samples in  $\mathcal{D}$  whose outcome is 1.

*Algorithmic Recourse.* Let  $\mathcal{A}$  denote a set of recourse methods,  $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{Z} \times Y \rightarrow Y$ . Let  $a((\vec{x}, z, y))$  be the counterfactual point output by the recourse method  $a \in \mathcal{A}$ . Given an original input  $(\vec{x}, z, y)$ , model  $f$ , and outcome  $y' = f((\vec{x}, z))$ , a recourse method  $a$  finds changes in the input needed such that  $f$  predicts the opposite outcome than  $y'$ . For instance, the method introduced by Wachter et al. [28] solves the following problem:

$$a_w((\vec{x}, z, y)) = \arg \min_{(\vec{x}', z')} (\lambda(f((\vec{x}', z')) - |y' - 1|)^2 + c((\vec{x}, z), (\vec{x}', z'))). \quad (1)$$

The first component  $(f((\vec{x}', z')) - |y' - 1|)^2$  pushes the counterfactual towards a different prediction than the original prediction and the latter part is a cost function,  $c$ .  $\lambda$  is a hyperparameter that balances the first and second components. By default we set  $\lambda = 1$ .

The *cost of recourse*,  $c$ , is the effort expended to change from the initial input  $(\vec{x}, z, y)$  to the counterfactual point  $a((\vec{x}, z, y))$ . For instance, if cost is the Euclidean distance between  $(\vec{x}, z, y)$  and  $a((\vec{x}, z, y))$ , then

$$c_e((\vec{x}, z, y), a((\vec{x}, z, y))) = \|(\vec{x}, z, y) - a((\vec{x}, z, y))\|_2.$$

Wachter et al.’s method uses the Manhattan distance weighted by the inverse median absolute deviation for  $c$ ; however, other cost functions are reasonable. To simplify notation, we write  $c((\vec{x}, z, y), a)$  instead of  $c((\vec{x}, z, y), a(\vec{x}, z, y))$ . The *objective* of a recourse method

is the cost function for which it optimizes, e.g.,  $c$  in Eq. (1). Conversely, the *objective* of a group is the cost function they prefer to minimize.

The preferred recourse objective of a group may be different from the objective of the method providing them with recourse. For instance, women in Example 1.1 may prefer receiving recourse with minimal Euclidean distance (Fig. 1a). However, they may be provided recourse from a method whose objective provides more realistic, i.e., plausible, recourse (Fig. 1c). We denote the set of cost functions/objectives that groups may choose from as  $C$ , and assume that  $C \neq \emptyset$ . Note that  $|C|$  does not have to equal the number of protected groups  $|Z|$ .

Prior works typically assume a single cost function for all users. Knowing the objectives that groups may choose from enables comparisons of how much worse, w.r.t. their potential preferences, the recourse they received from one method is versus the recourse they would have received from another method. For instance, in Example 1.1 we need to know that women may have a proximity objective in order to claim that the plausibility-based recourse method gave them worst recourse option w.r.t. proximity. Knowing  $C$  places more load on modelers on obtaining a set of reasonable cost functions, however, we argue that this is necessary for representing varied real-world preferences. Additionally, these cost functions can be obtained from human feedback through surveys or human-computer interaction and recourse studies [16].

### 3 Recourse and Discrimination

We tie recourse with three discrimination concepts: direct, indirect, and induced. The first two concepts draw inspiration from *unfair influence* and *fair relationship* as introduced by Grabowicz et al. [10].

**Illegal Influence** is the influence of protected feature(s) on outcomes deemed illegal via some legal instrument. In real-world contexts, decisions are often generated by features associated with the protected feature. Some of these features may be legally admissible for use if they are not *illegally influenced*, are relevant to decisions, and satisfy a business necessity [26].

**Legal relationship** relates protected feature(s) with non-protected feature(s). This relationship is deemed legal, e.g., the U.S. business necessity clause [26], in the context of some decision-making. Otherwise, it is deemed *illegal*.

Both notions describe cases where protected features are used by an automated decision maker, either directly or indirectly. Such uses can be sanctioned by law, but may still yield discriminatory outcomes.

Given a recourse method  $a \in \mathcal{A}$ , there is *direct discrimination* under cost  $c \in C$  if the expected cost of recourse changes when the protected feature is swapped for samples in  $\mathcal{D}$ , i.e., recourse is *illegally influenced* by the protected feature  $Z$ . In other words, there exists a point  $(\vec{x}, z, y)$  and a  $g \in Z$  such that  $|c((\vec{x}, g, y), a) - c((\vec{x}, z, y), a)| > 0$ .

The presence of direct recourse discrimination arguably violates U.S. anti-discrimination laws such as the Civil Rights Act of 1964 [26]. Indeed, recourse methods may exhibit direct discrimination even when the underlying classifier does not (see §3.1).

Direct recourse discrimination captures the illegal influence of the protected feature on outcomes; however, we need to measure indirect discrimination via other features, especially considering the historical use of features correlated with the protected feature as proxies (e.g., redlining).

**Definition 3.1** (Indirect recourse discrimination). Assuming that all non-protected features have an illegal relationship with the protected feature, given a recourse method  $a \in \mathcal{A}$ , the degree of *indirect discrimination* under cost  $c \in C$  is the difference in expected cost of recourse between two protected groups  $g_0, g_1 \in Z$ . We take the expectation over  $z$  to rule out any effects of the protected feature on the cost.

$$\mathbb{E}_{(\vec{x}, z, y) \sim U(\mathcal{D})} \left[ \left| \mathbb{E}_{(\vec{x}, \emptyset, y) \sim U(\mathcal{D}_{g_0}^-)} [c((\vec{x}, z, y), a)] - \mathbb{E}_{(\vec{x}, \emptyset, y) \sim U(\mathcal{D}_{g_1}^-)} [c((\vec{x}, z, y), a)] \right| \right]. \quad (2)$$

When this quantity is zero, there is no indirect discrimination, although it is more realistic to expect that the degree of indirect discrimination is small compared to the average cost of recourse, as *perfect* balance in cost is quite difficult to achieve. If the cost function  $c$  is the Euclidean distance, (2) is equivalent to Equalized Recourse [11]. For simplicity, we assume that all non-protected features have an illegal relationship with the protected feature, which is a consistent with the common “we are all equal” assumption [8]. Determining which features have illegal relationships is a separate task that requires domain expertise to judge properly [9]; among other challenges, it requires judging whether the influence on outcomes of non-protected features related with the protected feature satisfies a business necessity [26].

Direct and indirect recourse discrimination evaluate both the recourse method itself and the underlying decision-making model (or process). We disentangle the two such that we mainly capture the recourse method’s discrimination. Additionally, even if we find a recourse method that offers two groups an equalized expected cost of recourse, one of the groups might be denied lower-cost counterfactuals if it reduces the cost disparity between the groups.

For instance, in Fig. 3, assume that  $\vec{x}_1$  and  $\vec{x}_2$  have different protected group membership. If  $\vec{x}_1$  is provided  $\vec{c}_1$  and  $\vec{x}_2$  is provided  $\vec{c}_3$  for recourse, we have Equalized Recourse between the two points, i.e.,  $\|\vec{x}_1 - \vec{c}_1\|_2 = \|\vec{x}_2 - \vec{c}_3\|_2$ . However,  $\vec{x}_1$  was denied the closer point  $\vec{c}_2$ . Neither direct nor indirect discrimination measure this induced extra cost. To address these challenges, we measure the group-wise average difference in the cost of recourse for a given recourse method vs. the cost for the cheapest recourse method.

One can frame this as a form of *group disutility* in the choice of recourse from  $\mathcal{A}$ . More formally, given a set of recourse methods  $\mathcal{A}$ , a group  $g \in Z$ , and a recourse method  $a \in \mathcal{A}$ , we define the group  $g$ ’s disutility under cost  $c \in C$  as the difference between the lowest expected cost recourse method in  $\mathcal{A}$  and the expected cost of the given method,  $C(g, a, c)$  defined as:

$$\mathbb{E}_{(\vec{x}, z, y) \sim U(\mathcal{D}_g^-)} [c((\vec{x}, z, y), a)] - \min_{a' \in \mathcal{A}} \mathbb{E}_{(\vec{x}, z, y) \sim U(\mathcal{D}_g^-)} [c((\vec{x}, z, y), a')]. \quad (3)$$

**Definition 3.2** (Induced discrimination via recourse). The magnitude of *induced discrimination* by a recourse method  $a \in \mathcal{A}$  under

cost  $c \in C$  is the difference in disutility (as defined by Eq. (3)) between some two protected groups  $g_0, g_1 \in \mathcal{Z}$ ,

$$|C(g_0, a, c) - C(g_1, a, c)|. \quad (4)$$

When this quantity is zero, the recourse method  $a$  either does not induce discrimination or both groups have zero disutility when  $a$  is selected. If  $a$  optimizes for  $c$ , both groups will have zero disutility. Def. 3.2 captures the extra work one protected group would have to do over another for a given recourse method. For instance, if the cost of recourse is  $\ell_2$  distance, then this definition captures the extra distance needed to attain recourse.

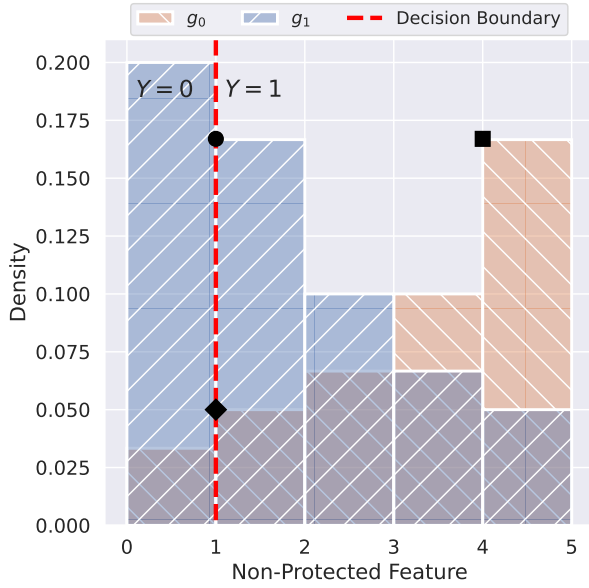


Figure 2: A scenario with protected groups  $g_0$  and  $g_1$  and non-protected feature  $x$ . Bar colors and lines represent conditional density at  $x$  for each group.

### 3.1 Recourse Objectives and Discrimination

We focus on two popular types of recourse objectives, *proximity* and *plausibility*, implemented in state-of-the-art methods and the tension between them [15].

**Proximity:** Recourse algorithms should provide the *nearest* counterfactual; that is, recourse should offer the closest counterfactual point w.r.t.  $\ell_2$  distance from the original point. In Fig. 3,  $\vec{x}_3$ 's closest point is  $\vec{c}_5$ . A proximity-based recourse method to pick  $\vec{c}_5$ , rather than the point along the data manifold  $\vec{c}_6$ .

**Plausibility:** Recourse algorithms should offer realistic recourse; that is, a counterfactual should lie in high-density regions or the data manifold, be part of the training data, or close to a prototypical example. Additionally, some plausibility objectives condition on the protected features. In Fig. 3,  $\vec{x}_3$  lies in the data manifold of the triangular points. Data manifold points will lead to a plausibility-based recourse method to pick  $\vec{c}_6$ , rather than the closer point  $\vec{c}_5$ . Individuals might prefer plausible recourse options over proximal options if they are less difficult to attain. For instance, to get a job

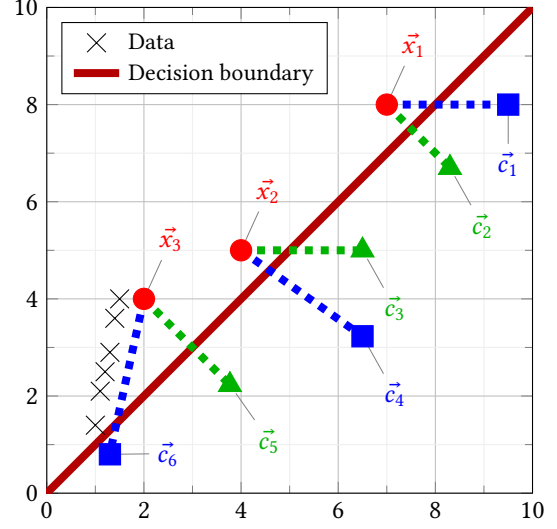


Figure 3: A two dimensional space with datapoints and potential counterfactuals. The three **circle** datapoints are negatively classified and have two possible counterfactuals, **square** and **triangle**. The triangle counterfactuals offer the lowest recourse cost w.r.t. the  $\ell_2$  distance. The square points represent some other recourse method. The black x's are points in the data manifold used by a plausibility objective.

that requires a bachelor's degree, it may be more plausible for a low-income high school graduate to attain an associate's and bachelor's degree. This recourse might be less proximal than seeking only a bachelor's degree, but the student may prefer it since tuition is typically less expensive this way in the U.S.

The different counterfactual points selected by methods under these objectives can lead to induced discrimination. For instance, in Fig. 2 there will be induced discrimination w.r.t. a proximity and a plausibility (conditional density) objective if  $\mathcal{A}$  contains a method for each objective,  $a$  and  $a'$  respectively. Assume that  $x = 0$  for all negatively classified ( $Y = 0$ ) points in this example. We refer to the circle, square, and diamond points in Fig. 2. When  $c$  is Euclidean distance, each group's disutility for  $a'$  is the difference between the distance from their densest positive point to initial point and the distance from their nearest positive point to initial point. In Fig. 2 this is the difference in the horizontal distances from the overlapping circles to  $x = 0$  for  $g_1$ , 0, and the difference in horizontal distances from the square and diamond to  $x = 0$  for  $g_0$ , 3. If  $c$  is negative density (to minimize density-based cost), each group's disutility for  $a$  is the difference in negative density from their nearest to densest positive points; the vertical difference between the overlapping circles for  $g_1$ , 0, and the vertical difference between the diamond to square point for  $g_0$ , 0.12. In both cases, we observe that  $g_0$  had higher disutilities than  $g_1$ , i.e., there was induced discrimination. If we add a new method that mixes proximity and plausibility to  $\mathcal{A}$ ,  $g_0$  will still have disutility. Any point where  $x \in [2, 4)$  will be further from  $g_0$ 's closest positive point (diamond) to  $x = 0$  and less dense than  $g_0$ 's densest positive point (square). In order to have zero disutility for both objectives, a recourse method would need

to offer both the most proximal point and the most plausible point. In Section 4 we propose a method that would accomplish this if provided these objectives. Our method offers multiple recourse options where each optimizes for a different objective.

*Relation with Fair Classification.* Fairness in recourse cost is not implied by statistically fair classifiers (and vice versa) [17]. This holds for our definitions of recourse discrimination. Recourse objectives often rely on information from the underlying dataset, ignoring any fairness criterion the classifier satisfies. Additionally, influence-based fair classifiers satisfying Equalized Recourse do not imply fair recourse w.r.t. our definitions. For instance, consider the jobs dataset in Example 1.1. If we apply a naive fair learning method that suppresses gender’s undesirable influence, men and women who have the same work experience and education will have the same expected outcomes. However, recourse methods with plausibility objectives requiring recourse examples present in the dataset, e.g., FACE [23], directly discriminate by disadvantaging women. Since men were better represented in the underlying dataset, men are likelier to have cheaper recourse options present in the dataset than women with the same non-protected features.

Similar behavior holds for the Equalized Recourse-calibrated classifier [11]. When this classifier is applied to the dataset in Example 1.1, negatively classified men and women have the same expected  $\ell_2$  distance to recourse, i.e. there is no indirect discrimination w.r.t.  $\ell_2$ . However, since this dataset contains illegal relationships, plausibility objectives penalize women’s closest recourse options as described in the example. So the more plausible recourse options will be selected over the closest  $\ell_2$  options that were required to satisfy Equalized Recourse’s indirect discrimination prevention. Even if we modify the Equalized Recourse classifier such that it reduces disparity for any given recourse cost, this behavior holds. If the selected recourse method optimizes for a cost function different than the Equalized Recourse classifier, the recourse it generates will not satisfy Equalized Recourse w.r.t. that method’s cost function.

*Actionability and Fairness.* One desiderata of recourse is that it should not call for changes in features that cannot reasonably be altered, i.e., they are *immutable*. Protected features fall under this immutable category. For instance, it is unreasonable to require a woman to change their gender identity to receive better recourse in Example 1.1. Actionability differs from objectives such as sparsity and plausibility as it is typically represented via constraints for constrained optimization on a recourse method’s cost function, i.e., it is not represented as cost function.

However, the actionability constraint does lead to disparate recourse costs, especially under plausibility objectives. Suppose we drop the actionability constraint in Example 1.1. If the dataset contains more men with a bachelor’s degree and no work experience than women, then plausible recourse for women with similar characteristics would be to swap their sex and then receive the men’s recourse. This results in less disparate costs, but calls for an unreasonable change.

For our proposed method and experiments, we implement actionability by rejecting counterfactual points that require non-actionable changes. However, we allow for changes on the protected feature when measuring direct and indirect recourse discrimination. This is necessary for capturing the effects of only the protected

feature (direct) and the effects of only the non-protected features (indirect) on recourse.

## 4 Welfare and Discriminatory Recourse

While Defs. 3.1 and 3.2 provide useful measures for group disparity and disutility in recourse, they are not suitable minimization objectives. This is because these discrimination definitions only measure the difference in recourse costs across groups, and not their actual costs. In other words, a recourse method may offer equally high recourse cost to all groups, and rank well on indirect and induced recourse discrimination. These challenges follow a common criticism of inequality measures as fairness objectives: they promote equality across groups but say nothing about actual group welfare [13].

To address this issue, we use social welfare measures to find recourse methods that balance both per-group recourse cost and parity. We define per-group recourse cost as,

$$C'(g, a, c) = \mathbb{E}_{(\vec{x}, z, y) \sim U(\mathcal{D}_g^-)} [c((\vec{x}, z, y), a)] .$$

We then aggregate per-group recourse cost with *power-mean malfare* functions, as argued for in fair ML by Cousins [5, 6], due to their convenient statistical properties, ease of optimization, and axiomatically-justified fairness properties. When recourse methods perform poorly w.r.t. malfare, while performing well w.r.t. indirect and induced discrimination, that means there was recourse equality, but poor performance over recourse costs for both groups.

There are numerous objectives groups may choose from, requiring a search over the cost functions  $C$ . We assume that all  $c \in C$  outputs are scaled s.t. their outputs can be compared against each other and  $\forall c \in C, c((\vec{x}, z, y), a) \geq 0$ . Note that, some cost functions return no solution if certain conditions are not met, making scaling and comparisons challenging. In our experiments, we only use cost functions that always return a solution and use min-max scaling w.r.t. all of the negative points’ costs. We propose the *cost-group nested power mean malfare*,

$$\mathbb{M}_{p,q}(a) = \sqrt[p]{\frac{1}{|\mathcal{Z}|} \sum_{g \in \mathcal{Z}} \left( \sqrt[q]{\sum_{c \in C} \mathbf{W}_{g,c} C'(g, a, c)^q} \right)^p} . \quad (5)$$

The inner malfare function calculates the disutility of a group using a  $q$ -mean. The outer malfare function computes overall disutility using a  $p$ -mean over all groups’ malfare. We distinguish between  $p$  and  $q$  as the welfare criterion to account for differences between how society views protected groups versus how each group views recourse costs. For example, society may prefer egalitarian treatment over groups for fairness ( $p = \infty$ ), while groups may be satisfied with good overall performance w.r.t. their own costs ( $q = 1$ ). We can encode a group  $g$ ’s cost preferences via a weight matrix  $\mathbf{W}$  as in Eq. (5). We assume that for every group  $g \in \mathcal{Z}$ ,  $\sum_{c \in C} \mathbf{W}_{g,c} = 1$  and that weights are non-negative.

If  $\mathbf{W}$  is unknown, it is reasonable to assume uniform preferences as a neutral choice, i.e.,  $\forall (g, c) \in \mathcal{Z} \times C : \mathbf{W}_{g,c} = 1/|C|$ . Malfare with uniform preferences also has convenient theoretical properties as in Thm. 4.1. In our experiments we assume uniform preferences.

When  $p = q$ , Eq. (5) is equivalent to the *cost-group  $p$ -power mean malfare* across the Cartesian product of the observed groups and

given cost functions  $C$ ,

$$\mathbb{M}_p(a) = \sqrt[p]{\frac{1}{|\mathcal{Z}|} \sum_{(g,c) \in \mathcal{Z} \times C} \mathbf{W}_{g,c} C'(g, a, c)^p}. \quad (6)$$

For the remainder of this manuscript, we set  $p = q$  for simplicity.

When  $p = q = 1$ ,  $\mathbb{M}_{1,1}(a)$  is equivalent to utilitarian malfare; when minimized, it offers the minimal total group-wise cost of recourse,

$$\arg \min_{a \in \mathcal{A}} \mathbb{M}_{1,1}(a) = \arg \min_{a \in \mathcal{A}} \frac{1}{|\mathcal{Z}|} \sum_{(g,c) \in \mathcal{Z} \times C} \mathbf{W}_{g,c} C'(g, a, c).$$

Utilitarian malfare can burden one group with a large share of the cost if other groups receive sufficiently low enough cost as a result. Alternatively, as  $p \rightarrow \infty$  and  $q \rightarrow \infty$ , the malfare function converges to egalitarian malfare,

$$\arg \min_{a \in \mathcal{A}} \mathbb{M}_{\infty,\infty}(a) = \arg \min_{a \in \mathcal{A}} \max_{(g,c) \in \mathcal{Z} \times C} \mathbf{W}_{g,c} C'(g, a, c).$$

This results in more equitable costs between groups, aligning it with induced discrimination minimization (Def. 3.2). Unlike induced discrimination, egalitarian malfare minimization may select recourse methods that unequally reduce the cost of recourse for advantaged and disadvantaged groups.

**Theorem 4.1.** Suppose  $p = q$ ,  $\forall (g, c) \in \mathcal{Z} \times C : \mathbf{W}_{g,c} = 1/|C|$ , and recourse method  $a^*$  is given by  $\arg \min_{a^* \in \mathcal{A}} \mathbb{M}_p(a^*)$ . Then for any other fair recourse objective  $r(a')$  taking  $a' = \arg \min_{a' \in \mathcal{A}} r(a')$ ; it holds that

$$\frac{1}{\sqrt[p]{|\mathcal{Z} \times C|}} \max_{\substack{g \in \mathcal{Z} \\ c \in C}} C'(g, a^*, c) \leq \mathbb{M}_p(a^*) \leq \max_{\substack{g \in \mathcal{Z} \\ c \in C}} C'(g, a', c).$$

The result follows by contradiction to the construction of  $a^*$  and the properties of power-mean functions (full proof in Appendix A). Thm. 4.1 bounds the worst-off group's disutility under  $a^*$ , for any  $p$  to never be more than a factor of  $\sqrt[p]{|\mathcal{Z} \times C|}$  worse than that of methods given by other fair recourse selection strategies. The largest gap in the bound for the worst-off group occurs when  $p = q = 1$ , motivating why higher values of  $p$  ( $p = q = \infty$ ) should be used. That is, for any fair recourse selection strategy that operates over  $\mathcal{A}$ , when  $p = q = \infty$  there is no way to achieve better recourse cost for the worst-off group than egalitarian malfare minimization. This offers a mathematically precise trade-off between group fairness and total cost. So when group objectives  $C$  are known and there are multiple recourse methods to choose from ( $|\mathcal{A}| > 1$ ), our analysis suggests selecting the method that minimizes egalitarian malfare over imposing explicit cost-equality constraints.

**Objective-Robust Recourse.** We propose an *objective-robust* recourse method (ORR), i.e., a method that minimizes malfare w.r.t. objectives in a set of *potential* cost functions  $C$ . Let  $\zeta$  be the cost given by  $c$  for point  $(\vec{x}, z, y)$ 's best counterfactual w.r.t. cost  $c'$ , i.e.,  $\zeta((\vec{x}, z, y), g, c, c')$  equals

$$c \left( (\vec{x}, z, y), \arg \min_{(\vec{x}', z', y') \in \mathcal{D}_g^+} c'((\vec{x}, z, y), (\vec{x}', z', y')) \right).$$

This first finds the minimum-cost counterfactual for the given point w.r.t.  $c'$ . Then, it computes the cost between the found counterfactual point and the given point w.r.t.  $c$ . Our method then finds a set of  $k$

objectives  $C_k \in \binom{\hat{C}}{k}$  that minimize malfare over all objectives in  $C$ , i.e.,  $C^*(C, \hat{C}, k, p, q)$  equals

$$\arg \min_{C_k \in \binom{\hat{C}}{k}} \sum_{c' \in C_k} \sqrt[p]{\frac{1}{|\mathcal{Z}|} \sum_{g \in \mathcal{Z}} \left( \sum_{c \in C} \mathbf{W}_{g,c} \left( \mathbb{E}_{\substack{(\vec{x}, z, y) \\ \sim U(\mathcal{D}_g^+)}} [\zeta((\vec{x}, z, y), g, c, c')] \right)^q \right)^{\frac{p}{q}}}.$$

That is, we generate recourse w.r.t. the objectives given by  $\hat{C}$  and calculate its malfare over the objectives in  $C$ . Then, we return the  $k$  objectives in  $\hat{C}$  that had minimal malfare. We specify  $k$  to avoid providing the user with too many recourse options later in Eq. (7). For example, if  $|\hat{C}| = 50$ , Eq. (7) would return 50 recourse options to users and overload them. Additional details on  $C^*$  such as actionability and runtime complexity are in Appendix B.

$\hat{C}$  and  $C$  are different since one may want to generate recourse w.r.t. different objectives than what is used to calculate malfare, e.g., recourse is generated w.r.t.  $\hat{C} = \{c_{\ell_0} + c_{\ell_2}, c_{\ell_{1.5}}\}$  and its malfare is calculated over  $C = \{c_{\ell_1}, c_{\ell_2}\}$ . We recommend that  $\hat{C} = C$  or that  $\hat{C}$  contains mixtures of the cost functions in  $C$ . If  $C$  is unknown, we recommend that practitioners populate  $C$  with cost functions related to proximity, sparsity, and plausibility objectives. This means that our method will only robust w.r.t. these three high-level objectives, however, these three are the most popular in the recourse literature [15]. Again, we assume that the costs in  $C$ 's outputs are scaled s.t. they can be compared against each other and are non-negative.

When  $p = q = \infty$ , we obtain the special case where  $C^*(C, \hat{C}, k, \infty, \infty)$  equals

$$\arg \min_{C_k \in \binom{\hat{C}}{k}} \sum_{c' \in C_k} \max_{(g,c) \in \mathcal{Z} \times C} \mathbf{W}_{g,c} \mathbb{E}_{(\vec{x}, z, y) \sim U(\mathcal{D}_g^+)} [\zeta((\vec{x}, z, y), g, c, c')].$$

Our recourse method then optimizes the above. In particular, we return  $k$  counterfactual points for each objective returned by  $C^*$ , i.e.,  $a_{\mathcal{M}}((\vec{x}, z, y))$  equals

$$\left\{ \arg \min_{(\vec{x}', z', y') \in \mathcal{D}_g^+} \left( \lambda(1 - f((\vec{x}', z'), z)) + c((\vec{x}, z), (\vec{x}', z')) \right) \mid c \in C^* \right\}. \quad (7)$$

Next, we specify a *default* objective-robust recourse method (D-ORR) to use when there is no knowledge of  $C$ . We populate  $C$  with scaled formulations of  $\ell_0$  ( $c_1$ ),  $\ell_2$  ( $c_2$ ), and the local outlier factor score with 15 neighbors ( $c_3$ ). We then include in  $\hat{C}$  all possible binary weightings of a linear combination of these objectives. That is,  $C = \{c_1, c_2, c_3\}$  and  $\hat{C} = \{c_1 + 0c_2 + 0c_3, c_1 + c_2 + 0c_3, \dots, 0c_1 + c_2 + c_3, c_1 + c_2 + c_3\}$ . We leave the choice of  $p$ ,  $q$ , and  $k$  for  $C^*$  to practitioners.

## 5 Experiments

We empirically evaluate seven recourse methods on two benchmark datasets, testing whether they suffer from indirect and induced discrimination on six recourse objectives. We measure utilitarian and egalitarian malfare to determine the malfare minimizing recourse methods. We extend the library by Hamer et al. [12] to support our discrimination and malfare benchmarks.

### 5.1 Datasets and Learning Methods

We evaluate recourse methods on real-world datasets common in fairness and explainability research — the Adult Census Income [3]

| Dataset        | Method      | $\ell_0$                   | $\ell_1$                   | $\ell_{1.5}$               | $\ell_2$                   | 5NN                        | Density                    |
|----------------|-------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| Credit Default | Wachter     | 0.6204 $\pm$ 0.0045        | 0.0468 $\pm$ 0.0016        | 0.0861 $\pm$ 0.0035        | 0.0698 $\pm$ 0.0024        | 0.0209 $\pm$ 0.0007        | 0.0029 $\pm$ 0.0003        |
|                | Elastic Net | 0.6285 $\pm$ 0.0050        | 0.0434 $\pm$ 0.0014        | 0.0789 $\pm$ 0.0033        | 0.0629 $\pm$ 0.0022        | 0.0271 $\pm$ 0.0012        | 0.0040 $\pm$ 0.0005        |
|                | DiCE        | <b>0.0066</b> $\pm$ 0.0006 | 0.2855 $\pm$ 0.0138        | 0.2797 $\pm$ 0.0127        | 0.2821 $\pm$ 0.0120        | 0.2354 $\pm$ 0.0107        | 0.0942 $\pm$ 0.0085        |
|                | FACE        | 0.7076 $\pm$ 0.0052        | 0.0478 $\pm$ 0.0016        | 0.1012 $\pm$ 0.0041        | 0.0765 $\pm$ 0.0026        | 0.0273 $\pm$ 0.0014        | 0.0038 $\pm$ 0.0005        |
|                | CCHVAE      | 0.9077 $\pm$ 0.0066        | 0.0421 $\pm$ 0.0014        | 0.1016 $\pm$ 0.0038        | 0.0735 $\pm$ 0.0023        | 0.0266 $\pm$ 0.0015        | <b>0.0019</b> $\pm$ 0.0003 |
|                | ORR         | 0.4850 $\pm$ 0.0040        | <b>0.0377</b> $\pm$ 0.0012 | <b>0.0660</b> $\pm$ 0.0026 | <b>0.0530</b> $\pm$ 0.0017 | 0.0223 $\pm$ 0.0011        | 0.0031 $\pm$ 0.0004        |
|                | D-ORR       | 0.4866 $\pm$ 0.0039        | 0.0439 $\pm$ 0.0015        | 0.0785 $\pm$ 0.0029        | 0.0632 $\pm$ 0.0020        | <b>0.0157</b> $\pm$ 0.0007 | 0.0023 $\pm$ 0.0003        |
| Adult Census   | Wachter     | 0.3130 $\pm$ 0.0028        | 0.1332 $\pm$ 0.0040        | 0.2974 $\pm$ 0.0089        | 0.2303 $\pm$ 0.0072        | 0.1754 $\pm$ 0.0061        | 0.0881 $\pm$ 0.0052        |
|                | Elastic Net | 0.3300 $\pm$ 0.0032        | 0.1173 $\pm$ 0.0035        | 0.2912 $\pm$ 0.0086        | 0.2215 $\pm$ 0.0069        | 0.1654 $\pm$ 0.0063        | 0.0826 $\pm$ 0.0052        |
|                | DiCE        | <b>0.0041</b> $\pm$ 0.0010 | 0.3420 $\pm$ 0.0176        | 0.3534 $\pm$ 0.0177        | 0.3502 $\pm$ 0.0178        | 0.4518 $\pm$ 0.0131        | 0.3028 $\pm$ 0.0165        |
|                | FACE        | 0.6193 $\pm$ 0.0061        | 0.2221 $\pm$ 0.0068        | 0.5559 $\pm$ 0.0166        | 0.3936 $\pm$ 0.0124        | 0.2335 $\pm$ 0.0096        | 0.1304 $\pm$ 0.0086        |
|                | CCHVAE      | 0.4120 $\pm$ 0.0058        | 0.1344 $\pm$ 0.0040        | 0.2756 $\pm$ 0.0090        | 0.2148 $\pm$ 0.0070        | 0.2314 $\pm$ 0.0080        | 0.0895 $\pm$ 0.0060        |
|                | ORR         | 0.2041 $\pm$ 0.0024        | 0.1034 $\pm$ 0.0030        | <b>0.2518</b> $\pm$ 0.0078 | <b>0.1954</b> $\pm$ 0.0062 | 0.1751 $\pm$ 0.0061        | 0.0844 $\pm$ 0.0051        |
|                | D-ORR       | 0.2349 $\pm$ 0.0071        | <b>0.1022</b> $\pm$ 0.0036 | 0.2654 $\pm$ 0.0106        | 0.2043 $\pm$ 0.0079        | <b>0.0910</b> $\pm$ 0.0036 | <b>0.0581</b> $\pm$ 0.0037 |

**Table 1: Overall recourse cost  $\pm$  standard error over 10 trials w.r.t. normalized objectives. Lower is better. Bold indicates the lowest-cost recourse method for each objective (column).**

| Dataset        | Method      | $\ell_0$                   | $\ell_1$                   | $\ell_{1.5}$               | $\ell_2$                   | 5NN                        | Density                    |
|----------------|-------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| Credit Default | Wachter     | 0.0244 $\pm$ 0.0033        | 0.0014 $\pm$ 0.0003        | 0.0012 $\pm$ 0.0003        | 0.0011 $\pm$ 0.0002        | 0.0019 $\pm$ 0.0005        | 0.0003 $\pm$ 0.0001        |
|                | Elastic Net | 0.0200 $\pm$ 0.0038        | 0.0011 $\pm$ 0.0003        | 0.0019 $\pm$ 0.0006        | 0.0017 $\pm$ 0.0005        | 0.0016 $\pm$ 0.0004        | 0.0004 $\pm$ 0.0001        |
|                | DiCE        | <b>0.0000</b> $\pm$ 0.0000 | 0.0109 $\pm$ 0.0029        | 0.0085 $\pm$ 0.0018        | 0.0089 $\pm$ 0.0023        | 0.0131 $\pm$ 0.0029        | 0.0079 $\pm$ 0.0015        |
|                | FACE        | 0.0127 $\pm$ 0.0022        | 0.0031 $\pm$ 0.0004        | 0.0047 $\pm$ 0.0006        | 0.0037 $\pm$ 0.0005        | 0.0017 $\pm$ 0.0004        | 0.0002 $\pm$ 0.0001        |
|                | CCHVAE      | 0.0444 $\pm$ 0.0004        | 0.0017 $\pm$ 0.0003        | 0.0014 $\pm$ 0.0003        | 0.0012 $\pm$ 0.0005        | 0.0017 $\pm$ 0.0004        | <b>0.0001</b> $\pm$ 0.0001 |
|                | ORR         | 0.0195 $\pm$ 0.0031        | <b>0.0000</b> $\pm$ 0.0000 | <b>0.0000</b> $\pm$ 0.0000 | <b>0.0000</b> $\pm$ 0.0000 | 0.0010 $\pm$ 0.0003        | 0.0002 $\pm$ 0.0001        |
|                | D-ORR       | 0.0198 $\pm$ 0.0031        | 0.0010 $\pm$ 0.0002        | 0.0010 $\pm$ 0.0002        | 0.0007 $\pm$ 0.0002        | <b>0.0000</b> $\pm$ 0.0000 | 0.0003 $\pm$ 0.0001        |
| Adult Census   | Wachter     | 0.0990 $\pm$ 0.0046        | 0.0094 $\pm$ 0.0017        | 0.0078 $\pm$ 0.0013        | 0.0045 $\pm$ 0.0017        | 0.0148 $\pm$ 0.0038        | 0.0056 $\pm$ 0.0011        |
|                | Elastic Net | 0.0947 $\pm$ 0.0048        | 0.0049 $\pm$ 0.0009        | 0.0084 $\pm$ 0.0010        | 0.0031 $\pm$ 0.0012        | 0.0161 $\pm$ 0.0028        | 0.0051 $\pm$ 0.0015        |
|                | DiCE        | <b>0.0000</b> $\pm$ 0.0000 | 0.0663 $\pm$ 0.0045        | 0.0173 $\pm$ 0.0036        | 0.0347 $\pm$ 0.0041        | 0.0382 $\pm$ 0.0049        | 0.0444 $\pm$ 0.0050        |
|                | FACE        | 0.0643 $\pm$ 0.0032        | 0.0139 $\pm$ 0.0010        | 0.0298 $\pm$ 0.0024        | 0.0283 $\pm$ 0.0014        | 0.0398 $\pm$ 0.0042        | 0.0131 $\pm$ 0.0019        |
|                | CCHVAE      | 0.0757 $\pm$ 0.0037        | 0.0066 $\pm$ 0.0012        | 0.0076 $\pm$ 0.0023        | 0.0089 $\pm$ 0.0018        | 0.0128 $\pm$ 0.0029        | 0.0130 $\pm$ 0.0017        |
|                | ORR         | 0.0694 $\pm$ 0.0028        | <b>0.0022</b> $\pm$ 0.0005 | <b>0.0000</b> $\pm$ 0.0000 | <b>0.0000</b> $\pm$ 0.0000 | 0.0120 $\pm$ 0.0026        | 0.0023 $\pm$ 0.0006        |
|                | D-ORR       | 0.0752 $\pm$ 0.0030        | 0.0037 $\pm$ 0.0016        | 0.0016 $\pm$ 0.0006        | 0.0020 $\pm$ 0.0008        | <b>0.0000</b> $\pm$ 0.0000 | <b>0.0000</b> $\pm$ 0.0000 |

**Table 2: Group-wise induced discrimination  $\pm$  standard error over 10 trials w.r.t. normalized objectives. Lower is better. Bold indicates the recourse method with the least induced discrimination for each objective (column).**

and Credit Card Default [29] datasets. We use data preprocessing functions modified from the AIF360 library [4] and standardize both datasets. We train a logistic regression model on both datasets; the protected feature we consider is sex (binary Male or Female).

**Adult Census:** a dataset where the outcome is whether an individual's income exceeds \$50K/yr. Features describe individuals' demographics, education, work status, and capital gains/losses. It contains a total of 48842 samples.

**Credit Card Default:** a financial dataset with 30000 samples. The outcome is whether a customer will default on their credit card. Features describe each customer's payment and billing history and demographics.

For both datasets, we set all demographic features as non-actionable features.

## 5.2 Evaluated Methods and Objectives

We analyze our objective-robust method and methods with model-agnostic implementations, each with differing objectives. We evaluate our objective-robust recourse method assuming  $C$  is known and the default objective-robust recourse method (ORR & D-ORR). For these methods we set  $p = q = \infty$  and  $\forall (g, c) \in \mathcal{Z} \times C : W_{g,c} = 1/|C|$ . We compare our methods to the following recourse methods.

**Wachter's Method [28]:** This method solves Eq. (1), using Manhattan distance normalized by the median absolute deviation for the cost function.

**DiCE [21]:** Diverse Counterfactual Explanations (DiCE) finds  $k$  counterfactual points that satisfy a variant of (1). This variant includes an objective encouraging repulsion between the counterfactual points. Once the counterfactual points are selected, DiCE performs a post-hoc operation encouraging sparsity.

| Dataset        | Method      | $\ell_0$                   | $\ell_1$                   | $\ell_{1.5}$               | $\ell_2$                   | 5NN                        | Density                    |
|----------------|-------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| Credit Default | Wachter     | 0.0240 $\pm$ 0.0033        | 0.0037 $\pm$ 0.0005        | 0.0106 $\pm$ 0.0006        | 0.0078 $\pm$ 0.0005        | 0.0024 $\pm$ 0.0004        | 0.0003 $\pm$ 0.0001        |
|                | Elastic Net | 0.0197 $\pm$ 0.0037        | 0.0047 $\pm$ 0.0005        | 0.0112 $\pm$ 0.0009        | 0.0084 $\pm$ 0.0008        | 0.0014 $\pm$ 0.0003        | 0.0003 $\pm$ 0.0001        |
|                | DiCE        | <b>0.0007</b> $\pm$ 0.0002 | 0.0147 $\pm$ 0.0031        | 0.0132 $\pm$ 0.0031        | 0.0137 $\pm$ 0.0030        | 0.0130 $\pm$ 0.0027        | 0.0080 $\pm$ 0.0015        |
|                | FACE        | 0.0124 $\pm$ 0.0022        | <b>0.0020</b> $\pm$ 0.0002 | <b>0.0063</b> $\pm$ 0.0005 | <b>0.0045</b> $\pm$ 0.0003 | 0.0014 $\pm$ 0.0003        | 0.0002 $\pm$ 0.0001        |
|                | CCHVAE      | 0.0442 $\pm$ 0.0003        | 0.0034 $\pm$ 0.0006        | 0.0110 $\pm$ 0.0010        | 0.0071 $\pm$ 0.0008        | 0.0013 $\pm$ 0.0002        | 0.0002 $\pm$ 0.0000        |
|                | ORR         | 0.0192 $\pm$ 0.0030        | 0.0051 $\pm$ 0.0004        | 0.0110 $\pm$ 0.0008        | 0.0082 $\pm$ 0.0006        | <b>0.0010</b> $\pm$ 0.0002 | 0.0002 $\pm$ 0.0000        |
|                | D-ORR       | 0.0196 $\pm$ 0.0030        | 0.0042 $\pm$ 0.0005        | 0.0114 $\pm$ 0.0011        | 0.0084 $\pm$ 0.0008        | 0.0011 $\pm$ 0.0002        | 0.0002 $\pm$ 0.0000        |
| Adult Census   | Wachter     | 0.0975 $\pm$ 0.0049        | 0.0072 $\pm$ 0.0024        | 0.0611 $\pm$ 0.0028        | 0.0380 $\pm$ 0.0025        | 0.0537 $\pm$ 0.0053        | 0.0393 $\pm$ 0.0028        |
|                | Elastic Net | 0.0932 $\pm$ 0.0048        | 0.0084 $\pm$ 0.0014        | 0.0623 $\pm$ 0.0025        | 0.0402 $\pm$ 0.0018        | 0.0525 $\pm$ 0.0049        | 0.0380 $\pm$ 0.0024        |
|                | DiCE        | <b>0.0020</b> $\pm$ 0.0005 | 0.0787 $\pm$ 0.0051        | 0.0738 $\pm$ 0.0050        | 0.0751 $\pm$ 0.0052        | 0.0897 $\pm$ 0.0044        | 0.0846 $\pm$ 0.0061        |
|                | FACE        | 0.0628 $\pm$ 0.0034        | <b>0.0027</b> $\pm$ 0.0006 | <b>0.0268</b> $\pm$ 0.0009 | <b>0.0121</b> $\pm$ 0.0005 | 0.0913 $\pm$ 0.0048        | 0.0534 $\pm$ 0.0039        |
|                | CCHVAE      | 0.0742 $\pm$ 0.0041        | 0.0064 $\pm$ 0.0013        | 0.0492 $\pm$ 0.0023        | 0.0315 $\pm$ 0.0015        | 0.0616 $\pm$ 0.0044        | <b>0.0273</b> $\pm$ 0.0029 |
|                | ORR         | 0.0679 $\pm$ 0.0029        | 0.0146 $\pm$ 0.0010        | 0.0565 $\pm$ 0.0021        | 0.0404 $\pm$ 0.0015        | 0.0635 $\pm$ 0.0034        | 0.0400 $\pm$ 0.0026        |
|                | D-ORR       | 0.0736 $\pm$ 0.0030        | 0.0088 $\pm$ 0.0013        | 0.0556 $\pm$ 0.0020        | 0.0384 $\pm$ 0.0014        | <b>0.0515</b> $\pm$ 0.0028 | 0.0402 $\pm$ 0.0030        |

**Table 3: Group-wise indirect discrimination  $\pm$  standard error over 10 trials w.r.t. normalized objectives. Lower is better. Bold indicates the recourse method with the least indirect discrimination for each objective (column).**

| Dataset        | Method      | $M_{1,1}(a)$               | $M_{\infty,\infty}(a)$     |
|----------------|-------------|----------------------------|----------------------------|
| Credit Default | Wachter     | 0.1412 $\pm$ 0.0014        | 0.6324 $\pm$ 0.0047        |
|                | Elastic Net | 0.1408 $\pm$ 0.0015        | 0.6383 $\pm$ 0.0060        |
|                | DiCE        | 0.1973 $\pm$ 0.0089        | <b>0.2995</b> $\pm$ 0.0142 |
|                | FACE        | 0.1607 $\pm$ 0.0014        | 0.7139 $\pm$ 0.0052        |
|                | CCHVAE      | 0.1922 $\pm$ 0.0015        | 0.9298 $\pm$ 0.0067        |
|                | ORR         | <b>0.1112</b> $\pm$ 0.0011 | 0.4946 $\pm$ 0.0036        |
|                | D-ORR       | 0.1150 $\pm$ 0.0012        | 0.4964 $\pm$ 0.0036        |
| Adult Census   | Wachter     | 0.2062 $\pm$ 0.0035        | 0.3669 $\pm$ 0.0038        |
|                | Elastic Net | 0.2013 $\pm$ 0.0033        | 0.3781 $\pm$ 0.0043        |
|                | DiCE        | 0.3007 $\pm$ 0.0094        | 0.4967 $\pm$ 0.0149        |
|                | FACE        | 0.3591 $\pm$ 0.0052        | 0.6540 $\pm$ 0.0070        |
|                | CCHVAE      | 0.2263 $\pm$ 0.0039        | 0.4491 $\pm$ 0.0067        |
|                | ORR         | 0.1690 $\pm$ 0.0028        | <b>0.2801</b> $\pm$ 0.0082 |
|                | D-ORR       | <b>0.1593</b> $\pm$ 0.0033 | 0.3047 $\pm$ 0.0077        |

**Table 4: Utilitarian ( $M_{1,1}(a)$ ) and egalitarian ( $M_{\infty,\infty}(a)$ ) mal-fare  $\pm$  standard error over 10 trials. Lower is better. Bold indicates the recourse method with the least of each mal-fare metric.**

**FACE [23]:** Feasible and Actionable Counterfactual Explanations (FACE) constructs a weighted graph of the data and runs Dijkstra’s algorithm from the input point to counterfactual points. Each point is a node and for any two nodes whose  $\ell_2$  distance is under a threshold  $\epsilon$ , an edge is added connecting them, encouraging plausibility. We set  $\epsilon$  to be 4.

**C-CHVAE [22]:** C-CHVAE uses a variational autoencoder to construct a latent neighborhood within a radius around the input point and iteratively expands this radius counterfactual point is found. The autoencoder approximates the conditional log likelihood of the mutable features given the immutable (protected) features. This encourages proximal and plausible recourse for C-CHVAE. We run C-CHVAE for 50 iterations with a step distance of 1.

**Elastic Net [7]:** Elastic Net solves Eq. (1) using cost

$$c((\vec{x}, z), (\vec{x}', z')) = \|(\vec{x}, z) - (\vec{x}', z')\|_1 + \|(\vec{x}, z) - (\vec{x}', z')\|_2^2,$$

adding an  $\ell_2$  proximity incentive over Wachter’s method.

We use the authors’ code for DiCE, and use FACE and C-CHVAE implementations from Hamer et al. [12]. For DiCE and the objective-robust methods (ORR & D-ORR), we set  $k = 3$  for both datasets and pick the lowest cost counterfactual for each sample, measure, and objective. Our selection process for  $k$  and  $\epsilon$  and other implementation details are in Appendix C. Other hyperparameters are set to their default values.

For each recourse method/dataset combination, we calculate indirect and induced recourse discrimination (Defs. 3.1 and 3.2). Additionally, we present the  $p = q = 1$  and  $p = q = \infty$  mal-fare results. For each method, we also calculate overall recourse cost, i.e., expected cost over all samples. We evaluate over 1000 negative samples requiring recourse and aggregate over 10 trials of different random seeds for data splits, models, and recourse methods for both datasets. We balance the negative samples by the protected feature we focus on. Samples where recourse fails are not used in our analysis, however, all methods had a near perfect or perfect success rate.

We evaluate six group recourse objectives, representing a scenario where both *groups* wish to minimize these objectives.

**Sparsity:** represented by the  $\ell_0$  distance.

**Proximity/Sparsity mixture:** represented by  $\ell_1$  &  $\ell_{1.5}$  distance.

**Proximity:** represented by the  $\ell_2$  distance.

**Plausibility:** we evaluate plausibility in two ways: (i) the average  $\ell_2$  distance from the counterfactual’s five nearest neighbors (labeled as “5NN”), and (ii) the negative log densities via a Kernel Density Estimator as another plausibility objective.

We scale these objectives’ outputs over all negatively classified points using min-max scaling. We calculate overall cost, mal-fare, and discrimination with respect to these objectives. Even though some of these group objectives have the same high-level proximity, sparsity, and plausibility goals as the evaluated recourse methods’

objectives, they do not measure the same effects. This represents how the objectives for groups do not perfectly align with the objectives of recourse methods.

### 5.3 Results

We present results for both datasets in Tabs. 1 to 4. We initially observed that all recourse methods induce discrimination under at least one group objective. Note that the best performing recourse method w.r.t. induced discrimination was typically used as the lowest cost method when computing Eq. (3). Additionally, each group objective had indirect recourse discrimination under at least one recourse method. We observed that for almost all discrimination measures in both datasets, women were discriminated against; signaling the need for careful recourse method selection.

Our results on utilitarian and egalitarian malfare empirically revealed malfare’s relationship with recourse cost and discrimination. For instance, FACE had the lowest or second-lowest indirect discrimination with respect to all the  $\ell$  norm group objectives. However, since FACE had higher costs and induced discrimination than most methods, it also had worse malfare than them. We also observed a special case under egalitarian malfare in the Credit Default dataset. DiCE had the best egalitarian malfare, despite having the worse performance over five of the group objectives. This was because all the other methods had exceptionally high  $\ell_0$  group cost, which dominated their egalitarian malfare. However, DiCE’s poor group costs over the other five objectives was captured by utilitarian malfare. Therefore, we recommend evaluating recourse with both utilitarian and egalitarian malfare. Methods that had lower malfare for *both* criteria had lower group costs and, subsequently, induced less discrimination.

Our observations about malfare measures support why the methods that optimize for malfare, ORR and D-ORR, should be used for recourse. As expected, ORR and D-ORR had the best malfare, except for egalitarian malfare in the Credit Default dataset where they were second best. Therefore, these methods had better group costs and induced discrimination than most other recourse methods over all group objectives, i.e. they had the most robust performance over the objectives. Additionally, D-ORR had similar malfare to ORR, even though it did not directly optimize malfare for all the measured group objectives. Compared to ORR, D-ORR had lower costs and induced discrimination over the plausibility objectives, while sacrificing minimal performance over the remaining  $\ell$  norm objectives. This supports the notion that if the objective-robust method optimizes for a different, but related, set of objectives to the group objectives, we can still obtain low malfare recourse.

## 6 Conclusion

We present a novel point-of-view for discrimination in algorithmic recourse. Rather than focusing on the disparate recourse costs across ML models, we introduce a definition that captures how recourse methods themselves induce discrimination. We describe how the objectives suggested by state-of-the-art recourse methods can induce discrimination, even when explaining fair ML methods and when there is no indirect recourse discrimination. Inspired by influence-based fairness notions, we define discrimination for recourse. Then, we take welfare-based measures used for fair ML to

measures recourse methods’ performance in recourse costs and parity w.r.t. multiple recourse objectives and groups. We then propose a recourse method that optimizes for malfare and yields recourse options over multiple objectives. Lastly, we evaluate our method and popular recourse methods on real-world datasets. The presented results revealed the recourse disparity between groups and how methods that minimized malfare perform well across recourse parity and overall cost measures.

## References

- [1] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *Proceedings of the 5th ACM Conference on Fairness, Accountability and Transparency (FAccT)*.
- [2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. <http://www.fairmlbook.org>.
- [3] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5XW20>.
- [4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943* (2018). arXiv:1810.01943 <https://arxiv.org/abs/1810.01943>.
- [5] Cyrus Cousins. 2021. An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 16610–16621.
- [6] Cyrus Cousins. 2023. Revisiting fair-PAC learning and the axioms of cardinal welfare. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 6422–6442.
- [7] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*. 590–601.
- [8] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv:1609.07236 [cs.CY]* <https://arxiv.org/abs/1609.07236>.
- [9] Przemyslaw Grabowicz, Nicholas Perello, and Yair Zick. 2023. Towards an AI Accountability Policy. *arXiv preprint arXiv:2307.13658* (2023).
- [10] Przemyslaw A. Grabowicz, Nicholas Perello, and Aarshree Mishra. 2022. Marrying Fairness and Explainability in Supervised Learning. In *Proceedings of the 5th ACM Conference on Fairness, Accountability and Transparency (FAccT)*. 1905–1916.
- [11] Vivek Gupta, Pegah Nokhiz, Chitradheep Dutta Roy, and Suresh Venkatasubramanian. 2020. Equalizing Recourse across Groups. *arXiv preprint arXiv:1909.03166* (2020). arXiv:1909.03166 <https://arxiv.org/abs/1909.03166>.
- [12] Jenny Hamer, Nicholas Perello, Jason Valladares, Vignesh Viswanathan, and Yair Zick. 2024. Simple Steps to Success: A Method for Step-Based Counterfactual Explanations. *Transactions on Machine Learning Research* (2024). <https://openreview.net/forum?id=R6ey5DKaoX>.
- [13] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability and Transparency (FAccT)*. Association for Computing Machinery, 535–545. doi:10.1145/3351095.3372857.
- [14] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-Agnostic Counterfactual Explanations for Consequential Decisions. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*. 895–905.
- [15] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Survey*, Article 95 (dec 2022), 29 pages.
- [16] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. In *Proceedings of the 5th ACM Conference on Fairness, Accountability and Transparency (FAccT)*.
- [17] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2022. On the Fairness of Causal Algorithmic Recourse. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 9 (Jun. 2022), 9584–9594. doi:10.1609/aaai.v36i9.21192.
- [18] Arnaud Van Looveren and Janis Klaise. 2019. Interpretable Counterfactual Explanations Guided by Prototypes. In *Proceedings of the 2019 European Conference on Machine Learning and Data Mining*. 650–665.

- [19] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 4768–4777.
- [20] Ayan Majumdar and Isabel Valera. 2024. CARMA: A practical framework to generate recommendations for causal algorithmic recourse at scale. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (Rio de Janeiro, Brazil) (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1745–1762. doi:10.1145/3630106.3659003
- [21] Ramaravind K. Muthilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 3rd ACM Conference on Fairness, Accountability and Transparency (FAccT)*. 607–617.
- [22] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of the 26th International World Wide Web Conference (WWW)*. 3126–3132.
- [23] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the 3rd AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIIES)*. 344–350.
- [24] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models. In *Proceedings of the 3rd AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIIES)*. 166–172.
- [25] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual explanations can be manipulated. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 62–75.
- [26] Title VII of the Civil Rights Act. 1964. 7, 42 U.S.C., 2000e et seq.
- [27] Giovanni De Toni, Paolo Viappiani, Stefano Teso, Bruno Lepri, and Andrea Passerini. 2024. Personalized Algorithmic Recourse with Preference Elicitation. *Transactions on Machine Learning Research* (2024). <https://openreview.net/forum?id=8sg2l9zXgO>
- [28] S Wachter, B Mittelstadt, and C Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2017), 842–887.
- [29] I-Cheng Yeh. 2016. Default of Credit Card Clients. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C55S3H>.
- [30] Jayanth Yetukuri, Ian Hardy, Yevgeniy Vorobeychik, Berk Ustun, and Yang Liu. 2024. Providing Fair Recourse over Plausible Groups. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 19 (Mar. 2024), 21753–21760. doi:10.1609/aaai.v38i19.30175

## A Theorem 3.3 Proof

PROOF. First observe that  $a' = \arg \min_{a' \in \mathcal{A}} r(a')$ , we implicitly assume that it exists. Now suppose by contradiction that the right-side of the conclusion does not hold, i.e.,  $\sqrt[p]{|Z \times C|} \mathcal{M}_p(a^*) > \sqrt[p]{|Z \times C|} \max_{(g,c) \in Z \times C} (C'(g, a', c))$ . It then holds that  $\mathcal{M}_p(a^*) > \mathcal{M}_p(a')$ , which contradicts the construction of  $a^*$ . Therefore, we conclude that  $\mathcal{M}_p(a^*) \leq \max_{(g,c) \in Z \times C} (C'(g, a', c))$ .

Next suppose by contradiction that the left-side of the conclusion does not hold, i.e.,  $\max_{(g,c) \in Z \times C} (C'(g, a^*, c)) > \sqrt[p]{|Z \times C|} \mathcal{M}_p(a^*)$ . It then holds that  $\mathcal{M}_p(a^*) > \sqrt[p]{|Z \times C|} \mathcal{M}_p(a^*)$ , a contradiction as  $|Z \times C| \geq 1$  and  $\sqrt[p]{|Z \times C|} \geq 1$  when  $p \geq 1$ . Therefore we conclude that  $\max_{(g,c) \in Z \times C} (C'(g, a^*, c)) \leq \sqrt[p]{|Z \times C|} \mathcal{M}_p(a^*)$ .  $\square$

## B Additional Objective-Robust Recourse Details

We detail our methodology for generating  $C^*$  and the runtime complexity of it. The steps we took are as follows:

- (1) We calculate all costs in  $\hat{C}$  and  $C$  between every pair of positively and negatively classified samples. During this step, we record the lowest-cost positive points under each objective for each negative point, i.e., we record  $|\hat{C}| \cdot |C|$  positive points for each negative point. Note that for each negative point we ignore positive points that require non-actionable changes.
- (2) For each recourse objective in  $\hat{C}$ , we calculate malfare w.r.t. each objective in  $C$  and group in  $\mathcal{Z}$ . We utilize the saved

lowest-cost positive points to quickly compute the expected value of  $\zeta$  when we calculate malfare.

- (3) Lastly, we return the  $k$  objectives in objective in  $\hat{C}$  that had the lowest malfare in the prior step.

Let  $n$  be the number of samples used to optimize the recourse method and let  $f$  and  $g$  be the worst-case complexity of the cost functions respectively in  $\hat{C}$  and  $C$ . The first part gives a runtime of  $O(|\hat{C}| \cdot n^2 \cdot f) + O(|C| \cdot n^2 \cdot g)$ . Since we saved the lowest-cost positive point under each objective for each negative point, the second part's runtime complexity is only  $O(|\hat{C}| \cdot |\mathcal{Z}| \cdot |C| \cdot n)$ . Note that, if  $n$  is large, we can sample from the dataset to compute  $C^*$ .

## C Additional Experimental Details

### C.1 DiCE and FACE Hyper-Parameters

We select  $k = 3$  for DiCE and  $\epsilon = 4$  for FACE by performing a basic grid search over those two parameters over  $[1, 5]$ . We roughly optimized for  $\ell_2$  distance, the success rate of attaining recourse, and each method's high-level objectives.

For DiCE, we searched over  $[1, 5]$ . Since we picked the best counterfactual points for each measure when analysing DiCE, having more  $k$ , i.e., options, generally improved  $\ell_2$  distance. However, too many points meant less repulsion between them.  $k = 3$  represented a balance between distance and repulsion. Any higher had minuscule  $\ell_2$  distance gains for large repulsion losses and the opposite held for lower values of  $k$ .

For FACE we searched over  $[1, 6]$  at intervals of 0.5. FACE was the only method we evaluated that could substantially fail to attain recourse for both groups. This was dependent on the  $\epsilon$  parameter. If  $\epsilon$  is too low, the graph generated by FACE is sparse and there are few paths from negatively classified points to counterfactual points. However, if  $\epsilon$  is too high, FACE reduces to (1) where cost is  $\ell_2$  distance. At  $\epsilon = 4$  we are able to attain recourse for most points while performing good w.r.t. a local outlier factor score with 15 neighbors (a plausibility objective).

### C.2 Implementation Details

We use the authors' code for DiCE [23], and use FACE and C-CHVAE implementations from Hamer et al. [12]. We implement the Wachter and Elastic Net methods following the details provided by Wachter et al. [28] for implementing (1). For data loading and preprocessing functions, we utilize the AIF360 library [4]. We use scikit-learn's implementation for the Logistic Regression classifier. We detailed the resources used to run these implementations on in Appendix D.

## D Computational Resources Used

- 8 CPU cores, 64GB RAM, NVIDIA 4070 GPU, and 2TB local SSD disk memory.
- 16 CPU cores 64GB RAM, NVIDIA 4080S GPU, and 4TB local SSD disk memory.