On the Impact of Algorithmic Recourse on Social Segregation

Ruijiang Gao [1] **Himabindu Lakkaraju** [2]

Abstract

As predictive models seep into several real-world applications, it has become critical to ensure that individuals who are negatively impacted by the outcomes of these models are provided with a means for recourse. To this end, there has been a growing body of research on algorithmic recourse in recent years. While recourses can be extremely beneficial to affected individuals, their implementation at a large scale can lead to potential data distribution shifts and other unintended consequences. However, there is little to no research on understanding the impact of algorithmic recourse after implementation. In this work, we address the aforementioned gaps by making one of the first attempts at analyzing the delayed societal impact of algorithmic recourse. To this end, we theoretically and empirically analyze the recourses output by state-of-the-art algorithms. Our analysis demonstrates that large-scale implementation of recourses by end users may exacerbate social segregation. To address this problem, we propose novel algorithms which leverage implicit and explicit conditional generative models to not only minimize the chance of segregation but also provide realistic recourses. Extensive experimentation with real-world datasets demonstrates the efficacy of the proposed approaches.

1. Introduction

Machine learning (ML) models are increasingly being deployed in domains such as finance, healthcare, and public policy to make a variety of consequential decisions. As a result, there is a growing emphasis on providing *recourse* to individuals who have been adversely impacted by the predictions of these models (Voigt & Von dem Bussche, 2017). Let us consider an individual who was denied a loan by a predictive model employed by a bank. It would be very helpful to provide a means for recourse to this individual instead of simply informing them that their loan application has been rejected. For example, this individual could be informed that if they increase their salary by $1000 and their credit score by 20 points and reapply for a loan, the bank's predictive model will approve it. Such a prescription is referred to as recourse. Several approaches in the recent literature tackled the problem of providing recourse by generating counterfactual explanations (Wachter et al., 2017; Ustun et al., 2019; Karimi et al., 2020a; Poyiadzi et al., 2020; Van Looveren & Klaise, 2019) which highlight what features need to be changed and by how much to flip a model's prediction.

Existing recourse algorithms account for various considerations when generating recourses. For instance, algorithms such as Wachter et al. (2017) aim to generate recourses that are low cost i.e., the distance between the counterfactual and the original instance is as small as possible. However, in an attempt to generate low-cost recourses, such algorithms end up outputting unrealistic counterfactual explanations which may not follow the true distribution of individuals who receive the desired outcome. To address this shortcoming, recent works attempted to find realistic recourses that adhere to the underlying data distribution by either leveraging generative models, e.g., Variational AutoEncoders (VAE) (Pawelczyk et al., 2020) or causal graphs (Karimi et al., 2020b) to ensure that the changes being requested are feasible and are inline with the data distribution.

Recent research also highlighted and addressed various challenges pertaining to the robustness (Dominguez-Olmedo et al., 2021; Slack et al., 2021; Pawelczyk et al., 2022; Rawal et al., 2021) and fairness (Gupta et al., 2019; von Kügelgen et al., 2022) of algorithmic recourse. Recourse robustness means that the counterfactual explanations should remain valid with respect to small changes in the environment. While Upadhyay et al. (2021) constructed recourses that are robust to small shifts in the underlying model, Dominguez-Olmedo et al. (2021) constructed recourses that are robust to small input perturbations. On the other hand, Gupta et al. (2019) developed an algorithm which ensures that the average recourse cost (distance between the original instance and its counterfactual) corresponding to the minority group, is not significantly worse than that of the majority group,

[1]University of Texas at Austin [2]Harvard University. Correspondence to: Ruijiang Gao <ruijiang@utexas.edu>, Himabindu Lakkaraju <hlakkaraju@hbs.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

which is the group-level recourse unfairness. While prior research on algorithmic recourse has clearly accounted for various considerations including low costs, realistic recourses, fairness and robustness of recourses, little attention has been paid to analyzing and understanding the delayed impacts of algorithmic recourse. While recourses can be extremely beneficial to affected individuals, their implementation at a large scale can lead to potential data distribution shifts and other unintended side effects such as social segregation.

In this work, we address the aforementioned gaps and make one of the first attempts at investigating the societal impacts of algorithmic recourse. More specifically, we analyze the recourse outputs by state-of-the-art algorithms empirically and theoretically in a synthetic loan lending example. Our analysis reveals that these algorithms may increase social segregation under some data distributions. For instance, most of these methods prescribe different recourses to individuals from different subgroups. To illustrate, let us consider a toy example with two groups (Figure 2) where the group 1 (e.g., male) is prescribed to increase credit scores and the group 2 (e.g., female) is asked to increase their education levels. Without loss of generality, we will use male and female as the sensitive groups throughout the paper. While state-of-the-art methods may account for the fact that it might be easier for male to increase credit scores and for female to increase education levels, such differences in prescribed recourses lead to social segregation, i.e., these groups exhibit rather distinct feature distributions (e.g., males with disproportionately higher credit scores and females with disproportionately higher education levels) after implementing prescribed recourses. Such social segregation does not only lead to differing standards w.r.t. what is required of individuals in each group to get a loan, but also has other undesirable consequences in the long term. For instance, if an external event or circumstance (e.g., global economic downturn) makes it very hard to increase credit scores during a particular year, then male will be disproportionately impacted by this and will not be able to secure loans. As pointed out in Heidari et al. (2019), segregation does not always imply the unfairness, since it can also be a result of specialization where different subpopulation intentionally invest in different qualifications while unfairness usually comes with some degree of segregation, thus it can be used as an effective means to measure potential unfairness.

To mitigate this, we propose novel recourse algorithms which ensure that implementing the prescribed recourses will result in individuals with similar feature distributions of counterfactuals across majority and minority groups. To this end, we sample counterfactuals from high density regions of feature distributions for both subgroups such that the resulting counterfactuals not only lie on the data manifold but also induce similar feature distributions for different subgroups in the long term, thereby minimizing social segregation. Since we aim to generate counterfactuals with similar feature distributions for majority and minority groups, i.e., "balanced" counterfactuals, we refer to these methods as balanced recourse algorithms. We also build two variants of balanced recourse algorithms with implicit and explicit density models for flexibility. The information of the sensitive attribute is removed from counterfactuals in implicit models using adversarial representation learning when using implicit density models such as VAE. We also conduct extensive experimentation on both synthetic and real-world datasets in credit lending, school admission and law enforcement to validate our proposed approaches. We find that existing recourse algorithms often increase social segregation and balanced recourses can reduce segregation effectively with no significant increase in recourse cost and cost disparity while remaining realistic. To the best of our knowledge, our work is the first to study the delayed societal impact of algorithmic recourse.

2. Related Work

Several approaches have been proposed in the recent literature to provide recourses to individuals who receive negative decisions from algorithms (Dhurandhar et al., 2018; Wachter et al., 2017; Ustun et al., 2019; Van Looveren & Klaise, 2019; Pawelczyk et al., 2020; Mahajan et al., 2019; Karimi et al., 2020a;b; Dandl et al., 2020). These approaches can be broadly categorized into different dimensions (Verma et al., 2020): *type of the underlying predictive model* (e.g., tree vs. differentiable classifier), *type of access* they require to the underlying predictive model (e.g., black box vs. gradient access), whether they encourage *sparsity* in counterfactuals (i.e., only a small number of features should be changed), whether counterfactuals should lie on the *data manifold*, whether the underlying *causal relationships* should be accounted for when generating counterfactuals, and whether the output produced by the method should be *multiple diverse counterfactuals* or a single counterfactual. In addition, Rawal & Lakkaraju (2020) considers how to generate global, interpretable summaries of counterfactual explanations. Some recent works also demonstrated that the recourse output by state-of-the-art techniques might not be robust, i.e., small perturbations to the original instance (Dominguez-Olmedo et al., 2021; Slack et al., 2021), the underlying model (Upadhyay et al., 2021; Rawal et al., 2021), or the recourse (Pawelczyk et al., 2022) itself may render the previously prescribed recourses invalid. These works also formulate and solve minimax optimization problems to find *robust* recourses to address the aforementioned challenges. Some research also consider fair recourse which equalizes the recourse cost required for different sensitive groups (Gupta et al., 2019; von Kügelgen et al., 2022), which is different from our setup which con-

siders the potential segregation impact in the population.

Some recent papers have studied the long-term impact of machine learning algorithms and fair interventions on the decision subjects (Liu et al., 2018; D'Amour et al., 2020; Kannan et al., 2019) where they show fair interventions may lead to undesired outcomes while our paper considers how decision subjects will be affected by algorithmic recourse methods through updating their mutable qualifications. More recently, Heidari et al. (2019) studied how decision subjects will change their qualifications through social learning (Bandura, 1962; 1978). However, their work does not consider recourse.

3. Preliminaries

Let $h : \mathcal{X} \to \mathcal{Y}$ denote a predictive model and $\mathcal{Y} = \{0, 1\}$ represents the desirable and undesirable outcome (e.g., loan approval). For an instance $x \in \mathcal{X}$, which has received an unfavorable outcome, as determined by $h(x) = 0$, the goal is to identify a set of changes that can be made to x in order to change the outcome from negative to positive. The task of modifying x requires to find a counterfactual x' which the predictive model outputs a positive label.

Recourse algorithms aim to provide the counterfactual x' such that the cost required to change x to x' is minimal and x' also flips the model prediction i.e., $h(x') = 1$. The cost can be defined as a distance metric such as l_p norm or learned from user preferences (Rawal & Lakkaraju, 2020). Some work also constrains the feature changes to conform to the underlying causal relationships (Karimi et al., 2020b) or the data manifold (Pawelczyk et al., 2020; Poyiadzi et al., 2020). We broadly categorize these algorithms as distance-based and realistic recourse algorithms.

Distance-based recourse methods: For a given distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, these methods try to minimize the distance between the original instance and the generated counterfactuals. The counterfactuals produced by these methods are typically close to the decision boundary of the classifier h. The objective employed by these methods can be written as follows(Laugel et al., 2017; Wachter et al., 2017): $\arg\min_{x'} d(x, x') \quad s.t. \quad h(x') \geq s$,

where s is the target score for x', such as 0.5.

Realistic recourse methods: Beyond the desideratum that the recourses should have low costs, several works have argued the need for recourse algorithms to generate counterfactuals that are realistic and avoid generating out-of-distribution examples. Such constraints can be framed as: $\arg\min_{x'} d(x, x') \quad s.t. \quad f(x', x) > c, \quad h(x') \geq s$,

where c is the constraint on the density of the generated counterfactual examples and f can be viewed as a realisticity constraint (e.g., density, or causal relationship). Such constraint can be realized by generating counterfactual examples by perturbing the latent dimension of a trained generative model (Pawelczyk et al., 2020).

4. Impact of Recourse on Social Segregation

To model the impact of algorithmic recourse on the underlying data distribution, we assume 1) the assumptions of each algorithmic recourse method are met and all counterfactual explanations are implementable; and 2) decision subjects will always implement the algorithmic recourse recommendations. The first assumption is used to reconcile many different assumptions made in different recourse algorithms and to show the segregation change under the optimal conditions of these methods. The second assumption corresponds to a case where human behaviors are subject to a micro-utility model with an infinite reward, it also corresponds to the worst-case segregation for each algorithm. We also assume there is a binary sensitive group. We emphasize that our behavior assumptions may not capture the real-world change perfectly, and we try to highlight a potential dynamic behavior change caused by recourse recommendations and its prospective harm in social segregation, even in such simplistic setting.

Having defined the user behavior model, we are able to investigate the societal change in terms of social segregation, a model-agnostic measure to see the change in decision subjects' qualifications. First, we use a toy example to illustrate the potential risk of existing recourse methods.

The underlying data distribution for a loan application population with two sensitive groups M and F is shown in Figure 1a. For each sensitive group, the distribution for repaying loans or default follows a uniform distribution in a two-dimensional feature space. Assume we have access to a perfect classifier $h(x)$ denoted by the blue line and aims to provide recourses to customers who will default to improve them in the future. For ease of our analysis, here we use a segregation measure defined as the cross-group distance minus the within-group distance. Formally, it is defined as $\mathbb{E}_{x_1 \sim \mathcal{X}_1, x_2 \sim \mathcal{X}_2} d(x_1, x_2) + \mathbb{E}_{x_1 \sim \mathcal{X}_1, x_2 \sim \mathcal{X}_4} d(x_1, x_2) - \mathbb{E}_{x_1 \sim \mathcal{X}_1, x_2 \sim \mathcal{X}_3} d(x_1, x_2)$, where $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4$ represent M-repay, F-repay, M-default and F-default respectively. We denote this measure as Segregation Index (SI). Intuitively, this is a clustering-like metric that measures the segregation between each neighborhood stratified by sensitive group and loan status. The adoption of SI here is for our theoretical analysis only, as we shall see later in Section 6, our conclusions also hold for more complex empirical segregation measures.

The optimal counterfactuals of distance-based and realistic methods are shown in Figure 1b and Figure 1c respectively. For distance-based recourse algorithm like Wachter et al.

(2017), the optimal resulted distribution is around the decision boundary and the optimal realistic counterfactuals like Pawelczyk et al. (2020) will be in the region with positive density and the shortest distance to the original data point. A numerical calculation estimates that the original SI is 2.90. For distance-based methods, SI increases to 3.23 and counterfactuals from realistic methods lead to SI of 3.42. Both types of algorithms lead to a large increase in SI, which indicates members in the same sensitive group may be closer in the feature space while members in different groups are relatively farther after all recourse recommendations are implemented. Formally, we show in Theorem 4.1, for such parallel rectangular-shaped uniform distributions, both types of recourse algorithms will always increase SI.

Theorem 4.1. *Assume repaying male $\mathcal{X}_1$ with (x, y) follows a uniform distribution $x \sim U[l, l+b], y \sim U[u, u+a]$; repaying female $\mathcal{X}_2$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[u, u+a]$; default male $\mathcal{X}_3$ follows $x \sim U[l, l+b], y \sim U[-u-a, -u]$; default female $\mathcal{X}_4$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[-u-a, u]$; $c < b$; $a, b, u, c > 0$. With perfect linear classifier $y = 0$ and segregation metric $d(\mathcal{X}_1, \mathcal{X}_2) + d(\mathcal{X}_1, \mathcal{X}_4) - d(\mathcal{X}_1, \mathcal{X}_1) - d(\mathcal{X}_1, \mathcal{X}_3)$ where $d(\cdot, \cdot)$ is the Hausdorff distance with Euclidean distance, the distance-based and realistic recourse algorithms will **always increase segregation** for any l, u, a, b, c.*

Remark 4.2 (Relation to Fairness Metrics). Heidari et al. (2019) shows enforcing traditional group-wise effort-based fairness constraint does not imply decreasing social segregation, which calls for methods specifically designed to reduce segregation (Figure 1b and Figure 1c are also examples where equal effort across groups may also increase segregation). As we shall see later, reducing segregation also does not imply an increase of the unfairness metric.

5. Our Framework: Balanced Recourse

To avoid the aforementioned issue, we propose two new recourse algorithms based on implicit and explicit density models by offering counterfactual explanations in the high-density regions of feature distributions of *both sensitive groups* with positive outcomes.

To see the benefit of balanced recourse, in the above example, balanced recourse offers counterfactual explanations to regions with positive densities for both male and female as shown in Figure 1d. On a high level, offering recourses that are realistic for either groups can move qualifications of different groups closer, therefore reducing social segregation. Empirically, SI reduces to 2.35 and theoretically, we show that SI will always decrease in this case.

Theorem 5.1. *Assume repaying male $\mathcal{X}_1$ with feature (x, y) follows a uniform distribution $x \sim U[l, l+b], y \sim U[u, u+a]$; repaying female $\mathcal{X}_2$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[u, u+a]$; default male $\mathcal{X}_3$ follows a uniform distribution $x \sim U[l, l+b], y \sim U[-u-a, -u]$; default female $\mathcal{X}_4$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[-u-a, u]$; $c < b$; $a, b, u, c > 0$. With perfect linear classifier $y = 0$ and segregation metric $d(\mathcal{X}_1, \mathcal{X}_2) + d(\mathcal{X}_1, \mathcal{X}_4) - d(\mathcal{X}_1, \mathcal{X}_1) - d(\mathcal{X}_1, \mathcal{X}_3)$ where $d(\cdot, \cdot)$ is the Hausdorff distance with Euclidean distance, the balanced recourse algorithm will **always decrease segregation** for any l, u, a, b, c.*

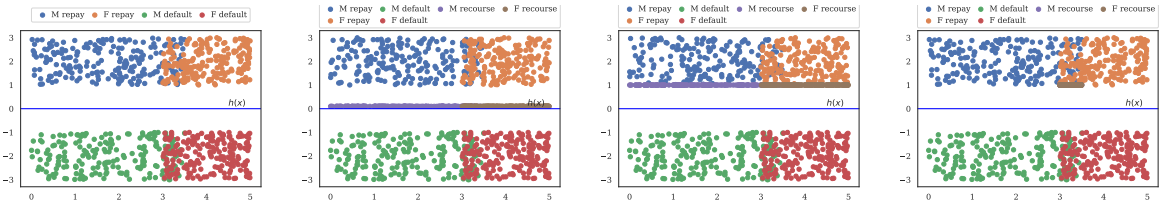
Algorithm 1 Explicit Balanced Recourse

1: Fit Conditional Density Model $P_\theta(x_f | x'_I, m, \hat{y} = 1)$.
2: For $x = (x_f, x'_I, m)$, sample x^i_f from the distribution $\propto P_\theta(x_f | x'_I, m = 1, \hat{y} = 1) P_\theta(x_f | x'_I, m = 0, \hat{y} = 1)$ by rejection sampling.
3: Accept samples with $P_\theta(x^i_f | x'_I, m = 1, \hat{y} = 1) P_\theta(x^i_f | x'_I, m = 0, \hat{y} = 1) > \beta$ and $f((x^i_f, x'_I, m)) = 1$.
4: $x'_f = \arg\min_{x^i_f} d(x, (x^i_f, x'_I, m))$
5: Return $x' = \{x'_f, x'_I, m\}$

5.1. Explicit Balanced Recourse (EBR)

Denote the sensitive group as $m \in \{0, 1\}$ and predictions as $\hat{y}$, we propose to output counterfactuals by sampling data points $\propto P(x | m = 1, \hat{y} = 1) P(x | m = 0, \hat{y} = 1)$.

To sample from the joint high-density region, we rely on a conditional density estimator such as nonparametric methods like kernel density estimation (KDE) (Sugiyama et al., 2010), normalizing flows (Trippe & Turner, 2018) or mixture density networks (MixD) (Bishop, 1994). We denote the immutable features of the instance such as age, sex, race as x_I, mutable features such as salary as x_f. At a high level, for each instance $x = (x_f, x_I)$, we estimate the conditional density of $P_\theta(x_f | x'_I, m, \hat{y} = 1)$, then sample new instances from high-density regions from $P_\theta(x^i_f | x'_I, m = 1, \hat{y} = 1) P_\theta(x^i_f | x'_I, m = 0, \hat{y} = 1)$ by rejection sampling. This would allow us to generate counterfactuals that have similar qualifications in both sensitive groups. After these samples are generated, we pick the one with the lowest recourse cost to the original instance x. We can also make the counterfactual examples to have a high density by constraining the density of x' to be greater than β. As we shall see in Section 6.4, the hyperparameter β can trade off recourse cost and social segregation. Intuitively, with the counterfactual explanations falling in higher density regions of the balanced area, the recourse cost should be higher with a better social segregation outcome. Practitioners can select β with respect to their practical need. The complete algorithm of Explicit Balanced Recourse is shown in Algorithm 1.



(a) Synthetic Data (b) Distance-based (c) Realistic (d) Balanced

Figure 1. Toy data with different recourse methods. Distance-based methods offer counterfactual explanations which are close to the decision boundary and Realistic counterfactual explanations are close to the underlying data manifold. Balanced counterfactual explanations are on the high-density region of both sensitive groups.

5.2. Implicit Balanced Recourse (IBR)

Many conditional generative models based on Generative Adversarial Network (GAN) (Goodfellow et al., 2014) and VAE (Tomczak & Welling, 2018) do not offer explicit density estimations, which cannot be used in EBR. However, we may still want to utilize these models for their high generation quality to generate realistic counterfactuals. First, we briefly introduce a VAE-based recourse algorithm CCHVAE (Pawelczyk et al., 2020).

CCHVAE: Similar to VAE, CCHVAE uses an encoder to map the mutable features x_f to a latent space z, to maximize the likelihood of mutable features given the immutables x_I, CCHVAE proposes to concatenate z with x_I to reconstruct x_f. The evidence lower bound (ELBO) can be written as $L_{CV} = \mathbb{E}_{q_\phi(z|x_f,x_I)} \log p_\theta(x_f|z,x_I) - KL(q_\phi(z|x_f,x_I)||p(z|x_I))$, then counterfactual explanations are generated by searching for the closet latent z' on a neighborhood of the encoded z from x_f, then decode x'_f from z' until a new counterfactual explanation with desired outcome is found.

To generate balanced recourses, we propose an additional regularization term inspired from adversarial representation learning (Xie et al., 2017; Roy & Boddeti, 2019) using an additional discriminator on z predicting m trained adversarially to remove the effect of the sensitive attribute on the generated samples, therefore the latent z does not contain any information about the sensitive attribute m.

The balanced CCHVAE objective can be written as

$$\min_{\eta} \max_{\theta,\phi} \mathbb{E}_{q_\phi(z|x_f)} \log p_\theta(x_f|z,x'_I) - KL(q_\phi(z|x_f)||p(z|x'_I)) - \beta \log(P_\eta(m|z)).$$

The counterfactual generation algorithm is the same as CCHVAE's, we refer to it as Implicit Balanced Recourse (IBR). To see the effect and the equilibrium of the adversarial training, theoretically it can be shown that for an auto-encoder, the optimal discriminator and decoder can be achieved given a fixed encoder. Without loss of generality, we assume $x'_I = \emptyset$ and write $x = x_f$.

Theorem 5.2. *Given a fixed deterministic encoder $z = f(x)$, the features are sampled from decoder $x' = g_\phi(x'|z)$. Training with objective $\min_\eta \max_{\theta,\phi} \mathbb{E}_{x\sim P(x), z\sim P_\theta(z|x)} \log P_\phi(x|z) - \beta \log(P_\eta(m|z))$ with sensitive attribute m, the optimal discriminator is $P_\eta(m|f(x)) = p(m|f(x))$ and the optimal decoder is $g_\phi(x|z) = p(x|z)$.*

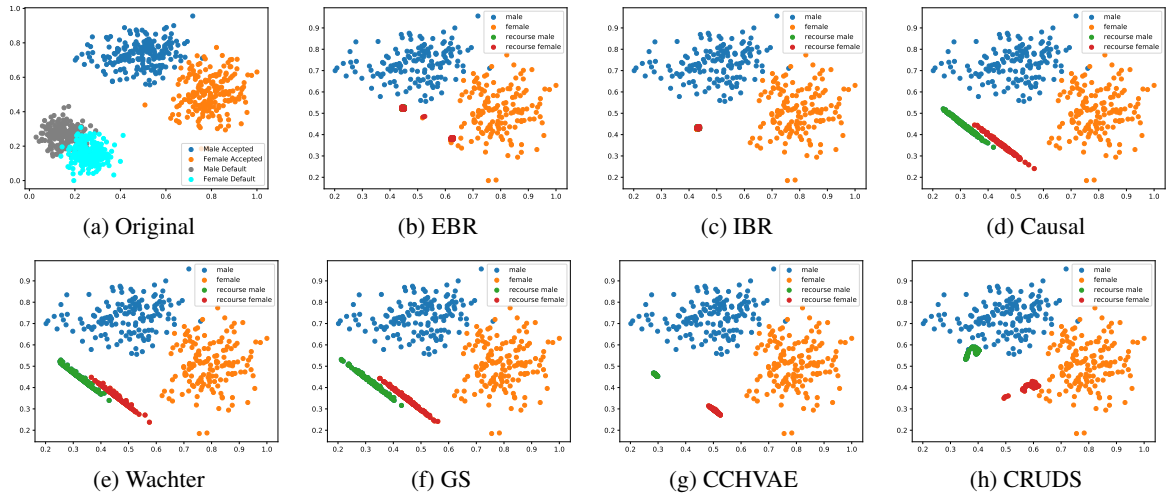
Then the optimization problem reduces to $\min_\theta \mathbb{E}_x - \log P(x|z) + \beta \log P(m|z)$, then the optimal encoder induces uniform distribution $p(m|z)$ when $x \perp\!\!\!\perp m$.

Corollary 5.3. *When $x \perp\!\!\!\perp m$, if the optimal discriminator and decoder as $p(m|z)$ and $p(x|z)$, then the optimal encoder induces uniform distribution $p(m|z)$.*

When the discriminator and latent code is a one-to-one mapping, Corollary 5.3 indicates the adversarial training can ensure that generated features induce a uniform distribution over sensitive classes. $P(x|m=0)P(x|m=1) \propto P(m=1|x)P(m=0|x)P(x)^2$ is also maximized when $p(m|z)$ is uniform, thus the implicit balanced recourse also allows us to sample from high-density regions of the unnormalized distribution $P(x|m=0)P(x|m=1)$. When $x \not\perp\!\!\!\perp m$, which means the mutable features are related to the sensitive attribute, then the optimization in encoder optimization cannot reach optimum at the same time, the relative optimality will depend on the underlying data distribution and β.

5.3. Recourse Cost

Ideally, we want the counterfactuals to have low costs so that the decision subjects can improve qualifications with minimum effort. By constraining that decision subjects with similar qualifications of different sensitive groups should have similar counterfactuals, the recourse costs are expected to be higher. However, as shown in Theorem 5.4, when the accepted distribution is already balanced, meaning all individuals from different sensitive groups have similar feature distributions, then the recourse cost of balanced recourse algorithms attains the same upper bound as realistic algorithms'. As we shall see in the experiments, balanced counterfactuals may even have lower recourse costs compared to



(a) Original (b) EBR (c) IBR (d) Causal

(e) Wachter (f) GS (g) CCHVAE (h) CRUDS

Figure 2. Counterfactuals generated by different algorithms. Blue: M - Accepted; Orange: F - Accepted; Grey: M - Default; Cyan: F - Default; Green: M - recourse; Red: F - Recourse; Balanced recourse algorithms offer low-cost counterfactuals in the overlap region of male and female distribution. Counterfactuals from Watcher, GS and Causal (with independence assumption) can be viewed as a visualization of the decision boundaries. CCHVAE and CRUDS offer counterfactuals within the data manifold, while the recourses for male and female candidates are distinct from each other.

realistic algorithms.

Theorem 5.4. *Assume male and female who repay are from a compact subspace $\mathcal{X}_1, \mathcal{X}_2$ and male and female who will default are from a compact subspace $\mathcal{X}_3, \mathcal{X}_4$ and the high-density region $\mathcal{X}' = \{x : P(x|m)P(x|1-m) > \beta\}$. Assume the recourse cost $d(\cdot,\cdot)$ is measured by l_p distance. The recourse cost of the balanced recourse cost is bounded by $d_H(\mathcal{X}_3, \mathcal{X}') + d_H(\mathcal{X}_4, \mathcal{X}')$. The recourse cost of the realistic recourse cost is bounded by $d_H(\mathcal{X}_3, \mathcal{X}_1) + d_H(\mathcal{X}_4, \mathcal{X}_2)$.*

6. Experimental Evaluation

Baselines and Methods[1] We compare the following recourse algorithms: Watcher (Wachter et al., 2017), Growing Spheres (GS) (Laugel et al., 2017), CCHVAE (Pawelczyk et al., 2020), CRUDS (Downs et al., 2020) and Causal recourse (Karimi et al., 2020b). Watcher and GS try to find the minimum-distance counterfactuals across decision boundaries, which can be viewed as distance-based recourse algorithms, while CCHVAE and CRUDS restrict the generated counterfactuals to the data manifold, which are examples of realistic recourse algorithms. Causal recourse assumes an underlying causal relationship between features. We use the default implementations of baselines in Pawelczyk et al. (2021) and use MLP as encoder and decoder architectures for VAE in CCHVAE, CRUDS and IBR. We use the Gaussian Mixture Network (Bishop, 1994) as our density model in EBR. For more details, see Appendix B.

Datasets First, we build a synthetic dataset to examine and visualize the social segregation effect of recourse algorithms by constructing a slightly more complex toy data than the toy example discussed in Section 4 by generating the features from a two-dimensional Gaussian distribution with the same four groups stratified by the sensitive attribute and default condition. Each group has 200 samples that are drawn from Gaussian distributions as shown in Figure 2a. To further investigate the potential delayed effect in practice, we use the real-world datasets German (Asuncion & Newman, 2007), Give-me-some-credit (GMC) (Kaggle, 2011), Law (Wightman, 1998) and COMPAS (Angwin et al., 2016) with sensitive attributes to evaluate existing and proposed recourse algorithms. German and GMC are used to predict whether decision subjects will default after receiving loans (shown in the main paper, results for other datasets are included in Appendix G). Law dataset is used to predict students' first-year average GPA. COMPAS is a recidivism prediction dataset to predict each decision subject will recommit crimes. In COMPAS and Law dataset, we use either sex or race as the sensitive attribute. We use sex and age as the sensitive attribute for German and GMC respectively. See Appendix C for more details about datasets.

Predictive Models We consider binary classification tasks. For the synthetic dataset, the classifier is trained on empirical data using a Logistic Regression. For real-world datasets, the classifier is MLP with `relu` activations.

Metrics To study the impact of different recourse methods, we compare various social segregation indexes, the segregation indexes are described in detail in Section 6.1. In order to understand proposed algorithms better, we also

[1] Our code is available at this URL.

Table 1. Quantitative evaluation of counterfactual explanations generated by different recourse algorithms on synthetic datasets. Results are averaged over 5 runs. Balanced recourse algorithms can effectively reduce social segregation while all other recourse methods may lead to social segregation increase in the long run. Greater yNN indicates the counterfactuals are closer to the positive data manifolds. We highlight the entry in green if the resulted segregation index is worse than the original distribution's.

	ORIGIN	IBR	EBR	CCHVAE	GS	Wachter	CRUDS	Causal
Centralization	0.4803	0.3075	0.2695	0.4953	0.4804	0.4817	0.4951	0.4804
Atkinson Index	0.9496	0.6957	0.7435	0.9971	0.9455	0.9446	0.9971	0.9638
Avg Prox	0.6409	0.7745	0.7840	0.6772	0.7012	0.7020	0.6983	0.7019
Recourse Cost	-	0.3291	0.3925	0.2448	0.2410	0.2391	0.3872	0.2430
Cost Gap	-	0.0087	0.0197	0.0166	0.0147	0.0103	0.0169	0.0144
yNN	-	0.9620	1	0.3003	0.2699	0.2871	0.9968	0.3008
# Inc. in Segregation	-	0	0	2	1	1	2	2

compare standard recourse metrics such as recourse cost, recourse cost difference between groups, invalidation rates and closeness to data manifold of these algorithms. The recourse cost difference between groups is used to measure whether minimizing segregation has effect on the usual unfairness measure in algorithmic recourse (Gupta et al., 2019; von Kügelgen et al., 2022). The closeness metric is a measure that evaluates the fraction of positively classified instances that are close to counterfactuals, which is calculated by a nearest-neighbor algorithm defined as $\text{yNN} = 1 - \frac{1}{nk}\sum_{x'}\sum_{j\in\text{KNN}(x')}|h_b(x') - h_b(x_j)|$, where $h_b = \mathbb{I}[h(x) > 0.5]$ and we choose $k = 5$ (Pawelczyk et al., 2021). A larger yNN indicates that the neighborhoods of counterfactuals are closer to positively classified instances.

A lower recourse cost is preferred since decision subjects need less effort to get the desired outcome. We use l_2-distance as the distance metric. Since many algorithms are not guaranteed to find counterfactuals for every individual, the invalidation rate measures the proportion of decision subjects who do not receive counterfactuals with desired outcomes. All results are averaged over 5 runs.

6.1. Quantifying Social Segregation

Social segregation measures how far two ethnic or racial groups are separated from each other, which is extensively studied in sociology as an indicative measure for unfairness in society (Heidari et al., 2019). Following the seminal work of Massey & Denton (1988) where they propose five dimensions of (residential) social segregation: centralization, evenness, clustering, exposure, and concentration, we overview these measures and adapt them into quantifying segregation in users' feature distributions.

Centralization measures the degree that a group is located near the center of an urban area (usually declining). Centralization Index is defined as $\sum_{i\in\text{central}} m_i/m$ where m is the total number of minorities and m_i is an indicator suggesting whether the person is a minority. **Evenness** measures how uneven minority population is distributed over the areal units. It is maximized when, in all areal units, minority and majority have the same relative number of members. For N areal units, define T as the total population, P the fraction of minority, t_i, m_i the total and minority population in area i, $p_i = m_i/t_i$, the Atkinson Index (Atkinson et al., 1970) can be defined as $1-\frac{P}{1-P}(\frac{1}{N}\sum_{i=1}^{N}(1-p_i)^{1-\beta}p_i^{\beta}t_i/TP)^{1/(1-\beta)}$, which measures the unevenness of the minority population. **Clustering** measures the degree that areal units of minority members are clustered together. **Exposure** measures the degree of potential contact or interaction between majority and minority groups. As noted in Massey & Denton (1988), exposure is related to evenness and depends on the relative sizes of the two groups. **Concentration** refers to the amount of space taken by the minority group in the urban region (Massey & Denton, 1988). If a minority group takes less space with the same size as the majority group, it is considered to be more concentrated, therefore more segregated.

Following Heidari et al. (2019), we measure centralization, evenness and clustering dimensions in our experiments. For centralization, we define the central area as an ϵ-ball around each minority member, where the neighborhood has the fraction of minority members greater than 0.5 and ϵ is set as the 20% quantile of within-group distance. For Atkinson Index, each neighborhood is determined using a K-Means clustering algorithm with 30 clusters fitted on the original data and $\beta = 0.5$, meaning minority and majority members contribute the same to segregation. Clustering is measured by the spatial proximity (White, 1986) between two groups to measure the clustering of groups in space. The average proximity is estimated as $\sum_{i=1}^{N}\sum_{j=1}^{N} m_i M_j c_{i,j}/(mM)$, where M_i is the number of majority members in neighborhood i and M is the total number of majority members. Each individual is treated as a neighborhood and $c_{i,j} = \exp(-d_{i,j})$, where $d_{i,j}$ is the distance between feature i, j. For centralization and Atkinson Index, a greater value indicates more

segregation. For average proximity, a larger value means user qualifications are less segregated.

6.2. Experimental Results with Synthetic Data

In this section, we evaluate how different recourse methods reduce segregation with the synthetic data. Qualitative results for each method are shown in Figure 2. EBR and IBR perfectly generate counterfactuals that are realistic, low-cost and balanced in groups. Distance-based recourse methods like Wachter and GS and Causal recourse with independence assumption simply route rejected instances to go across the decision boundary, forming two linear lines in the feature space, which seems to increase social segregation. Similarly, since realistic methods often generate counterfactual explanations conditioned on immutable features, it generates two distinct clusters for each group and makes two groups even more isolated from each other. While CCHVAE and IBR have a similar training and sampling procedure, with the help of adversarial training regarding the sensitive attribute, the generated counterfactuals of IBR are placed at the overlapping region of both groups.

We also quantitatively evaluate each method in terms of social segregation indexes and recourse cost in Table 1, all methods have invalidation rates of 0. If the social segregation indexes get worse after the the prescribed recourses were implemented compared to the original population's (higher for Centralization and Atkinson Index; lower for Avg Proximity), the entry is marked in green. EBR and IBR can effectively reduce social segregation, while all other methods worsen some social segregation index. Realistic algorithms may lead to worse segregation outcomes due to the conditional generation process. The balanced recourse algorithms have a higher recourse cost, while interestingly, empirically IBR even has a smaller recourse cost than CRUDS, which indicates balanced recourse algorithms have similar recourse costs compared to realistic methods'. Among all methods, realistic methods like CRUDS have higher yNN, indicates the generated counterfactuals are closer to the positive data manifolds. Similarly, IBR and EBR also have high yNN, which shows our methods can generate counterfactuals that are realistic while mitigating segregation. While distance-based algorithms offer the lowest-cost recommendation, they also have a risk of being unrealistic or non-actionable (Ustun et al., 2019). All methods have a similar recourse cost gap across groups. IBR has the lowest cost gap and EBR has the highest, while most of them have quite different segregation performance and IBR and EBR share similar segregation metrics. This confirms our previous intuition that segregation metrics are orthogonal to the group disparity in recourse costs. Balanced recourse algorithms will not necessarily increase recourse cost gap compared to other recourse algorithms. In general, balanced recourse algorithms do not introduce active harm in social segregation while still providing realistic counterfactuals.

6.3. Experimental Results with Real World Data

We further examine the potential social segregation impact brought by different recourse algorithms in a real-world context. The results are shown in Table 2, we observe in most cases, all existing recourse algorithms being evaluated will lead to some increment in social segregation (see additional results with Causal Recourse in Appendix E with similar findings). This finding is similar to Heidari et al. (2019) with a social learning dynamic. It is expected since many realistic recourse algorithms also try to find the nearest accepted samples in the underlying population, therefore the decision subjects in the same sensitive group may become closer to each other in the impacted population, which leads to greater segregation. We also include results of more datasets with similar findings in Appendix G.

By offering balanced explanations, we find that both EBR and IBR can reduce the investigated segregation indexes effectively on all datasets with different sensitive attributes while EBR has a high invalidation rate on German dataset (see Appendix D for invalidation rate results), which is probably due to the small dataset size of the dataset, which leads to poor fitting of the density models. While IBR and EBR are expected to have a higher recourse cost compared to existing recourse methods, we find that realistic algorithms such as CCHVAE and CRUDS may have a much greater recourse cost, which is consistent with our theoretical result in Theorem 5.4. At the cost of higher recourse cost, we find realistic recourse methods CCHVAE and CRUDS, along with proposed balanced recourse algorithms have a higher yNN in general across settings, which means they offer counterfactuals that are closer to the positive data manifolds, which can be considered as more realistic. Similar to our findings with synthetic data, balanced recourse algorithms do not increase the recourse cost gap and may even have a smaller recourse cost gap compared to other methods.

6.4. Ablation Study

Here we conduct ablation studies on the hyperparameters of EBR and IBR on our synthetic examples in Section 6.2 to examine their effect on social segregation and recourse cost. The results are shown in Appendix F. With a higher value of β, the counterfactual explanations generated by EBR have a higher density in the feature distribution of both sensitive groups, therefore, the recourse costs are expected to be higher, which is validated by our experimental findings. Meanwhile, we also find that increasing β in EBR in general reduces social segregation indexes. For IBR, we find that the impact of β on both social segregation and recourse costs is small, a potential reason can be that VAE uses the mean of the posterior distribution for generating new samples, which

Table 2. Quantitative evaluations of counterfactuals generated by different recourse algorithms on real datasets. Balanced recourse algorithms can effectively reduce social segregation indexes while all other recourse methods may increase social segregation in the long run. Greater value for Centralization and Atkinson Index and smaller value in Average Proximity indicate more segregation. Greater yNN indicates the counterfactuals are closer to the positive data manifolds. See Appendix G for all datasets. We highlight the entry in green if the resulted segregation index is worse than the original distribution's.

Dataset (Sen)	Metric	Origin	IBR	EBR	CCHVAE	GS	Wachter	CRUDS
German (Sex)	Centralization	0.0487	0.0452	0.0104	0.0384	0.0452	0.0803	0.1260
German (Sex)	Atkinson Index	0.1790	0.1731	0.1041	0.2342	0.1754	0.1965	0.2247
German (Sex)	Avg Prox	0.3051	0.3769	0.3357	0.2974	0.2592	0.2726	0.4548
German (Sex)	Recourse Cost	-	1.2402	1.4853	1.0995	1.2575	0.5123	1.7954
German (Sex)	Cost Gap	-	0.1404	0.1258	0.1374	0.1422	0.0874	0.0718
German (Sex)	yNN	-	0.2142	0.1439	0.1690	0.2094	0.0888	0.0562
GMC(Age)	Centralization	0.3414	0.3314	0.3374	0.3331	0.3411	0.3417	0.3320
GMC(Age)	Atkinson Index	0.1218	0.1193	0.1216	0.1248	0.1220	0.1218	0.1252
GMC(Age)	Avg Prox	0.5820	0.5902	0.5853	0.5884	0.5826	0.5825	0.5895
GMC(Age)	Recourse Cost	-	0.6335	0.3565	0.5156	0.1483	0.1606	0.6272
GMC(Age)	Cost Gap	-	0.0596	0.1082	0.0810	0.0240	0.0203	0.0780
GMC(Age)	yNN	-	0.9741	0.9541	0.9623	0.5265	0.5056	1
	# Inc. in Segregation	-	0	0	3	2	4	3

is always the highest density region. The recourse cost gap also remains small with varying β.

7. Conclusions and Future Work

In this work, we made one of the first attempts at analyzing the societal impact of algorithmic recourse. To this end, we theoretically and empirically analyzed the recourses output by state-of-the-art algorithms and demonstrated that large-scale implementation of recourses by end users may lead to social segregation in the long term. To address this problem, we proposed novel algorithms which leverage implicit and explicit conditional generative models to not only minimize the chance of segregation but also to provide realistic recourses. We also carried out extensive empirical analysis to establish the efficacy of the proposed algorithms. Our work paves the way for several interesting future research directions. For instance, it is unlikely that recourses will be implemented with full compliance in practice. It would be interesting to study the societal impacts of recourses with partial compliance on social segregation.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. In *Ethics of Data and Analytics*, pp. 254–264. Auerbach Publications, 2016.

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Atkinson, A. B. et al. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

Bandura, A. Social learning through imitation. *Nebraska Symposium on Motivation*, 1962.

Bandura, A. Social learning theory of aggression. *Journal of communication*, 28(3):12–29, 1978.

Bishop, C. M. Mixture density networks. Technical report, Aston University, 1994.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pp. 448–469. Springer, 2020.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Dominguez-Olmedo, R., Karimi, A.-H., and Schölkopf, B. On the adversarial robustness of causal algorithmic recourse. *arXiv:2112.11313*, 2021.

Downs, M., Chu, J. L., Yacoby, Y., Doshi-Velez, F., and Pan, W. Cruds: Counterfactual recourse using disentangled subspaces. *ICML WHI*, 2020:1–23, 2020.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.

Gupta, V., Nokhiz, P., Roy, C. D., and Venkatasubramanian, S. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

Heidari, H., Nanda, V., and Gummadi, K. P. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *arXiv preprint arXiv:1903.01209*, 2019.

Kaggle. Give me some credit, kaggle challenge, 2011.

Kannan, S., Roth, A., and Ziani, J. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 240–248, 2019.

Karimi, A.-H., Barthe, G., Balle, B., and Valera, I. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020a.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., and Valera, I. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020b.

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Mahajan, D., Tan, C., and Sharma, A. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Massey, D. S. and Denton, N. A. The dimensions of residential segregation. *Social forces*, 67(2):281–315, 1988.

Pawelczyk, M., Broelemann, K., and Kasneci, G. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pp. 3126–3132, 2020.

Pawelczyk, M., Bielawski, S., Van den Heuvel, J., Richter, T., and Kasneci, G. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *Advances in Neural Information Processing Systems (NeurIPS) (Benchmark and Datasets Track)*, volume 34, 2021.

Pawelczyk, M., Datta, T., van-den Heuvel, J., Kasneci, G., and Lakkaraju, H. Algorithmic recourse in the face of noisy human responses. *arXiv preprint arXiv:2203.06768*, 2022.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

Rawal, K. and Lakkaraju, H. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, 33:12187–12198, 2020.

Rawal, K., Kamar, E., and Lakkaraju, H. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv:2012.11788*, 2021.

Roy, P. C. and Boddeti, V. N. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2586–2594, 2019.

Slack, D., Hilgard, S., Lakkaraju, H., and Singh, S. Counterfactual explanations can be manipulated. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., and Okanohara, D. Conditional density estimation via Least-Squares density ratio estimation. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 781–788, Chia Laguna Resort, Sardinia, Italy, 2010. PMLR.

Tomczak, J. and Welling, M. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223. PMLR, 2018.

Trippe, B. L. and Turner, R. E. Conditional density estimation with bayesian normalising flows. February 2018.

Upadhyay, S., Joshi, S., and Lakkaraju, H. Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

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

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

Verma, S., Dickerson, J., and Hines, K. Counterfactual explanations for machine learning: A review. *arXiv:2010.10596*, 2020.

Voigt, P. and Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676): 10–5555, 2017.

von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., and Schölkopf, B. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9584–9594, 2022.

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

White, M. J. Segregation and diversity measures in population distribution. *Population index*, pp. 198–221, 1986.

Wightman, L. F. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Xie, Q., Dai, Z., Du, Y., Hovy, E., and Neubig, G. Controllable invariance through adversarial feature learning. *Advances in neural information processing systems*, 30, 2017.

A. Proofs

Theorem 4.1. Assume repaying male $\mathcal{X}_1$ with feature (x, y) follows a uniform distribution $x \sim U[l, l+b], y \sim U[u, u+a]$; repaying female $\mathcal{X}_2$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[u, u+a]$; default male $\mathcal{X}_3$ follows a uniform distribution $x \sim U[l, l+b], y \sim U[-u-a, -u]$; default female $\mathcal{X}_4$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[-u-a, u]$; $c < b$;$a, b, u, c > 0$. With perfect linear classifier $y = 0$ and segregation metric $d(\mathcal{X}_1, \mathcal{X}_2) + d(\mathcal{X}_1, \mathcal{X}_4) - d(\mathcal{X}_1, \mathcal{X}_1) - d(\mathcal{X}_1, \mathcal{X}_3)$ where $d(\cdot, \cdot)$ is the Hausdorff distance with Euclidean distance, the distance-based and realistic-based recourse algorithms will **always increase segregation** for any l, u, a, b, c.

Proof of Theorem 4.1:

Proof. Hausdorff distance is defined as $d_H(X, Y) = \max\{\sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y)\}$, where $d(x, Y) = \inf_{y \in Y} d(x, y)$. Here we define $d(\cdot, \cdot)$ as Euclidean distance.

Original Segregation Index:

$d(\mathcal{X}_1, \mathcal{X}_2) = b - c$; $d(\mathcal{X}_1, \mathcal{X}_1) = 0$; $d(\mathcal{X}_1, \mathcal{X}_3) = 2\mu + a$; $d(\mathcal{X}_1, \mathcal{X}_4) = \sqrt{(b-c)^2 + (2\mu + a)^2}$;

Distance-based Segregation Index: $d(\mathcal{X}_1, \mathcal{X}_2) = b - c$; $d(\mathcal{X}_1, \mathcal{X}_1) = 0$; $d(\mathcal{X}_1, \mathcal{X}_3) = \mu + a$; $d(\mathcal{X}_1, \mathcal{X}_4) = \sqrt{(b-c)^2 + (\mu + a)^2}$;

Realistic Segregation Index: $d(\mathcal{X}_1, \mathcal{X}_2) = b - c$; $d(\mathcal{X}_1, \mathcal{X}_1) = 0$; $d(\mathcal{X}_1, \mathcal{X}_3) = a$; $d(\mathcal{X}_1, \mathcal{X}_4) = \sqrt{(b-c)^2 + a^2}$;

$\text{SI} = d(\mathcal{X}_1, \mathcal{X}_2) + d(\mathcal{X}_1, \mathcal{X}_4) - d(\mathcal{X}_1, \mathcal{X}_1) - d(\mathcal{X}_1, \mathcal{X}_3)$

Thus $\text{SI}_{\text{original}} - \text{SI}_{\text{dist}} = -\mu + \sqrt{(b-c)^2 + (2\mu + a)^2} - \sqrt{(b-c)^2 + (\mu + a)^2}$

$$\text{SI}_{\text{original}} - \text{SI}_{\text{dist}} < 0 \rightleftharpoons -\mu + \sqrt{(b-c)^2 + (2\mu + a)^2} < \sqrt{(b-c)^2 + (\mu + a)^2} \tag{1}$$

$$\rightleftharpoons 2\mu + a < \sqrt{(b-c)^2 + (2\mu + a)^2} \tag{2}$$

$\text{SI}_{\text{original}} - \text{SI}_{\text{real}} = -2\mu + \sqrt{(b-c)^2 + (2\mu + a)^2} - \sqrt{(b-c)^2 + a^2}$

$$\text{SI}_{\text{original}} - \text{SI}_{\text{real}} < 0 \rightleftharpoons -2\mu + \sqrt{(b-c)^2 + (2\mu + a)^2} < \sqrt{(b-c)^2 + a^2} \tag{3}$$

$$\rightleftharpoons 2\mu + a < \sqrt{(b-c)^2 + (2\mu + a)^2} \tag{4}$$

which concludes the proof.

□

Theorem 5.1. Assume repaying male $\mathcal{X}_1$ with feature (x, y) follows a uniform distribution $x \sim U[l, l+b], y \sim U[u, u+a]$; repaying female $\mathcal{X}_2$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[u, u+a]$; default male $\mathcal{X}_3$ follows a uniform distribution $x \sim U[l, l+b], y \sim U[-u-a, -u]$; default female $\mathcal{X}_4$ follows $x \sim U[l+b-c, l+2b-c], y \sim U[-u-a, u]$; $c < b$; $a, b, u, c > 0$. With perfect linear classifier $y = 0$ and segregation metric $d(\mathcal{X}_1, \mathcal{X}_2) + d(\mathcal{X}_1, \mathcal{X}_4) - d(\mathcal{X}_1, \mathcal{X}_1) - d(\mathcal{X}_1, \mathcal{X}_3)$ where $d(\cdot, \cdot)$ is the Hausdorff distance with Euclidean distance, the balanced recourse algorithm will **always decrease segregation** for any l, u, a, b, c.

Proof of Theorem 5.1:

Proof. **Original Segregation Index:**

$d(\mathcal{X}_1, \mathcal{X}_2) = b - c$; $d(\mathcal{X}_1, \mathcal{X}_1) = 0$; $d(\mathcal{X}_1, \mathcal{X}_3) = 2\mu + a$; $d(\mathcal{X}_1, \mathcal{X}_4) = \sqrt{(b-c)^2 + (2\mu + a)^2}$;

Balanced Segregation Index:

$d(\mathcal{X}_1, \mathcal{X}_2) = b - c$; $d(\mathcal{X}_1, \mathcal{X}_1) = 0$; $d(\mathcal{X}_1, \mathcal{X}_3) = b - c$; $d(\mathcal{X}_1, \mathcal{X}_4) = b - c$;

$\text{SI} = d(\mathcal{X}_1, \mathcal{X}_2) + d(\mathcal{X}_1, \mathcal{X}_4) - d(\mathcal{X}_1, \mathcal{X}_1) - d(\mathcal{X}_1, \mathcal{X}_3)$

$\text{SI}_{\text{original}} - \text{SI}_{\text{balanced}} = -(2\mu + a) + \sqrt{(b-c)^2 + (2\mu + a)^2} > 0.$ □

Theorem 5.2. Given a fixed deterministic encoder $z = f(x)$, the features are sampled from decoder $x' = g_\phi(x'|z)$. Training with objective $\min_\eta \max_{\theta,\phi} \mathbb{E}_{x\sim P(x), z\sim P_\theta(z|x)} \log P_\phi(x|z) - \beta \log(P_\eta(m|z))$ with sensitive attribute m, the optimal discriminator is $P_\eta(m|f(x)) = p(m|f(x))$ and the optimal decoder is $g_\phi(x|z) = p(x|z)$.

Proof of Theorem 5.2:

Proof. **Discriminator:** The objective of the discriminator is

$$V_D = \mathbb{E}_x - \log(P_\eta(m|z)) = -\sum_{x,m} p(x,m)\log(P_\eta(m|z)) \quad s.t. \sum_m P_\eta(m|z) = 1, P_\eta(m|z) \geq 0$$

The Lagrangian dual can be written as

$$L_D = -\sum_{x,m} p(x,m)\log(P_\eta(m|z)) + \sum_z \lambda_z(\sum_m P_\eta(m|z) - 1) \tag{5}$$

Taking gradient with respect to $P_\eta(m|z)$, we have

$$\lambda_z P_\eta(m|z) = p(m,z) \tag{6}$$

Since $\sum_m P_\eta(m|z) = 1$, $\lambda_z = p(z)$, then we get $P_\eta(m|z) = P(m|z)$.

Decoder: The objective of decoder is

$$V_{Dec} = \mathbb{E}_x - \log(P_\phi(x|z)) = -\sum_{x,m,z} p(x,m,z)\log(P_\phi(x|z)) \quad s.t. \sum_x P_\phi(x|z) = 1, P_\phi(x|z) \geq 0$$

The Lagrangian dual can be written as

$$L_{Dec} = -\sum_{x,m,z} p(x,m,z)\log(P_\phi(x|z)) + \sum_z \lambda_z(\sum_x P_\phi(x|z) - 1) \tag{7}$$

Taking the gradient with respect to $P_\phi(x|z)$, we have

$$\lambda_z P_\phi(x|z) = p(x,z) \tag{8}$$

Since $\sum_x P_\phi(x|z) = 1$, $\lambda_z = p(z)$, then we get $P_\phi(x|z) = P(x|z)$.

□

Corollary 5.3. *When $x \perp\!\!\!\perp m$, if the optimal discriminator and decoder as $p(m|z)$ and $p(x|z)$, then the optimal encoder induces uniform distribution $p(m|z)$.*

Proof of Corollary 5.3:

Proof. The objective is

$$\min_\theta \mathbb{E}_{x,z=f_\theta(x)} - \log(P(x|z)) + \log(P(m|z)) \tag{9}$$
$$s.t. \sum_m P(m|z) = 1, P(m|z) \geq 0$$

Then Lagrangian can be written as

$$\mathbb{E}_{x,z=f_\theta(x)} - \log(P(x|z)) + \log(P(m|z)) - \lambda(\sum_m P(m|z) - 1) \tag{10}$$

when $x \perp\!\!\!\perp m$, $P(x|z)$ is independent of $P(m|z)$, then we can take the gradient w.r.t. $P(m|z)$, we have $\lambda P(m|z) = 1$, since $\sum P(m|z) = 1$, then $\lambda = m$, thus $P(m|z) = 1/m$. □

Theorem 5.4. Assume male and female who will repay are from a compact subspace $\mathcal{X}_1, \mathcal{X}_2$ and male and female who will default are from a compact subspace $\mathcal{X}_3, \mathcal{X}_4$ and the high-density region $\mathcal{X}' = \{x : P(x|m)P(x|1-m) > \beta\}$. Assume the recourse cost $d(\cdot,\cdot)$ is measured by l_p distance. The recourse cost of the balanced recourse cost is bounded by $d_H(\mathcal{X}_3, \mathcal{X}') + d_H(\mathcal{X}_4, \mathcal{X}')$. The recourse cost of the realistic recourse cost is bounded by $d_H(\mathcal{X}_3, \mathcal{X}_1) + d_H(\mathcal{X}_4, \mathcal{X}_2)$.

Proof of Theorem 5.4:

Proof. The recourse cost r for balanced algorithm can be written as

$$\mathbb{E}_{x\sim\mathcal{X}_3, y=\arg\min_{y\in\mathcal{X}'} d(x,y)} d(x,y) + \mathbb{E}_{x\sim\mathcal{X}_4, y=\arg\min_{y\in\mathcal{X}'} d(x,y)} d(x,y) \tag{11}$$

$$\leq d_H(\mathcal{X}_3, \mathcal{X}') + d_H(\mathcal{X}_4, \mathcal{X}') \tag{12}$$

The recourse cost r for realistic algorithm can be written as

$$\mathbb{E}_{x\sim\mathcal{X}_3, y=\arg\min_{y\in\mathcal{X}_1} d(x,y)} d(x,y) + \mathbb{E}_{x\sim\mathcal{X}_4, y=\arg\min_{y\in\mathcal{X}_2} d(x,y)} d(x,y) \tag{13}$$

$$\leq d_H(\mathcal{X}_3, \mathcal{X}_1) + d_H(\mathcal{X}_4, \mathcal{X}_2) \tag{14}$$

It is easy to see when $\mathcal{X}_1$ and $\mathcal{X}_2$ are already balanced ($\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}'$), realistic recourse and balanced recourse attain the same upper bound in terms of recourse cost. □

B. Recourse Methods

Here we offer an overview of the recourse methods used in the paper. We use the same notation in our main paper and the implementations in Pawelczyk et al. (2021).

Wachter (Wachter et al., 2017): Wachter solves the following optimization problem using :

$$\arg\min_{x'} d(x, x') \quad s.t. \quad h(x') \geq s, \tag{15}$$

where x is the original instance, then this optimization problem can be written as

$$\arg\min_{x'} l(h(x'), s) + \lambda d(x, x'), \tag{16}$$

λ is the Lagrangian multiplier and l is a loss function such as mean squared error. Since Wachter cannot preserve the immutable features, we manually hard-code this constraint to counterfactual explanations generated by Wachter. λ is set as 0.01 in the experiments.

Growing Spheres (GS) (Laugel et al., 2017): GS uses a random search algorithm to search the closet counterfactuals labels using growing hyperspheres. If the input feature is binary, GS uses bernoulli sampling to search for counterfactual explanations. Unlike Wachter, GS can keep the immutable features by forbidding them to change in the search procedure. The step size of the growing hyperspheres is set as 0.2.

CCHVAE (Pawelczyk et al., 2020): Similar to variational auto-encoder (VAE), CCHVAE uses an encoder to map the mutable features x_f to a latent space z, to maximize the likelihood of mutable features given the immutables x_m, CCHVAE proposes to concatenate z with x_m to reconstruct x_f. The evidence lower bound (ELBO) can be written as

$$L_{CV} = \mathbb{E}_{q(z|x_m)} \log p(x_f|z, x_I) - KL(q(z|x_m, x_I)||p(z|x_m)),$$

then counterfactual explanations are generated by searching for the closet latent z' on a neighborhood of the encoded z from x_f, then decode x'_f from z' until a new counterfactual explanation with desired outcome is found.

CRUDS (Downs et al., 2020): Similar to CCHVAE, CRUDS leverages VAE to capture realisticity. CRUDS uses a Conditional Subspace VAE (CSVAE) to disentangle the latent space into w that is predictive of the label y and z that does not affect the outcome. Given a instance x, samples are drawn from the posterior distribution $z \sim q_\phi(z|x)$ and $w \sim p_\theta(w|y=1)$, the counterfactual explanations are finally given by $x' \sim g_\eta(w, z)$, where g_η is a decoder which can be parameterized by a Gaussian distribution as in Downs et al. (2020).

Causal Recourse (Karimi et al., 2020b): Causal recourse assumes that the features follow a known causal relationship provided by a causal graph. For synthetic data, we use the implementation in the CARLA package(Pawelczyk et al., 2021) and assume feature independence (since our toy example cannot be represented by known structured equations with additive noise). For the german credit dataset, we use the official implementation in `https://github.com/amirhk/recourse` and use the causal graph provided in Karimi et al. (2020b). The experimental results for german credit are included in Appendix E separately since Karimi et al. (2020b) only used 7 features in the experiments.

C. Dataset

Here we discuss the dataset statistics and features we used in the paper.

German (Asuncion & Newman, 2007): We use sex as the sensitive attribute. The dataset includes 1000 samples, 12 immutable features such as sex, age, purpose of the loan, 8 mutable features including number of people being liable to provide maintenance for, installment rate, employment time, job qualification, credit amount. The prediction target is whether the customer has a good or bad credit risk.

Give-me-some-credit (GMC) (Kaggle, 2011): We use age as the sensitive attribute. The dataset includes 115527 samples, 1 immutable feature age, and 9 mutable features such as the monthly income, debt ratio, number of times that payment is 90 days late, number of real estate loans, revolving utilization of unsecured lines. We sample 10000 samples per run due to the high computation cost of some social segregation indexes. The prediction target is whether the customer will experience 90 days past due delinquency or worse.

Law (Wightman, 1998): We use either sex or race as the sensitive attribute. The dataset includes 21791 samples, 3 immutable features including sex, race, and region, 4 mutable features including SAT score, prior GPA, sander index and first-year average GPA. We sample 10000 samples per run due to the high computation cost of some social segregation indexes. The prediction target is whether the student passes bar exam.

COMPAS (Angwin et al., 2016): COMPAS is a recidivism prediction dataset to predict each decision subject will recommit crimes. We use either sex or race as the sensitive attribute. The dataset includes 6172 samples, 4 immutable features including age, sex, race, and charge degree, 3 mutable features including two-year recidivism, number of prior counts, and length of stay. The prediction target is whether the suspect has a high recidivism risk.

D. Experiment Results for Invalidation Rates

We include the invalidation results here due to the space constraint. The results are shown in Table 3. All methods receives low invalidation rates in general except Wachter since it does not consider immutable constraint. Since German dataset size is small, we observe all realistic methods except GS start to have invalidated recourse recommendations, which can be explained due to the difficulty of fitting density models given limited data.

Table 3. Invalidation Rate for Different Recourse Methods.

	IBR	EBR	CCHVAE	GS	Wachter	CRUDS
German (Sex)	0.0013	0.3385	0.0013	0	0.9893	0.8216
GMC (Age)	0	0	0	0	0	0
Law (Sex)	0	0	0	0	0	0
Law (Race)	0	0	0	0	0	0
COMPAS (Sex)	0	0	0	0	0.8448	0
COMPAS (Race)	0	0	0	0	0.8448	0

E. Additional Experiment Results for Causal Recourse

In addition, we examine whether Causal Recourse may increase social segregation. We use the causal graph and default implementation in Karimi et al. (2020b) on german dataset and treat sex as the sensitive attribute. The features used include age, education, loan amount, loan duration, income and savings. The classifier class is chosen to be MLP.

The results are shown in Table 4. Similarly, we observe Causal Recourse may also increase social segregation, which is

expected since it does not consider the potential user behavior changes.

Table 4. Results for Causal Recourse on German Dataset. Causal recourse algorithm also leads to more segregation.

German (Sex)	Centralization	Atkinson Index	Avg Prox	Recourse Cost	Cost Gap	Invalidation Rate
Origin	0.2700	0.0544	0.0032	-	-	-
Causal Recourse	0.2848	0.0548	0.0029	0.0846	0.0243	0

F. Experiment Results for Ablation Studies

We include our results for ablation studies in Table 5. Here we conduct ablation studies on the hyperparameters of EBR and IBR on our synthetic examples in Section 6.2 to examine their effect on social segregation and recourse cost. With a higher value of β, the counterfactual explanations generated by EBR have a higher density in the feature distribution of both sensitive groups, therefore, the recourse costs are expected to be higher, which is validated by our experimental findings. Meanwhile, we also find that increasing β in EBR in general reduces social segregation indexes. For IBR, we find that the impact of β on both social segregation and recourse costs is small, a potential reason can be that VAE uses the mean of the posterior distribution for generating new samples, which is always the highest density region. The recourse cost gap also remains small with varying β.

Table 5. Ablation study on hyperparameters of EBR and IBR. β in EBR can affect social segregation and recourse cost of generated recourses in opposite ways while counterfactual explanations from IBR are more stable with respect to β.

EBR	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	IBR	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 100$
Centralization	0.4836	0.2695	0.3068	0.3068	Centralization	0.3075	0.3075	0.3075	0.3075
Atkinson Index	0.7950	0.7435	0.7402	0.6914	Atkinson Index	0.6957	0.6957	0.6926	0.6975
Avg Prox	0.7524	0.7840	0.7935	0.8042	Avg Prox	0.7747	0.7745	0.7753	0.7746
Recourse Cost	0.3652	0.3925	0.4114	0.4470	Recourse Cost	0.3297	0.3291	0.3320	0.3292
Cost Gap	0.0238	0.0197	0.0153	0.0123	Cost Gap	0.0095	0.0087	0.0100	0.0088
yNN	1	1	1	1	yNN	0.9465	0.9620	0.9620	0.9620

G. Experiment Results for All Datasets

Here, we include results for all datasets in Table 6. Some datasets are omitted in the main paper due to space constraint. Similarly, we observe existing recourse methods may increase social segregation on most settings and our proposed recourse methods generate realistic counterfactuals (with high yNN) that can reduce segregation indexes effectively. In few case like Law with race as the sensitive attribute, all recourse methods do not increase segregation. There is no clear relationship between the recourse cost gap and segregation, which confirms our intuition that they are orthogonal notions of unfairness and future work can consider how to optimize them jointly.

Table 6. Quantitative evaluations of counterfactual explanations generated by different recourse algorithms on Real Datasets. Results are averaged over 5 runs. Balanced recourse algorithms can effectively reduce social segregation indexes while all other recourse methods may increase social segregation in the long run. Greater value for Centralization and Atkinson Index and smaller value in Average Proximity indicate more segregation. Greater yNN indicates the counterfactuals are closer to the positive data manifolds.

Dataset (Sen)	Metric	Origin	IBR	EBR	CCHVAE	GS	Wachter	CRUDS
German (Sex)	Centralization	0.0487	0.0452	0.0104	0.0384	0.0452	0.0803	0.1260
German (Sex)	Atkinson Index	0.1790	0.1731	0.1041	0.2342	0.1754	0.1965	0.2247
German (Sex)	Avg Prox	0.3051	0.3769	0.3357	0.2974	0.2592	0.2726	0.4548
German (Sex)	Recourse Cost	-	1.2402	1.4853	1.0995	1.2575	0.5123	1.7954
German (Sex)	Cost Gap	-	0.1404	0.1258	0.1374	0.1422	0.0874	0.0718
German (Sex)	yNN	-	0.2142	0.1439	0.1690	0.2094	0.0888	0.0562
GMC(Age)	Centralization	0.3414	0.3314	0.3374	0.3331	0.3411	0.3417	0.3320
GMC(Age)	Atkinson Index	0.1218	0.1193	0.1216	0.1248	0.1220	0.1218	0.1252
GMC(Age)	Avg Prox	0.5820	0.5902	0.5853	0.5884	0.5826	0.5825	0.5895
GMC(Age)	Recourse Cost	-	0.6335	0.3565	0.5156	0.1483	0.1606	0.6272
GMC(Age)	Cost Gap	-	0.0596	0.1082	0.0810	0.0240	0.0203	0.0780
GMC(Age)	yNN	-	0.9741	0.9541	0.9623	0.5265	0.5056	1
Law (Sex)	Centralization	0.0858	0.0694	0.0805	0.0743	0.0829	0.0833	0.0932
Law (Sex)	Atkinson Index	0.0319	0.0314	0.0318	0.0345	0.0318	0.0326	0.0387
Law (Sex)	Avg Prox	0.7027	0.7177	0.7076	0.7173	0.7055	0.7070	0.7168
Law (Sex)	Recourse Cost	-	0.3945	0.1334	0.3914	0.0974	0.1168	0.5384
Law (Sex)	Cost Gap	-	0.0231	0.0111	0.0302	0.0078	0.0092	0.0188
Law (Sex)	yNN	-	1	0.7098	1	0.5475	0.6231	1
Law (Race)	Centralization	0.0288	0.0028	0.0137	0.0025	0.0245	0.0175	0.0020
Law (Race)	Atkinson Index	0.2512	0.1939	0.2418	0.2374	0.2436	0.2401	0.2130
Law (Race)	Avg Prox	0.6660	0.7103	0.6833	0.7035	0.6739	0.6782	0.7111
Law (Race)	Recourse Cost	-	0.4308	0.1699	0.3914	0.0977	0.1168	0.4629
Law (Race)	Cost Gap	-	0.0301	0.0206	0.1690	0.0402	0.0444	0.1064
Law (Race)	yNN	-	1	0.8422	1	0.5510	0.6231	1
COMPAS (Sex)	Centralization	0.0020	0.0018	0.0020	0.0009	0.0025	0.0027	0.0057
COMPAS (Sex)	Atkinson Index	0.0581	0.0478	0.0537	0.0521	0.0508	0.0522	0.0500
COMPAS (Sex)	Avg Prox	0.6371	0.6511	0.6484	0.6474	0.6409	0.6405	0.6545
COMPAS (Sex)	Recourse Cost	-	0.5114	0.2642	0.2243	0.1247	0.0958	0.6302
COMPAS (Sex)	Cost Gap	-	0.0274	0.0535	0.0513	0.0227	0.0150	0.1760
COMPAS (Sex)	yNN	-	0.8792	0.9668	0.8786	0.3233	0.1860	1
COMPAS (Race)	Centralization	0.2947	0.2846	0.2943	0.2892	0.2986	0.3010	0.3054
COMPAS (Race)	Atkinson Index	0.0709	0.0563	0.0589	0.0656	0.0627	0.0654	0.0500
COMPAS (Race)	Avg Prox	0.6236	0.6372	0.6358	0.6366	0.6288	0.6281	0.6435
COMPAS (Race)	Recourse Cost	-	0.2331	0.2225	0.2243	0.1243	0.0958	0.6302
COMPAS (Race)	Cost Gap	-	0.0287	0.0589	0.0832	0.0129	0.0130	0.0931
COMPAS (Race)	yNN	-	0.8743	0.9859	0.8786	0.3233	0.1860	1
	# Inc. in Segregation	-	0	0	4	4	7	7