Options Without Agency: Diversity as a Requirement for Fair Actionable Recourse

PETER M. VANNOSTRAND, Worcester Polytechnic Institute, USA
DENNIS M. HOFMANN, Worcester Polytechnic Institute, USA
LEI MA, Worcester Polytechnic Institute, USA
ELKE A. RUNDENSTEINER, Worcester Polytechnic Institute, USA

Machine learning systems are increasingly being used to automate life-changing decisions in domains such as finance and recruitment, motivating the development of a myriad of explainable AI techniques. Among them, counterfactual explanations are widely promoted for enabling recourse by suggesting actions that individuals could take to change an unfavorable outcome. Despite their growing adoption, the fairness implications of counterfactual-based recourse remain underexplored. Existing notions of fairness in recourse focus primarily on the number of counterfactuals provided or the magnitude of the changes required to enact each counterfactual to achieve recourse. In this work, we argue that these notions alone are insufficient. Instead, we posit that the diversity of the provided counterfactuals is critical to ensuring that individuals are given meaningful and equitable choices for altering their outcomes. To achieve this, we introduce and analyze a quantitative metric for measuring diversity in recourse on both individual and group levels. Through extensive evaluation across multiple datasets and model architectures, we demonstrate that recourse diversity reliably captures fairness implications not considered by existing fairness metrics and can serve as a valuable signal in guiding model design and selection. This suggests that diversity constitutes an important complementary dimension for assessing fairness in actionable recourse. To facilitate examinations of recourse fairness by the community, we release our auditing tools on GitHub as an open-source framework.

CCS Concepts: • **Human-centered computing** → **User studies**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Fairness, Recourse, Counterfactual Explanation, Explainable AI, Algorithmic Justice

ACM Reference Format:
Peter M. VanNostrand, Dennis M. Hofmann, Lei Ma, and Elke A. Rundensteiner. 2026. Options Without Agency: Diversity as a Requirement for Fair Actionable Recourse. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3805689.3812337

1 Introduction

Motivation. Machine learning systems are widely used to fully or partially automate decision-making in consequential domains such as finance [11, 35], recruitment [19, 29], housing [18], policing [2, 24], and healthcare [38]. In such high-stakes domains, negative decision outcomes can have significant adverse consequences for affected individuals, often resulting in the loss of access to life-changing resources or critical opportunities. While we should critically assess whether such systems are appropriate for specific tasks, their widespread adoption has increasingly led to regulatory requirements – mandating that affected individuals be provided with explanations of how decisions were made [4, 11, 16]. Regulations generally argue for transparency and, when possible, options for *recourse* which allow the decision subject to take actions that alter the decision outcome [50, 55].

Authors' Contact Information: Peter M. VanNostrand, pvannostrand@wpi.edu, Worcester Polytechnic Institute, Worcester, USA; Dennis M. Hofmann, dmhofmann@wpi.edu, Worcester Polytechnic Institute, Worcester, USA; Lei Ma, lma5@wpi.edu, Worcester Polytechnic Institute, Worcester, USA; Elke A. Rundensteiner, rundenst@wpi.edu, Worcester Polytechnic Institute, Worcester, USA .

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812337

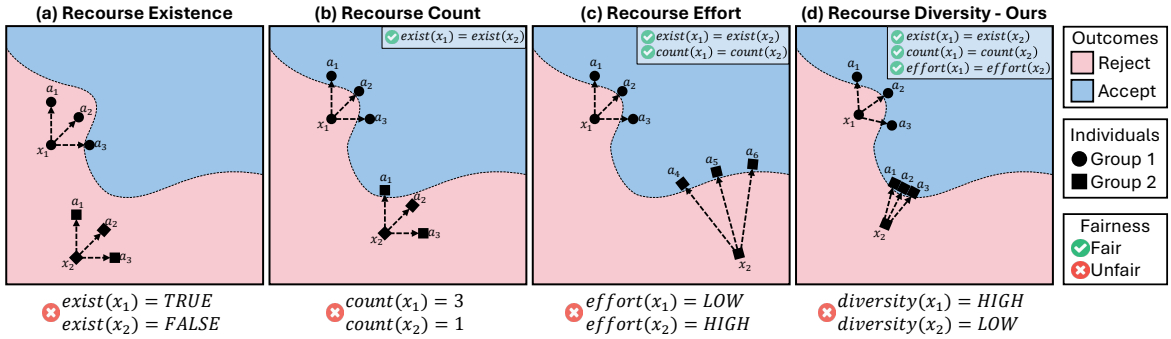

Fig. 1. Examples of unfair recourse explanations favoring rejected loan applicant x_1 over x_2 for several recourse fairness metrics. Moving left to right the explanations become progressively more fair according to the preceding metric(s).

Background. Of particular interest in recourse research is the development of a wide range of techniques for generating *counterfactual explanations*. These approaches aim to provide recourse by specifying changes to the features of an individual's (typically tabular) instance that would lead to a positive outcome—for example, suggesting that a loan applicant increase their income to a specific level to obtain approval [26, 58]. While the development and deployment of such explanation systems is significant, it remains unclear if these explanations lead to just outcomes. Recent work has shown that machine learning models often encode biases that lead to unequal distributions of predictions across groups [42]. Whether arising from chance or from discriminatory patterns in training data, these unwarranted associations can produce unjust outcomes and lead to discrimination on the basis of protected attributes like race or gender, in violation of anti-discrimination laws [16]. Thus, a variety of fairness metrics have been proposed to measure inequalities in the distribution of classification decisions and, in some cases, to modify the training data, model, or specific predictions to mitigate the effect of bias [42].

Limitations of the State-of-the-Art. Despite the prevalence of works that demonstrate bias in classification [10], few works have investigated the degree to which recourse-generation methods applied to these models may similarly exhibit bias. The key challenge in using counterfactual explanations for recourse is to ensure that the explanations are *actionable*, that is, that they suggest changes that are practical and reasonably achievable for the individual in question. Thus the literature on the fairness of counterfactuals for actionable recourse primarily focuses on *recourse existence* [3, 15, 53], i.e., whether an action exists that enables an individual to obtain recourse (Fig 1a); or on *recourse effort* [9, 15, 22, 28, 53, 57], i.e., the cost required for an individual to execute that action (Fig 1c). Fairness is then assessed based on disparities in these factors between individuals/groups.

While valuable, these works fall short of fully examining the real-world practicality of counterfactual explanations for recourse. Namely, as individuals face complex and personal constraints that affect the actionability of a counterfactual, they must be given multiple alternate counterfactual explanations, representing different potential paths to obtaining recourse [7, 40, 54]. Some existing works aim to address this by considering the *recourse choice count*, i.e., the number of such paths (Fig 1b), as a secondary fairness factor [9, 28] – defining fairness as an equal count between individuals/groups. However, to date, the problem of assessing *the impact of similarity between recourse paths on recourse fairness remains unexplored in the literature.*

Motivating Example. Consider two individuals, as in Fig 1d, each of whom has applied for a loan and been rejected. Both are presented with three possible actions, each constituting a valid and actionable path to recourse (e.g., achieved by increasing some combination of income and savings) and with the same cost (i.e., move an equal distance in the feature space). All three metrics discussed above would determine this to be *fair* in terms of recourse existence, effort, and count (See top right of Fig 1d). However, we notice that individual x_1 is presented with three actions that are *substantially different* from each other, while all three actions for individual x_2 are *nearly identical*. Therefore, we argue that x_1 is provided a much *higher degree of agency* in their choice than x_2.

For example, x_2 could receive explanations which all require increasing income, while x_1 is given the choice to increase income, increase savings, or decrease debt. As a result, x_1 is more likely to find a satisfying path to recourse and thus more likely to receive a loan. *Intuitively, this is not fair.* Yet existing recourse fairness notions overlook this, and instead indicate that this situation is fair.

Proposed Approach. To address this gap, we propose measuring the *diversity* of recourse actions available to different individuals to better assess the degree of agency they have in obtaining recourse. Specifically, we design a *diversity* metric that quantifies the similarity among a set of counterfactual actions. As shown above (Fig 1d), we develop diversity to be orthogonal to existing recourse fairness metrics. We argue that for truly fair recourse, individuals must be given explanations that are comparable not only in effort and count but also in diversity. While diversity has previously been considered as a desiderata in counterfactual explanation [30, 37, 40, 49, 58], as we discuss in Sec 2.1, these approaches do not develop diversity as a metric, as they are not concerned with fairness and are either ad-hoc or attempt to maximize the distance between points. Instead, we define diversity solely in terms of the similarity between counterfactual actions, ensuring it is independent of explanation distance.

In Sec 3, we provide technical preliminaries and establish notation. In Sec 4, we analyze diversity as a criterion for fairness, formulate our newly proposed diversity metric, and define our notion of individually fair recourse diversity. We extend this notion to the group level in Sec 5, covering both micro- and macro-level diversity. In Sec 6, we evaluate recourse diversity against existing fairness metrics across three model types and four datasets. We evaluate on the individual level (Sec 6.2) and ask questions such as: *Would individuals benefit from changing sensitive attributes like race and sex?*. We find this to be true, and demonstrate that real-world machine learning models can be individually fair in terms of recourse effort and count while exhibiting unfair disparities in recourse diversity. We also examine group-level fairness (Sec 6.3) for questions like: *Do different groups consistently receive more or less similar explanations than others?*; and *Does recourse diversity align with existing fairness metrics, and if so, how?*. Our results show that models can systematically (dis)favor particular groups with respect to recourse effort and diversity. This includes groups who receive favorable diversity and effort, groups who are doubly disfavored, and groups who are favored with respect to one metric but disfavored by the other. Further examining the effects of model selection on recourse fairness in Sec 6.4, we observe that choosing a model type not only selects for classification performance but also sets the overall range of recourse diversity available to individuals and concurrently makes implicit choices about which groups will be (dis)favored in recourse.

Finally, in Sec 7, we discuss the implications of these findings, illustrating that diversity shows clear value as a metric for assessing fairness in actionable recourse. We recommend that recourse diversity be used alongside other existing metrics to provide auditors with a more comprehensive picture of how decision-making systems treat the individuals they affect. We also encourage machine-learning practitioners to make use of recourse diversity and other recourse fairness metrics during the model development process so that they can understand the otherwise invisible effects of design choices on people's ability to fairly access recourse. To support those goals, we release our evaluation tools on GitHub and make suggestions for future work in Sec 8.

2 Related Work

2.1 Methods for Generating Diverse Counterfactual Explanations

While recourse diversity has *not* been considered as a criterion for fairness, some explanation techniques use notions of diversity when generating explanations. However, these works do not propose a metric for measuring the diversity of a set of counterfactuals.

Diversity as Avoiding Repeats. Wachter et al. [58] first consider counterfactual diversity and use gradient descent optimization with multiple initializations to create multiple "diverse" counterfactuals. Unfortunately, this approach cannot guarantee the generation of multiple distinct counterfactuals as it is random chance whether each initialization will converge to a different local minima. Similarly, Russell [49] generates counterfactuals via

integer programming optimization for decisions by linear classifiers. They provide "diversity" through a simple heuristic that prevents the same feature alterations as a counterfactual already in the set. This avoids only exact duplicate counterfactuals and does not preclude feature alterations that are substantially similar. Thus, neither work actually attempts to measure the diversity of the counterfactuals they create.

Diversity as Maximized Distance. Mothilal et al. [40] and Mohammadi et al. [37] develop optimization approaches that include some notion of similarity as a loss factor. The former develops DICE, which generates multiple explanations for differentiable models via integer programming. DICE's loss function aims to minimize the (mean absolute) distance of each counterfactual to the explained instance while maximizing the average pairwise distance between counterfactuals. As these goals are in tension, their loss function includes a parametrized trade-off between objectives. The latter approach [37] uses predicted model logits instead of gradients to generate multiple counterfactuals for neural networks. They adopt the pairwise distance approach from DICE but relax the constraint to enforce only a fixed minimum distance between pairs of counterfactuals.

Leofante and Potyka [30] examine counterfactual robustness (i.e., the tendency for an explanation to shift after perturbations to the explained instance) and suggest that returning many counterfactuals can improve robustness. As prohibitively many are required to guarantee robustness, they develop a greedy algorithm that takes a large set of positive samples and selects a subset of (50-1000) "diverse" positive samples. They add the subset to maximize either a) the distance of the new sample to the existing samples, or b) the difference in cosine angle between the new sample and the existing samples. To generate counterfactuals, they linearly transform each selected point towards the explained instance until moving closer would produce a negative prediction.

Notably, the above approaches are largely based explicitly on explanation distance. This makes their notions of diversity inappropriate for use as a fairness metric for two reasons. First, the average distance between pairs of points lacks a clear interpretation. The meaning of this value is highly dependent on the distance function used and the scale of the data. Thus, it is difficult to discern what qualifies as "diverse" as there is no well-defined value for ideal diversity. This also precludes comparisons across different datasets/configurations. Second, relying on explanation distance entangles the notions of recourse diversity and effort, making diversity unreliable. For example, the set of counterfactuals $\{income+\$100, income+\$1M, income+\$10M\}$ would produce a large "diversity" value yet provides no real choice about which features to alter. The cosine selection approach of Leofante and Potyka [30] is closest to providing a useful notion of diversity. However, their approach evaluates the similarity of a single counterfactual to an existing set rather than assessing the diversity of a set as a whole. In Sec 4, we propose a diversity metric based on our own analyses and a combination of ideas inspired by these works.

Interactive Counterfactual Explanation. An emerging area of research explores interactive counterfactual explanation [13, 17, 43, 52, 54, 59]. Rather than selecting a fixed set of explanations, these methods present the user with multiple options for obtaining recourse and allow the user to manipulate those options by applying constraints. The key benefit of such approaches is that they avoid the difficulty of determining an appropriate cost function (Sec 3) for measuring recourse effort. While these systems generally don't consider diversity directly, they support users in searching through a wider explanation space.

2.2 Fairness Auditing in Counterfactual Explanation

Below, we explore the different notions of fairness in counterfactual explanation introduced in the literature. This study of the literature points to an open gap where no works in fairness auditing have thus far considered the similarity of a set of counterfactual explanations as a factor in recourse fairness, as tackled by our work.

Recourse Effort. Ustun et al. [53] first consider the fairness of counterfactual explanations on single explanations for decisions made by linear classifiers. They measure fairness based on the *effort* required for individuals to obtain recourse, using either the maximum percentile shift of any feature or the total log percentile shift across features. Gupta et al. [22] build on this work to develop a method for using regularization during classifier training

to enforce fair recourse effort between groups at the cost of moderately reduced classification accuracy. Their regularization approach generates synthetic twins for each sample by flipping the value of a protected attribute and penalizing the model for the synthetic twin changing class or having a higher/lower (Euclidean) distance to the decision boundary. They enforce equal group-wise recourse effort directly for linear models or indirectly for arbitrary models using LIME [48] to locally approximate the underlying model.

von Kügelgen et al. [57] extend Gupta et al. [22] to incorporate a structural causal model (SCM), arguing that simply generating synthetic twins by flipping the sensitive attribute results in unrealistic instances, as in practice, many other features used for classification are causally related to sensitive attributes. Bynum et al. [9] extend von Kügelgen et al. [57] by incorporating *backtracking counterfactuals*, which use a SCM to consider what the hypothetical value of unobserved variables in the SCM would need to be to provide an individual with the desired outcome. This is less directly applicable to decision subjects obtaining recourse, but does not directly alter the value of observed variables and thus avoids breaking the relationships between variables in the SCM.

Recourse Robustness. Artelt et al. [3] identify counterfactual robustness as an individual fairness notion. They posit that similar individuals should be given counterfactual explanations that lead to similar recourse outcomes. They provide mathematical and empirical analyses of the effect of small perturbations to an instance on the resulting counterfactual explanation. Ehyaei et al. [15] connect the work of von Kügelgen et al. [57] and Artelt et al. [3] to show that flipping a sensitive attribute from one value to another (e.g., changing race from white to black) is a subset of the types of perturbations that counterfactual robustness considers, and that classifiers which are robust to these perturbations in sensitive features have implicitly fair recourse effort.

Recourse Opportunity and Choice. Kavouras et al. [28] consider additional criteria for recourse fairness outside of recourse effort. Namely, they identify that recourse opportunity (the portion of a group which can feasibly obtain recourse) and recourse choice (the number of feasible paths to recourse) are overlooked factors in recourse fairness. They also refine recourse fairness notions into micro (each individual acting independently) and macro (individuals collectively taking the same action) views and link this to approaches for global counterfactual explanation [32, 33, 47]. They further demonstrate that these are orthogonal concepts to recourse effort, and show that for the macro case, recourse opportunity and recourse effort are in a trade-off relationship.

3 Preliminaries

Decision System. Given an n-dimensional feature space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and a decision-making classifier $f : \mathcal{X} \to \mathcal{Y}$. Let $Y \in \mathcal{Y}$ be the set of possible classes containing at least a positive desired outcome y^+ and negative undesired outcome y^- (e.g., loan applicants in $\mathcal{X}$ classified as y^+ accept or y^- reject)[1]. Let $\mathcal{S} \subseteq \mathcal{X}$ be a set of one or more sensitive features (e.g., applicant race).

Recourse Actions. We consider a set of possible actions $A = \{a_1, a_2, \dots\}$ where each action a is a set of changes to the feature values in $\mathcal{X}$ (e.g., $a = \{\textit{income+\$100}, \textit{rent-\$50}\}$). For an instance $x \in \mathcal{X}$, we say action a provides recourse for x when the instance receives a negative outcome, $f(x) = y^-$, and applying the action to x as $x' = do(a, x)$ results in a positive outcome $f(x') = y^+$. While many $do()$ functions are possible, for simplicity, we consider $do()$ to be addition for numeric and ordinal features and assignment for categorical features as:

$$x' = do(a, x) = \begin{cases} x.\mathcal{X}_i + a.\mathcal{X}_i & \mathcal{X}_i \; numeric \\ pos^{-1}(pos(x.\mathcal{X}_i) + a.\mathcal{X}_i) & \mathcal{X}_i \; ordinal \\ a.\mathcal{X}_i & otherwise \end{cases} \tag{1}$$

where pos returns the position of the value in the given feature's ordering and pos^{-1} returns the value at the given position in the ordering. We can thus view the action a as a translation vector from x to x'.

[1]While we frame $\mathcal{Y}$ as binary help/harm for simplicity, this is not required. We assume only that there exists a pair of outcomes in $\mathcal{Y}$ where a user desires one more than another. We explicitly encourage using recourse for harm reduction when a fully positive outcome is impossible.

Action Feasibility. For an instance x, let $feas(a, x)$ be a function which returns whether the individual represented by x can practically enact the action a. For a set of actions A, let $A_x \subseteq A$ be the set of actions in A that are feasible for x and lead to recourse.

$$A_x = \{a \in A \mid feas(a, x) \wedge f(do(a, x)) = y^+\} \tag{2}$$

Action Cost. Existing works have measured the fairness of recourse based on the minimum effort required to obtain a positive outcome. For a specific action a, let $cost(a, x)$ be a function which measures the cost of performing the intervention $do(a, x)$. Many existing works evaluate cost as the distance between x and $x' = do(a, x)$ using some distance function δ. Popular choices for δ are L1 or L2 norms, mean absolute distance (MAD), and quintile shift [27]. We note that, as observed in [55], the determination of an action's cost and feasibility depends on the actionability of features for a specific individual, and the cost of performing alterations on those features. As the authors discuss, features may not always be jointly actionable. Cost functions may need to be personalized, making the determination of these functions themselves an interesting challenge. Nevertheless, we follow existing works in recourse fairness [22, 28, 53, 57] and assume that such feasibility and cost functions are known.

While our framework supports any cost function, for simplicity, we assume cost corresponds to the magnitude of changes in a. Following Kavouras et al. [28], we measure this as the Gower distance between x and $x' = do(a, x)$ as $cost(a, x) = \delta(x, x')$ given

$$\delta(x, x') = \sum_{i=1}^{n} \begin{cases} |norm(x.\mathcal{X}_i) - norm(x'\mathcal{X}_i)| & \mathcal{X}_i \; numeric \\ |pos(x.\mathcal{X}_i) - pos(x'\mathcal{X}_i)|/npos & \mathcal{X}_i \; ordinal \\ 1 * (x.\mathcal{X}_i \neq x'\mathcal{X}_i) & \mathcal{X}_i \; categorical \end{cases} \tag{3}$$

where *norm* normalizes each feature $[0, 1]$, *pos* returns the position of the value in the given feature's ordering, and *npos* is the number of positions in the ordering for $\mathcal{X}_i$. This is functionally equivalent to L1 distance for mixed feature type data.

Groups. Given a dataset of individuals $D \in \mathcal{X}$, we define groups of individuals $G_p \subseteq D$ based on one or more predicates p, (e.g., $p = income < \$1{,}000$ or $p = (gender == male)$) to be all individuals which satisfy p, i.e., $G_p = \{x \in D \mid p(x)\}$. We say G_p is a sensitive group if p conditions one or more of the protected features in S.

4 Individually Fair Recourse Diversity

For two individuals to be treated fairly, they must have the same ability to choose between multiple courses of action to obtain recourse. While existing works have considered fairness as the *number* of choices for feasible actions that lead to recourse [28] and the cost of these choices [22, 53, 57], these factors alone are insufficient, as they do not measure individuals' ability to choose between meaningfully distinct paths to recourse. For example, consider individual x_1 with feasible actions $feas(A, x_1) = \{\{income + \$100\}, \{income + \$101\}\}$ and Individual x_2 with feasible actions $feas(A, x_2) = \{\{income+\$100\}, \{savings+\$100\}\}$. While both individuals have the same number of choices (two), and those choices have nearly identical costs ($\sim$ \$100 each), individual x_2 obviously has much greater agency to choose their desired path to recourse as their feasible actions are substantially different from each other, while the feasible actions for x_1 are functionally identical.

To measure this notion of diversity for a single instance, we propose to compute the mean absolute pairwise dot product (relative to the explained instance) of all feasible actions for the individual. This represents the degree to which an individual is able to choose between meaningfully different actions to obtain recourse, with a diversity of 0 indicating that there is one possible path to obtaining recourse, and 1 indicating that there are many possible paths. Let the diversity of a set of actions A be defined as

$$diversity(A) = 1 - \frac{1}{\binom{k}{2}} \sum_{i=1}^{k} \sum_{j>i}^{k} |\hat{a}_i \bullet \hat{a}_j| \tag{4}$$

where $\hat{a} = \frac{a}{|a|}$ is the normalized unit vector a, $\hat{a}_i \bullet \hat{a}_j$ is the dot product of those vectors, and $k = |A|$ is the number of actions. The recourse diversity for an instance x is the diversity of the set of feasible actions for x.

$$diversity(A, x) = diversity(feas(A, x)) \tag{5}$$

Diversity is thus intuitively a measure of the spread of a set of actions and is independent of the magnitude of those actions (and therefore orthogonal to recourse effort). We note that the computation of ideal diversity is dependent on the number of counterfactuals in relation to the number of features in the data. Thus, there are three cases with distinct maximal diversity sets.

Case 1: Fully Orthogonal. Given $k \leq n$, then the ideal set of counterfactual actions is fully orthogonal and

$$\sum_{i=1}^{k} \sum_{j>i}^{k} |\hat{a}_i \bullet \hat{a}_j| = 0 \tag{6}$$

Note that when this is true, the set of actions is guaranteed to be optimal with respect to diversity (i.e., no set of k actions will have higher diversity). Further, all such sets will be simple rotational transformations of each other.

Case 2: Orthogonal and Antiparallel. Given $n < k \leq 2n$, then the ideal set of counterfactual actions contains n fully orthogonal vectors and $k - n$ vectors which are orthogonal to all but one of the other vectors in the set (i.e., of the k total vectors, each of the $k - n$ additional vectors is orthogonal to $k - 2$ vectors, parallel to one vector being itself, and antiparallel to exactly one vector)

$$\sum_{i=1}^{k} \sum_{j>i}^{k} |\hat{a}_i \bullet \hat{a}_j| = k \tag{7}$$

Note that while this holds for the ideal set of vectors to achieve maximal diversity, it also holds for many other non-ideal sets. Thus, this is necessary but not sufficient to check that a set of actions provides optimal diversity.

Case 3: Non-Orthogonal. Given $k > 2n$, the ideal set of counterfactual actions is a solution to the Tammes Problem [39], which seeks to maximize the minimum distance between a pair of points. This is equivalent to finding the largest possible radius such that k identical spherical caps can be placed on a sphere and not overlap. Formally, let $x'_i = do(a_i, x)$ be the result of applying action a_i to instance x. Then an ideal set of actions A^* is

$$A^* = max\left\{ \min_{1 \leq i < j < k} \left\{ |x'_i - x'_j|_2 \right\} \right\} : i = 1, \ldots, k \tag{8}$$

Note that there are many possible sets A^* that are valid solutions to the Tammes problem, any of which are sufficient to achieve optimal diversity, and as in Case 1, are simple rotational transformations of each other. Given human limitations on processing multiple data points at once and the comparatively high dimensionality of modern datasets, we limit our analysis to Case 1.

Individually Fair Recourse Diversity. Two individuals x_i and x_j have individually fair recourse diversity when the diversity of their feasible actions is the same, i.e., the two individuals are able to choose from a set of equivalently diverse actions:

$$diversity(feas(A, x_i)) - diversity(feas(A, x_j)) = 0 \tag{9}$$

Counterfactually Fair Recourse Diversity. While the notion of individually fair recourse diversity captures the degree to which two instances have the same degree of choice, the selection of which instances to compare is greatly significant, as substantially different instances are likely to obtain substantially different recourse diversity. Similar to Ustun et al. [53] and Gupta et al. [22], we argue that if a model is fair with respect to a sensitive feature $s \in \mathcal{S}$, then altering the value of feature s for an instance x should not impact the diversity of recourse offered to that instance. More formally, let x^{CF} (a counterfactual twin) be a modification of x such that the value of a sensitive attribute is altered (e.g., flipping a race white/black), but all other features retain their existing value. Then the model can be said to offer fair recourse diversity with respect to s if $diversity(feas(A, x)) - diversity(feas(A, x^{CF})) = 0$. If this does not hold, then the model has learned some bias in favor of one or more protected attributes. An "optimally" fair model would provide counterfactually fair recourse diversity for all features in $\mathcal{X}$ and all combinations of $s \in \mathcal{S}$, though this is highly unlikely in practice. We use this approach to measure the fairness of different models in Sec 6 with respect to selected attributes. Creating the counterfactual twin for such evaluations requires the sensitive attribute to a) be visible to the model, or b) be visible to the auditor, but not the model, and a SCM is available which can map changes in $\mathcal{S}$ to $\mathcal{X}$.

5 Group Fair Recourse Diversity

Given a group of instances $G = \{x_1, x_2, \dots\}$, we assess the diversity of the actions for that group according to both macro and micro perspectives following the macro/micro taxonomy introduced in [28].

Group Recourse Diversity (micro). Given a group G, the micro perspective views each individual in G acting independently to choose between their own set of feasible recourse actions. In this case, we compute group diversity as the average diversity of the actions for each instance. This represents the average degree of choice given to individuals in the group.

$$groupdiv_{micro}(G, A) = \frac{1}{|G|} \sum_{x \in G} diversity(feas(A, x)) \tag{10}$$

Group Recourse Diversity (macro). Given a group G, the macro perspective considers all individuals in G taking the same action. Consequently, we define the group's feasible actions as $A_G = feas(A, G)$, where $feas(A, G)$ returns the set of actions in A that provide recourse for all instances in G, i.e., $\forall a \in A_G : \{\forall x \in G : f(do(a, x)) = y^+\}$. Macro group recourse diversity is thus:

$$groupdiv_{macro}(G, A) = diversity(feas(A, G)) \tag{11}$$

This macro perspective on group diversity is valuable as it enables auditing possible structural counterfactual interventions, e.g., a policy maker examining a set of financial policies to assess which would best support black residents in obtaining mortgage loans.

However, if the set of common feasible actions for the group is small, then measuring the recourse fairness of that set has low informative value (and is impossible if the set is empty). We include macro recourse diversity for completeness and to relate to existing works [28, 47] which consider actions feasible for most, but not all, of a group (where macro diversity may be meaningful).

Group Fair Recourse Diversity. Two groups G_i and G_j have fair recourse when their diversity is equal.

$$groupdiv(G_i, A) - groupdiv(G_j, A) = 0 \tag{12}$$

Considering the macro and micro perspectives thus yields two distinct metrics for group fair recourse diversity, with the relative significance of each metric being determined by the end goal of the recourse task at hand.

	Dataset	Flip Atr	Logistic Regression			Neural Network			Random Forest		
			Δ acc ↓	Δ effort ↓	Δ div ↓	Δ acc ↓	Δ effort ↓	Δ div ↓	Δ acc ↓	Δ effort ↓	Δ div ↓
Heuristic	adult	race	0.03±.01	0.02±.00	0.15±.01	0.02±.00	0.02±.00	0.15±.01	0.06±.01	0.03±.00	0.16±.01
	adult	sex	0.08±.01	0.02±.00	0.15±.01	0.10±.02	0.02±.00	0.15±.01	0.03±.01	0.02±.00	0.15±.01
	adult	race, sex	0.08±.01	0.02±.00	0.14±.00	0.10±.02	0.02±.00	0.15±.01	0.06±.01	0.02±.00	0.14±.00
	compas	race	0.05±.00	0.02±.00	0.21±.01	0.02±.00	0.02±.00	0.19±.01	0.04±.01	0.03±.00	0.23±.01
	compas	sex	0.04±.01	0.02±.00	0.22±.01	0.02±.00	0.02±.00	0.20±.00	0.05±.01	0.03±.00	0.23±.01
	compas	race, sex	0.03±.01	0.02±.00	0.21±.01	0.02±.00	0.02±.00	0.19±.00	0.06±.01	0.03±.00	0.22±.01
	german	gender	0.00±.00	0.03±.00	0.12±.01	0.00±.00	0.06±.01	0.12±.02	0.03±.02	0.03±.00	0.12±.00
DICE	adult	race	0.03±.01	0.04±.00	0.19±.01	0.02±.00	0.03±.00	0.05±.00	0.06±.01	0.04±.00	0.20±.01
	adult	sex	0.08±.01	0.04±.00	0.22±.00	0.10±.02	0.03±.00	0.05±.00	0.03±.01	0.04±.00	0.19±.01
	adult	race, sex	0.08±.01	0.04±.00	0.20±.01	0.10±.02	0.03±.00	0.05±.00	0.06±.01	0.04±.00	0.21±.01
	compas	race	0.04±.00	0.03±.00	0.18±.01	0.02±.01	0.04±.00	0.09±.00	0.04±.01	0.04±.00	0.23±.01
	compas	sex	0.04±.01	0.04±.00	0.19±.01	0.02±.01	0.06±.00	0.10±.00	0.05±.01	0.04±.00	0.24±.01
	compas	race, sex	0.03±.01	0.04±.00	0.19±.01	0.02±.01	0.05±.00	0.07±.00	0.07±.01	0.04±.00	0.24±.02
	german	gender	0.00±.00	0.05±.00	0.15±.01	0.00±.00	0.03±.00	0.06±.01	0.03±.02	0.05±.00	0.16±.01

Table 1. Mean and standard error of the absolute change in balanced accuracy (acc), recourse effort (effort), and recourse diversity (div) caused by flipping an individual's race attribute (binary black/white), sex attribute (female/male), or both. For each instance, $k = 3$ explanations were generated with Heuristic (upper) and DICE [40] (lower) explainers. Results are averaged across five stratified random samples of individuals. A counterfactually fair model would result in no change in the model's accuracy, recourse effort, or recourse diversity, i.e., metric is zero.

6 Experimental Study

6.1 Experimental Setup and Methodology

To evaluate the fairness of recourse explanations, we need to 1) obtain appropriate data about individuals containing sensitive feature information, 2) train an automated decision-making system on this data, then 3) generate multiple counterfactual explanations for each individual. We detail our methods below and release our auditing tools on GitHub as an extensible open source framework: https://github.com/PeterVanNostrand/DiverseRecourse

Datasets. We select several datasets for evaluation based on their usage in similar fairness auditing works.

- ***UCI Adult*[5]**: Financials and demographics for 48842 individuals predicting if annual income exceeds $50k.
- ***COMPAS*[1]**: Criminal charge and demographics for 6167 individuals predicting binary recidivism in 2 years.
- ***German Credit*[25]**: Financials and demographics for 1000 loan applicants predicting reject or accept.
- ***GiveMeCredit* [20]**: Financials and demographics for 117831 individuals predicting debt delinquency in 2 years.

Models. We train logistic regression (LR), random forest (RF), and neural network (NN) classifiers on an 80/20 train-test split of each dataset. We use Scikit [46] with default parameters for LR and RF (100 trees, no max depth). We use Keras [12] to create a NN with 2, 10-node internal layers and train for 10 epochs with learning rate 0.001.

Explanation Methods. As there are many competing approaches to generate multiple counterfactuals for an instance, here we implement a straightforward heuristic adapted from [30]. Given an instance x with class y

(1) Sample a set of $v = 100$ uniform random action vectors.
(2) Walk along each vector away from x. At each step i, compute the stepped point $x^i = do(x, a^i)$ and check if it is counterfactual to x (i.e., $f(do(x, a^i)) \neq y$). Stop when x^i is counterfactual, or x^i exceeds the feature range.
(3) Drop action vectors for which x^i is not counterfactual.
(4) For each action vector a, starting from the counterfactual point x^i, binary walk along a towards x to find the point x' which is the closest point to a that remains counterfactual.
(5) Sort the final points x' by their distance to x, and select the k nearest points.

We refer to the above method as the "Heuristic" explainer. In Sec 6.2, we also compare the fairness of explanations generated by this technique to those generated by the popular DICE [40] method discussed in Sec 2.

Number of Explanations. For Sec 6.2-6.4, we generate $k = 3$ counterfactual explanations for each instance, as this is a sufficient number of explanations to compute recourse diversity while being a small enough amount of information that the user is not likely to be overwhelmed. Indeed, using a small set (3-5) of counterfactuals has been shown to improve user understanding compared to single counterfactuals [7, 51]. We explore the fairness effects of varying the number of counterfactuals in Appdx A.4.

Extracting Groups. In Sec 6.3, we analyze group recourse fairness. Per our definition in Sec 3, a group G_p is defined as a subset of the dataset D described by some predicate(s) p (e.g., $p = income < \$1{,}000$). To learn these predicates, we follow existing work [47] and perform associative rule mining on each dataset to extract populations of individuals with common factors. As in [47], we use a minimum support threshold of 0.01 for all datasets except German, which uses 0.1, as its smaller number of instances led to an excessive number of groups. We list the number of groups and the mean number of individuals per group in Appdx A.2 Tab 3.

6.2 Auditing Individual Recourse Fairness

Methodology. To evaluate individual recourse fairness, we use the notion of counterfactual recourse fairness discussed in Sec 4, which states that a model is unfair if altering the value of an individual's sensitive attribute(s) provides them with more/less desirable explanations. As explaining the entire dataset multiple times is cost-prohibitive, we follow the approach from [57] and take a stratified random sample of 20 individuals per intersection of sensitive attributes from the testing set (e.g., 20 black men, 20 black women, etc). We then generate $k = 3$ explanations for each individual using the Heuristic (Sec 6) and DICE [40] explainers and compute their recourse diversity and mean recourse effort. We also compute the mean balanced accuracy of each intersection as a

measure of prediction fairness. Then, for each individual, we create all possible counterfactual twins by altering their protected attributes. E.g., for a black man, we create three counterfactual twins by first altering race, then sex, then both while leaving all other features unchanged. We then explain each of the individual's counterfactual twins and compare these explanations to those of its original instance to determine if the model is favoring certain sensitive attribute values. The results of this process are shown in Table 1 using $S = race, sex$ as sensitive features (note the GiveMe credit dataset is excluded as it does not provide these features). For consistency, we repeat the process on five random samples of individuals and compute the mean and standard error change.

Individual Fairness. For a completely fair model, all counterfactual twins of each instance would obtain explanations which provide equivalent diversity and require equivalent effort, and the classifier would be equally accurate on each set of counterfactual twins. Examining Table 1, we see that this is not always true, with many cases exhibiting a non-zero difference. However, for balanced accuracy and recourse effort, we see the scale of the difference between each instance and its counterfactual twin is comparatively low for most cases. This indicates the classifiers are not substantially more accurate for one sensitive group than another and that instances which are comparable across each sensitive group are given recourse explanations that require comparable amounts of effort to enact. In contrast, we observe that the change in recourse diversity is often much larger. For the Logistic Regression and Random Forest models, both the heuristic and DICE explanation techniques produce a substantial difference in diversity between an instance and its counterfactual twins. Thus, these models are favoring certain values of race/sex in a way that existing fairness notions do not capture.

Effects of Explanation Technique. Comparing the explanation techniques, we find that the scale of change in recourse effort and recourse diversity is mostly similar. However, for the Neural Network models, the Heuristic explainer produces a large change in diversity between instances and their twins, while DICE produces explanations that are much more consistent in diversity. We discuss likely causes and implications of this in Sec 7.

6.3 Auditing Group Recourse Fairness

In this section, we evaluate our notion of micro group recourse diversity fairness as proposed in Sec 5. Given the training process and groups of instances described in Sec 6.1, we first classify all instances in the dataset with a logistic regression model and explain each decision with the Heuristic explainer ($k = 3$). Then, we compute the mean micro recourse diversity and effort of each group. To determine what types of groups are (dis)favored, we divide the set of all groups into subsets based on their shared predicate values for each feature. Then, we plot the frequency of recourse effort and recourse diversity for these subsets in Fig 2-5. For example, the groups G_1:($Sex = Female \wedge 17 < Age \leq 26$) and G_2:($Race = White \wedge 17 < Age \leq 26$) share the predicate $17 < Age \leq 26$ and would thus both be members of the blue distribution in the first subplot of Fig 2. Note that this is distinct from the distribution of individuals as one individual can fall in multiple groups (e.g., a white woman age 25 in $G_1 \& G_2$).

Considering these results, we first observe that the distribution of recourse diversity and effort follow distributions which are roughly normal or skewed-normal. This shows that both metrics are able to discriminate effectively between groups. Secondly, we find that recourse diversity and effort do not merely reproduce the same distribution of scores and therefore measure distinct aspects of fairness. By examining both metrics, we can identify cases where the metrics are in alignment (i.e., (dis)favoring the same groups) and cases where groups are favored with respect to one metric and disfavored by the other. For example, in Fig 4, the groups predicated $45 < Age \leq 75$ (purple) receive both the lowest recourse diversity and the highest recourse effort. This indicates that such individuals are doubly disfavored, as high diversity and low effort are desirable, and will have a harder time obtaining recourse and less freedom to choose their path than those age 19-26 (blue) or 26-30 (orange).

In contrast, for several cases, recourse diversity detects differences in fairness that recourse effort does not. For example, in credit (Fig 2), the 17-26 age bracket receives notably lower diversity than the other age brackets, a differential treatment not discernible from recourse effort. This also holds for Marital Status, Relationship Status, and Hours Per Week; AgeCat in the Compas dataset (Fig 3); Loan Duration and Years At Current Home

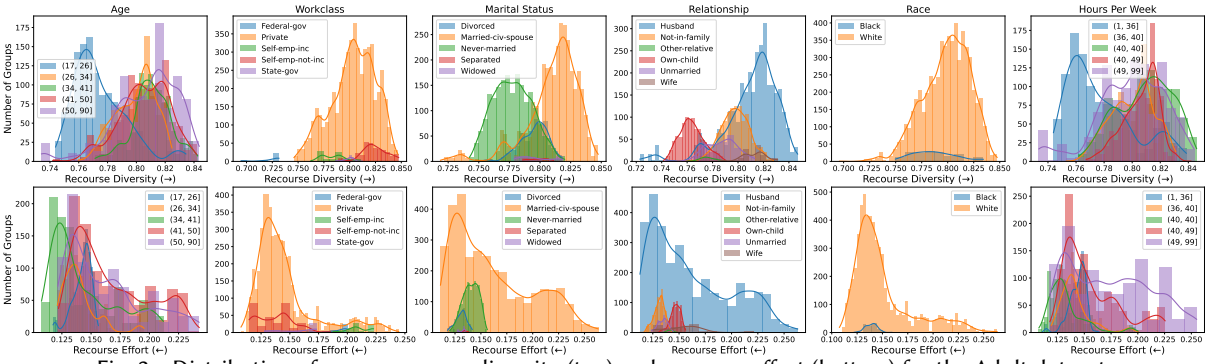

Fig. 2. Distribution of group recourse diversity (top) and recourse effort (bottom) for the Adult dataset.

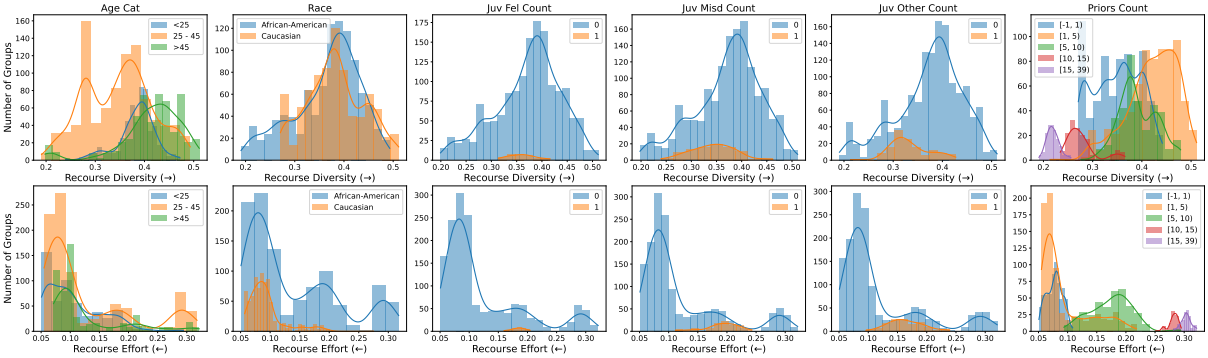

Fig. 3. Distribution of group recourse diversity (top) and recourse effort (bottom) for the Compas dataset.

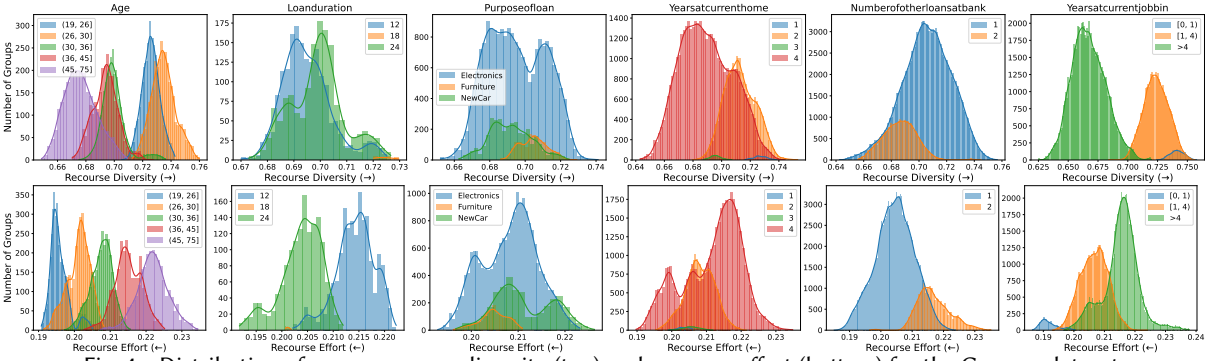

Fig. 4. Distribution of group recourse diversity (top) and recourse effort (bottom) for the German dataset.

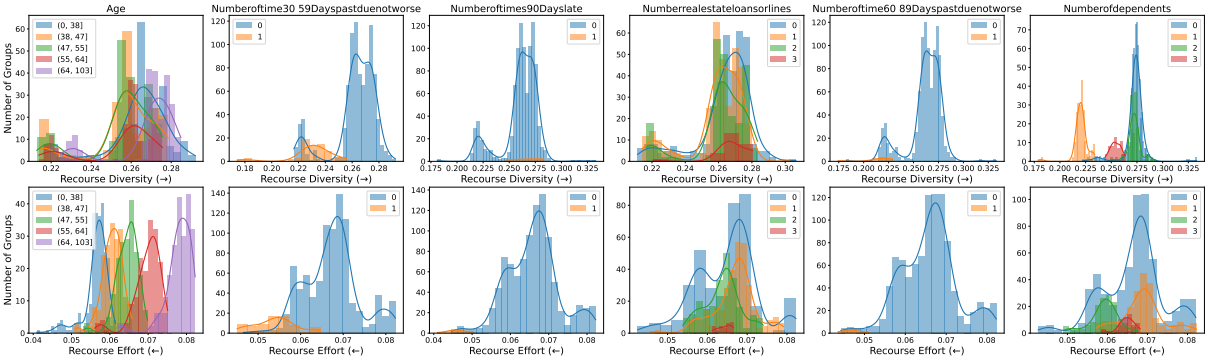

Fig. 5. Distribution of group recourse diversity (top) and recourse effort (bottom) for the GiveMe dataset.

in the German dataset (Fig 4); Number of Dependents in the GiveMe credit dataset (Fig 3); and more. Thus, using recourse diversity alongside effort empowers practitioners to both detect otherwise overlooked fairness implications and analyze the alignment between metrics to better understand their models.

6.4 Examining Effects of Model Selection

Next, we examine the effect of model architecture on recourse fairness. We train a LR, NN, and RF and explain each as in Sec 6.3. Fig 6-9 show the distribution of micro group recourse diversity (recourse effort is in Appdx A.3).

Examining these results, we find that the distribution of diversity is coarsely similar across the three models. This implies that the (un)fairness we observe is driven in part by biases in the underlying dataset and not simply a chance artifact produced by a specific model. Despite this general similarity, a closer inspection reveals cases where models differ significantly in diversity for certain groups. For example, on the Adult dataset in Fig 6, the groups with $(17 < Age \leq 26)$, (*MartialStatus = Never-married*), (*Relationship=Own-child*), and $(1 < HoursPerWeek \leq 36)$ show substantially lower diversity for LR than the NN and RF cases. Indeed, the $(1 < HoursPerWeek \leq 36)$ bracket has the lowest diversity of all brackets for LR, a median diversity for NN, and the highest diversity for RF. This indicates that the treatment of individuals in those groups is highly dependent on model selection.

Similar effects appear in the Compas dataset in Fig 7 for (*JuvFelCount=1*), (*JuvMisdCount=1*), and (*JuvOtherCount=1*), where LR places these groups centered with respect to their (*Count=0*) alternatives, while NN places the (*Count=1*) groups higher and RF lower. The $(AgeCat \leq 25)$ and (*Race=African-American*) groups also skew left for RF, indicating that this model has notable age and race bias against these groups compared to the LR and NN. While these effects show differential treatment of *certain* groups, a model can also have higher/lower diversity across *all* groups. We observe this in Fig 8, where LR and RF have diversity ranging ∼0.65-0.75 while NN ranges ∼0.25-0.6. Thus, even the group with the highest diversity in the NN will have less diverse explanations than the lowest diversity group in the LR and RF. This effect is compounded by differential treatment of groups with NN disfavoring some groups that LR and RF favor, such as (*YearsAtCurrentHome=2*) and $(1 < YearsAtCurrrentJob \leq 4)$.

7 Discussion

In this section, we discuss our experimental findings from Sec 6, with a focus on deriving recommendations for performing fairness audits with recourse diversity. We also explore fairness implications of model- and explainer-related design decisions, and provide suggestions for machine learning practitioners for system development.

Individual Fairness Implications. From the results of our individual fairness auditing in Sec 6.2, we find that the models we evaluate are nearly counterfactually fair in recourse effort for race and sex, but generally show distinct unfairness with respect to recourse diversity. Qualitatively, this indicates that for many individuals, completing the statement *"If I were a different race/sex..."* with *"it would be **easier** for me to get my desired outcome"* is false, but *"I would have **more freedom to choose** how I get my desired outcome"* is true. Further, examining the effects of flipping race, sex, or both, we find the effect sizes to be comparable across race and sex, and that flipping both does not yield a substantially larger impact than either alone. This indicates that the models may treat each intersection of feature-values discretely rather than favoring certain values in all cases. This suggests that fairness auditors should identify the sensitive features for their application, then evaluate recourse fairness across all intersections of those features, rather than only considering top-level aggregates for each.

Choice of Explanation Technique. In Sec 6.2, we examined the effect of different explanation techniques on individual recourse fairness. We found that in most cases, DICE and the Heuristic explainers had similar recourse fairness. However, for NN, only DICE was counterfactually fair in recourse diversity, with DICE's explanations offering the same degree of choice to individuals differing solely in race/sex, while the Heuristic explainer favored certain race/sex values. This may be because the Heuristic explainer seeks only to minimize recourse effort, while DICE seeks explanations with low effort *and* high diversity. Further, while DICE provided fair diversity for NN, it produced unfair diversity for LR and RF, possibly a result of DICE using different implementations based on

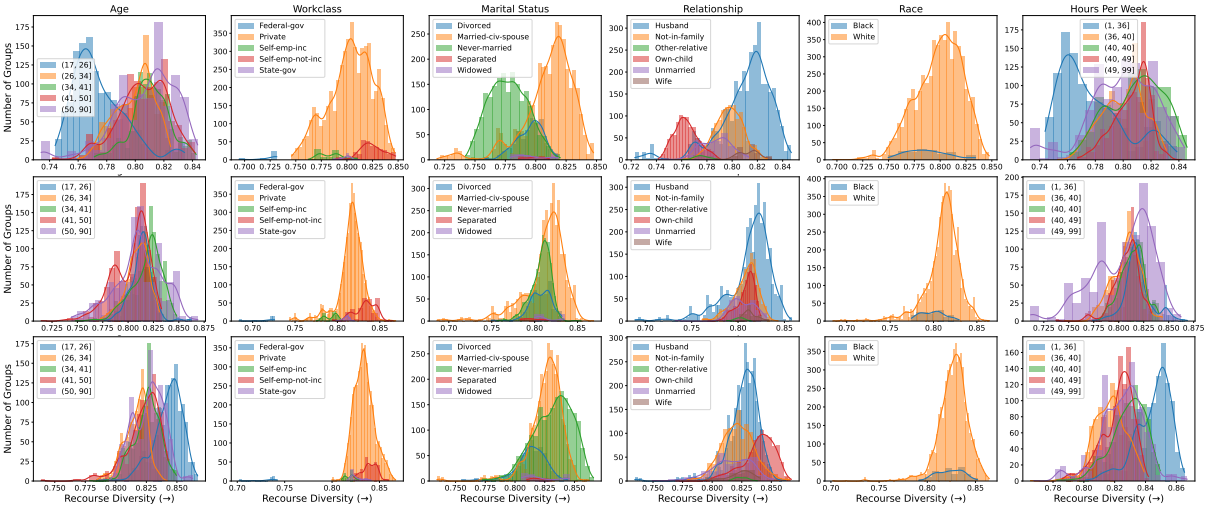

Fig. 6. Distribution of group recourse diversity for the Adult dataset trained on LR (top), NN (middle), and RF (bottom).

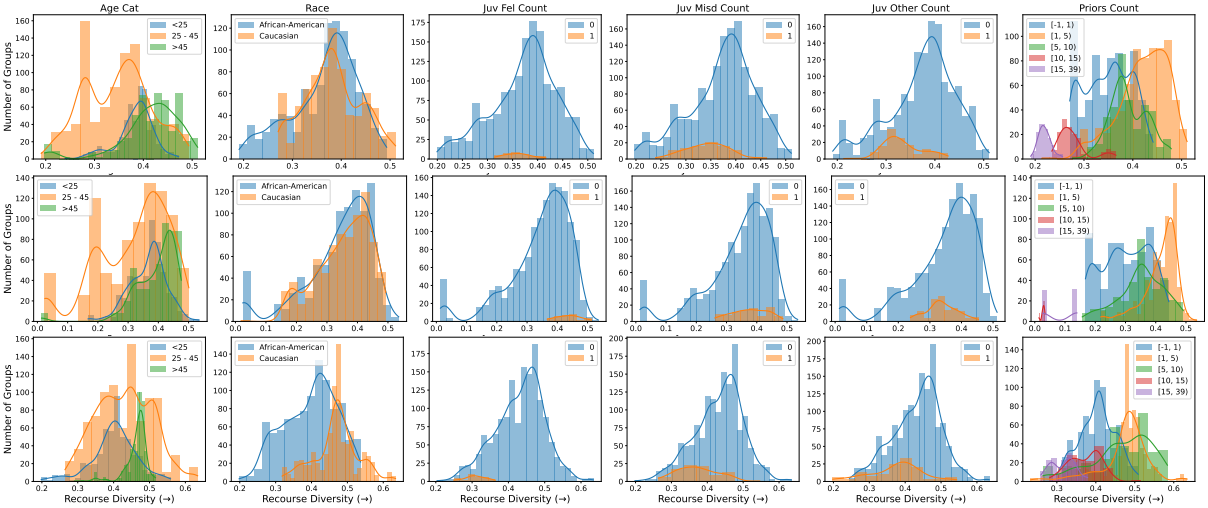

Fig. 7. Distribution of group recourse diversity for the Compas dataset trained on LR (top), NN (middle), and RF (bottom).

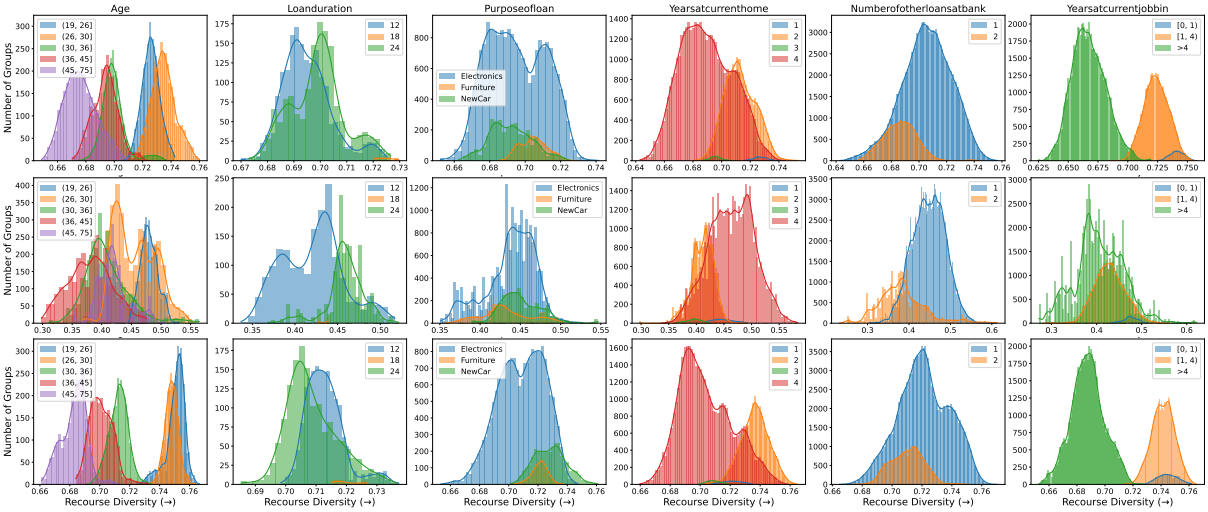

Fig. 8. Distribution of group recourse diversity for the German dataset trained on LR (top), NN (middle), and RF (bottom).

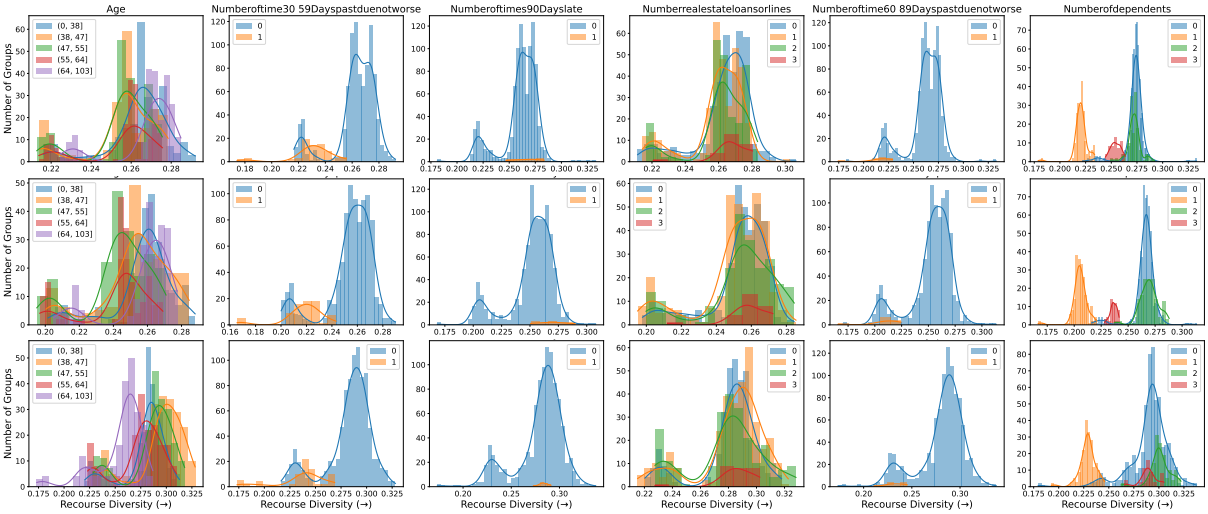

Fig. 9. Distribution of group recourse diversity for the GiveMe dataset trained on LR (top), NN (middle), and RF (bottom).

model type (namely, a gradient-based approach for NN, and a genetic algorithm for RF and LR). Combined, these observations suggest that machine learning practitioners should consider recourse fairness metrics, including diversity, when selecting an explanation technique, as this choice can strongly affect user treatment. Given that techniques can optimize for different goals and perform differently across model types, a system similar to model cards [36] may be useful for XAI methods to declare their desiderata and behavior in different scenarios.

Group Fairness Implications. At the group level (Sec 6.3), we found that across all datasets, there are features for which groups receive higher/lower recourse effort and micro recourse diversity depending on their feature-value. Notably, for race in the Adult and Compas datasets, we do not observe an overall disparity between racial groups. Comparing this to our individual fairness evaluation, this suggests that there is a correlation between race and other features in the dataset. This means that there is no single racial group that is definitively given more diverse explanations, but rather the favored value of race depends on each individual's specific features. However, we do find that there are features beyond race where certain groups have categorically more diverse explanations (e.g., Relationship Status in Adult, Age and YearsAtCurrentJob in German). We also observe that using recourse diversity and recourse effort in parallel reveals additional insights into the quality of explanations that each group receives. For example, auditors can identify groups that are doubly disfavored (low diversity and in high effort) or doubly favored (high diversity and low effort), as well as groups where explanations trade off diversity and effort.

When Does Disparity Become Discrimination? Considering these findings raises the question of how much disparity between groups constitutes discrimination. While this is undoubtedly subjective, we can draw inspiration from existing notions in equality law. For example, EU equal protection law defines indirect discrimination when an *"apparently neutral ... practice would put persons of one [protected group] at a particular disadvantage compared with [other persons], unless that ... practice is objectively justified by a legitimate aim"* [41]. Legal scholars argue that such a disadvantage can be proven when the *"likelihood of a member of a protected group being positively evaluated is at most 75% of the likelihood of a member of the privileged group"* [23]. Similarly, the US EEOC considers hiring rates to be substantially different when *"a selection rate ... is less than ... 80% of the selection rate for the group with the highest selection rate"*. While these notions are enforceable only in specific contexts, we can develop an analogous rule of thumb for auditors to assess differences in recourse fairness. For example, we could say that a group G_i is disadvantaged in recourse diversity if $groupdiv(G_i)/groupdiv(G_m) < r$, where G_m is the group with the highest diversity and r is a ratio in say $[0.75, 0.8]$. Then, assuming that no legitimate justification can be found, we could consider this disparity to be discrimination. Combining this with similar analyses for recourse

effort and count would enable auditors to detect potentially discriminatory disadvantages in recourse across multiple orthogonal metrics, with a clear disadvantage in any indicating a need for further investigation.

Fair Model Development. In Sec 6.4, we explore the effects of different models on recourse diversity. At the high level, we find that each dataset encodes relationships that lead models towards certain overall fairness implications. This is sensible as certain features are likely better predictors for the given task and thus more salient to the model. However, we also observe that each model learns distinct patterns, leading to different groups being favored. This includes cases where a feature-value may yield the highest diversity for one model but the lowest for another. Additionally, we found that different models can produce widely different diversity ranges across all groups. As such, choosing a decision-making model not only selects the quality of predictions but also implicitly decides the overall explanation quality and determines which groups will be (dis)favored. Based on this, we argue that in addition to serving as a post-hoc fairness-auditing metric, machine learning practitioners consider recourse fairness during model development. Notably, recent work has shown that there are many models with equally high performance for most problems [6, 21]. However, these models can differ substantially in their treatment of specific groups and, as suggested by our findings, may exhibit large differences in recourse fairness. Thus, using recourse fairness metrics during development will make explicit the otherwise invisible fairness impacts and enable machine learning practitioners to make informed model design decisions.

8 Limitations and Future Work

While the number of counterfactuals provided, the effort they require, and their diversity are necessary factors for providing fair recourse, they are not sufficient in themselves. In particular, while having a range of distinct paths to recourse is desirable, the direction of those paths can also be significant. E.g., for loan approval, an option increasing income is more desirable than one decreasing income. Thus, there are qualitative factors in recourse that are task- and individual-specific that these metrics do not capture. We therefore encourage fairness researchers not only to expand on these metrics, but also to make use of interactive explanation techniques (Sec 2.1) when appropriate, as such methods may enable users to express preferences for qualitative factors directly and thus remove the need for one-size-fits-all metrics. Further, we recommend that interactive methods consider using diversity as part of their workflows, as presenting users with a diverse set of initial explanations may offer useful starting points and help them understand the scope of possible paths to recourse. Finally, in some cases, multiple fairness metrics may be in tension. E.g., to decide which handful of options to present to an individual, we may be forced to trade off selecting low effort explanations to make the set of options diverse. We encourage machine learning practitioners to develop techniques for pre-processing and in-processing fairness interventions to help mitigate these trade-offs; e.g., by manipulating the training data to correct for bias or adding constraints during training. Future work should also seek to measure recourse fairness, including diversity, for additional tasks and perform more in-depth studies of the effects of design decisions like model parametrization.

9 Conclusion

In this work, we propose *diversity* as a critical metric for measuring the fairness of counterfactual explanations for actionable recourse. We analyze diversity as a fairness criterion, and provide definitions for individual and group fairness in recourse diversity. Comparing diversity to existing notions of fairness in recourse, we argue that diversity measures an important and overlooked aspect of fairness orthogonal to notions of recourse effort and choice count. Through comprehensive experimentation, we demonstrate that real machine-learning models can learn to be fair in recourse effort and choice count, but unfair in recourse diversity for both individuals and groups. We further explore the value of recourse diversity by examining the impact of model selection on recourse fairness and find large disparities in diversity between model types and differing biases towards certain groups. This underscores the need to consider recourse fairness during model design and selection. Finally, we provide our tools as open-source code and encourage practitioners to conduct fairness audits in additional settings.

Ethical Considerations. The auditing of recourse fairness should have largely positive social impacts by assisting machine-learning developers, internal auditors, regulators, and activists to identify concerning patterns in recourse explanations and, when possible, push for positive interventions. However, as with any tool, misuse is possible. Below, we identify potential hazards associated with diverse recourse explanations and recourse explanations in general.

- ***Misrepresentation of Fair Recourse Diversity.*** Identifying that counterfactual explanations for recourse have fair diversity does *not* imply that the explained machine-learning model makes fair or ethical decisions, nor does it imply that the explanations fairly distribute the burden of changes across groups. Additionally, the performance of a fairness audit alone should not be used as a band-aid or checkbox to substantiate claims of responsible AI use unless the audit results are properly considered and fairness violations are addressed.
- ***Bait-and-Switching Explanations.*** As we discuss in Sec 7, there are likely many possible models that are equally capable of performing a given classification task. Indeed, we encourage practitioners to consider recourse diversity when selecting a model. However, that does not mean that a fixed set of diverse explanations is equally valid for providing recourse for all these models. Recent work has explored how selecting another model or retraining the same model can substantially change the explanations of that model [8]. Once a user is given a counterfactual explanation, we agree with the view that these explanations are functionally promises that *if you perform the suggested alterations, you will definitely receive your desired outcome* [58]. Thus, changing the model after providing users with explanations can break that promise. Fair policies should either provide the user with a time frame for which the counterfactual is guaranteed, honor the previous counterfactuals regardless of the new model's decision, or provide counterfactuals that are robust to model updates [45].
- ***Failed or Noisy Execution.*** Another concern related to counterfactual explanations is the (in)ability of users to precisely control the feature-values of their instance. Once given a counterfactual, the user may attempt to enact this counterfactual by altering their features to obtain recourse. However, they may only be able to approximately control their features and thus deviate from the exact changes suggested by the counterfactual explanation. This is sometimes referred to as noisy execution [31]. Depending on the degree of perturbation from the counterfactual, the user may or may not obtain recourse. It is possible that providing the user with a diverse set of counterfactuals could exacerbate this issue, as failing to enact one of these counterfactuals would result in changes orthogonal to those suggested by the rest of the set. That said, several works explore methods for generating counterfactuals robust to such noisy execution [14, 34, 44, 54, 56]. Thus, applying such a technique ensures that if a user fails to enact one counterfactual from a diverse set, they would still have a reasonable chance of obtaining recourse. As a best practice, we suggest that machine learning practitioners consider the robustness of their counterfactuals to noisy execution and suggest that users of such systems obtain an updated set of counterfactuals if they are unable to follow their selected path to recourse.

Generative AI Usage. No Generative AI tools were used in the creation of the code, figures, or text of this work.

Acknowledgments. This research was supported in part by the US National Science Foundation under grants CSSI-2103832, IIS-1910880, IIS-2007932, and NRT-HDR-2021871. Thanks also to members of the DAISY research group for their valuable feedback on this project.

References

[1] AI Fairness 360. 2024. Compas Dataset. https://aif360.readthedocs.io/en/latest/modules/generated/aif360.sklearn.datasets.fetch_compas.html

[2] Julia Angwin, Surya Mattu, Lauren Kirchner, and Propublica. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. 2021. Evaluating Robustness of Counterfactual Explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, New York City, United States, 01–09.

[4] Article 29 Data Protection Working Party. 2016. Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679.

[5] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[6] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 850–863. doi:10.1145/3531146.3533149

[7] Clara Bove, Marie-Jeanne Lesot, Charles Albert Tijus, and Marcin Detyniecki. 2023. Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: An Explanation User Interface Proposition and User Study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 188–203. doi:10.1145/3581641.3584082

[8] Marc-Etienne Brunet, Ashton Anderson, and Richard Zemel. 2022. Implications of Model Indeterminacy for Explanations of Automated Decisions. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 7810–7823.

[9] Lucius E. J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. 2024. A New Paradigm for Counterfactual Reasoning in Fairness and Recourse. arXiv:2401.13935 [cs, stat] doi:10.48550/arXiv.2401.13935

[10] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7 (April 2024), 166:1–166:38. doi:10.1145/3616865

[11] CFPB. 2022. CFPB Acts to Protect the Public from Black-Box Credit Models Using Complex Algorithms. https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/.

[12] François Chollet et al. 2015. Keras. https://keras.io.

[13] Giovanni De Toni, Paolo Viappiani, Stefano Teso, Bruno Lepri, and Andrea Passerini. 2024. Personalized Algorithmic Recourse with Preference Elicitation. arXiv:2205.13743 [cs] doi:10.48550/arXiv.2205.13743

[14] Ricardo Dominguez-Olmedo, Amir H. Karimi, and Bernhard Schölkopf. 2022. On the Adversarial Robustness of Causal Algorithmic Recourse. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, Baltimore, Maryland, USA, 5324–5342.

[15] Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schoelkopf, and Setareh Maghsudi. 2023. Robustness Implies Fairness in Causal Algorithmic Recourse. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 984–1001. doi:10.1145/3593013.3594057

[16] Equal Credit Opportunities Act. 1974. Public Law, 15 C.F.R § 1691, Regulation B 12 C.F.R. § 1002.

[17] Seyedehdelaram Esfahani, Giovanni De Toni, Bruno Lepri, Andrea Passerini, Katya Tentori, and Massimo Zancanaro. 2024. Preference Elicitation in Interactive and User-centered Algorithmic Recourse: An Initial Exploration. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 249–254. doi:10.1145/3627043.3659556

[18] Wendy Fry, Ryan Tate, Vicki Haddock, Sisi Wei, and Adriana Heldiz. 2024. Landlords Are Using AI To Raise Rents - and Cities Are Starting To Push Back. https://themarkup.org/locked-out/2024/12/02/landlords-are-using-ai-to-raise-rents-and-cities-are-starting-to-push-back

[19] Joseph B Fuller, Manjari Raman, Eva Sage-Gavin, and Kristen Hines. 2021. *Hidden Workers: Untapped Talent*. Technical Report. Harvard Business School Project.

[20] Credit Fusion and Will Cukierski. 2011. Give Me Some Credit. https://kaggle.com/competitions/GiveMeSomeCredit. Kaggle.

[21] Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. 2025. Systemizing Multiplicity: The Curious Case of Arbitrariness in Machine Learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 2 (Oct. 2025), 1032–1048. doi:10.1609/aies.v8i2.36609

[22] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing Recourse across Groups. arXiv:1909.03166 [cs, stat] doi:10.48550/arXiv.1909.03166

[23] Philipp Hacker. 2018. Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law. *Common Market Law Review* 55, 4 (Aug. 2018), 1143–1186.

[24] Douglas Heaven. 2020. Predictive Policing Algorithms Are Racist. They Need to Be Dismantled. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/.
[25] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.
[26] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *Comput. Surveys* 55, 5 (Dec. 2022), 95:1–95:29. doi:10.1145/3527848
[27] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.* 55, 5 (Dec. 2022), 95:1–95:29. doi:10.1145/3527848
[28] Loukas Kavouras, Konstantinos Tsopelas, Giorgos Giannopoulos, Dimitris Sacharidis, Eleni Psaroudaki, Nikolaos Theologitis, Dimitrios Rontogiannis, Dimitris Fotakis, and Ioannis Emiris. 2023. Fairness Aware Counterfactuals for Subgroups. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 58246–58276.
[29] Colin Lecher. 2019. How Amazon Automatically Tracks and Fires Warehouse Workers for 'Productivity'. https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations.
[30] Francesco Leofante and Nico Potyka. 2024. Promoting Counterfactual Robustness through Diversity. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 19 (March 2024), 21322–21330. doi:10.1609/aaai.v38i19.30127
[31] Francesco Leofante and Matthew Wicker. 2025. Robustness of Counterfactual Explanations. In *Robust Explainable AI*, Francesco Leofante and Matthew Wicker (Eds.). Springer Nature Switzerland, Cham, 17–40. doi:10.1007/978-3-031-89022-2_3
[32] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. 2022. Global Counterfactual Explanations: Investigations, Implementations and Improvements. arXiv:2204.06917 [cs, stat] doi:10.48550/arXiv.2204.06917
[33] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. 2023. GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Honolulu, Hawaii, USA, 19315–19342.
[34] Donato Maragno, Jannis Kurtz, Tabea E. Röber, Rob Goedhart, Ş. İlker Birbil, and Dick den Hertog. 2024. Finding Regions of Counterfactual Explanations via Robust Optimization. *INFORMS J. on Computing* 36, 5 (Sept. 2024), 1316–1334. doi:10.1287/ijoc.2023.0153
[35] Emmanuel Martinez and Lauren Kirchner. 2021. The Secret Bias Hidden in Mortgage-Approval Algorithms. https://apnews.com/article/lifestyle-technology-business-race-and-ethnicity-mortgages-2d3d40d5751f933a88c1e17063657586.
[36] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. doi:10.1145/3287560.3287596
[37] Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. 2021. Scaling Guarantees for Nearest Counterfactual Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 177–187. doi:10.1145/3461702.3462514
[38] Beth Mole. 2023. UnitedHealth Uses AI Model with 90% Error Rate to Deny Care, Lawsuit Alleges. https://arstechnica.com/health/2023/11/ai-with-90-error-rate-forces-elderly-out-of-rehab-nursing-homes-suit-claims/.
[39] Edith Mooers. 1994. Tammes's problem.
[40] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. doi:10.1145/3351095.3372850
[41] Council of the European Union. 2000. Council Directive 2000/43/EC of 29 June 2000 Implementing the Principle of Equal Treatment between Persons Irrespective of Racial or Ethnic Origin. http://data.europa.eu/eli/dir/2000/43/oj
[42] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. 2023. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing* 7, 1 (March 2023), 15. doi:10.3390/bdcc7010015
[43] Indu Panigrahi, Sunnie S. Y. Kim, Amna Liaqat, Rohan Jinturkar, Olga Russakovsky, Ruth Fong, and Parastoo Abtahi. 2025. Interactivity x Explainability: Toward Understanding How Interactivity Can Improve Computer Vision Explanations. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3706599.3719730
[44] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, Valencia, Spain, 4574–4594.
[45] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, Virtual, 809–818.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[47] Kaivalya Rawal and Himabindu Lakkaraju. 2020. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Vancouver, Canada, 12187–12198.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1135–1144. doi:10.1145/2939672.2939778

[49] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 20–28. doi:10.1145/3287560.3287569

[50] Gesina Schwalbe and Bettina Finzel. 2023. A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts. *Data Mining and Knowledge Discovery* 38, 5 (Jan. 2023), 3043–3101. doi:10.1007/s10618-022-00867-8

[51] Nina Spreitzer, Hinda Haned, and Ilse van der Linden. 2022. Evaluating the Practicality of Counterfactual Explanations. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*. Curran Associates, Inc., New Orleans, Louisiana, USA, 20 pages.

[52] Muhammad Suffian, Pierluigi Graziani, Jose M. Alonso, and Alessandro Bogliolo. 2022. FCE: Feedback Based Counterfactual Explanations for Explainable AI. *IEEE Access* 10 (2022), 72363–72372. doi:10.1109/ACCESS.2022.3189432

[53] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. doi:10.1145/3287560.3287566

[54] Peter M. VanNostrand, Huayi Zhang, Dennis M. Hofmann, and Elke A. Rundensteiner. 2023. FACET: Robust Counterfactual Explanation Analytics. *Proc. ACM Manag. Data* 1, 4 (Dec. 2023), 242:1–242:27. doi:10.1145/3626729

[55] Suresh Venkatasubramanian and Mark Alfano. 2020. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 284–293. doi:10.1145/3351095.3372876

[56] Marco Virgolin and Saverio Fracaros. 2023. On the Robustness of Sparse Counterfactual Explanations to Adverse Perturbations. *Artificial Intelligence* 316 (March 2023), 103840. doi:10.1016/j.artint.2022.103840

[57] 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 (June 2022), 9584–9594. doi:10.1609/aaai.v36i9.21192

[58] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard journal of law & technology* 31 (April 2018), 841–887. doi:10.2139/ssrn.3063289

[59] Zijie J. Wang, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. 2023. GAM Coach: Towards Interactive and User-centered Algorithmic Recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3544548.3580816

A Additional Results

A.1 Additional Experimental Setup Details

Dataset Name	Abbr. Name	N	n
Adult	adult	45222	12
ProPublica COMPAS Recidivism Data Set	compas	6167	8
German Credit Data	german	1000	25
Give Me Credit	giveme	117831	10

Table 2. Dataset number of instances (N) and number of features (n).

A.2 Auditing Group Fairness

Below we show the number of groups extracted by associative rule mining for use in our group fairness evaluation in Sec 6.3 as well as the mean and standard error of the number of individuals per group.

Dataset	Num Groups	Mean Group Size
adult	28467	400.50 ± 7.48
compas	3620	49.90 ± 4.01
german	641204	46.27 ± 0.11
giveme	10809	10.90 ± 4.87

Table 3. Number of groups extracted by associative rule mining and their mean size and standard error.

As the German credit dataset has more than six discrete or categorical features, we show the distribution of group-wise fairness for only a subset of these features in Sec 6.3 due to limited page space in the main paper. We show the remaining features in Fig 10 for a logistic regression model. Kernel Density Estimation (KDE) for all distribution plots in this work was done with Scott's rule and a smoothing factor of 1.0. Here, we find similar results as in Sec 6.3 with some features (e.g., *NumberOfLiableIndividuals*) showing moderate to no favoring of certain values, while other features (e.g., *CheckingAccountBalanceBin* and *LoanRateAsPercentOfIncome*) showing groups with certain feature-values receiving more desirable explanations. Here, we see that individuals with a positive checking balance receive higher diversity explanations than those with a negative balance, and individuals applying for a loan rate of 2% of their income similarly receiving more diverse explanations than loan rates of other values.

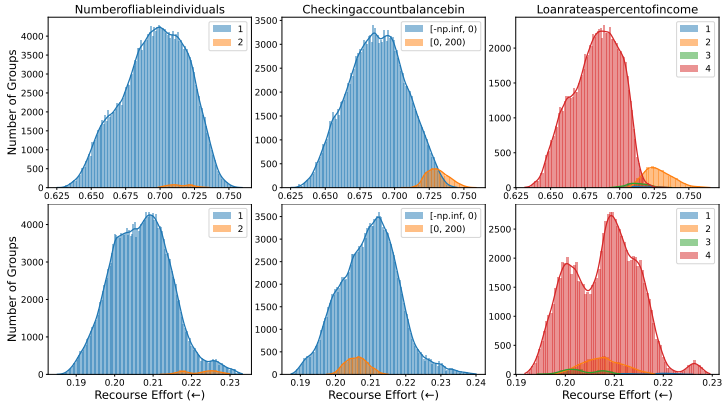

Fig. 10. Distribution of group recourse diversity (top) and recourse effort(bottom) for the German dataset.

A.3 Model Selection

Figs. 11-14 are the plots of recourse effort corresponding to the micro recourse diversity plots in our comparison of machine learning model architectures from Sec 6.4.

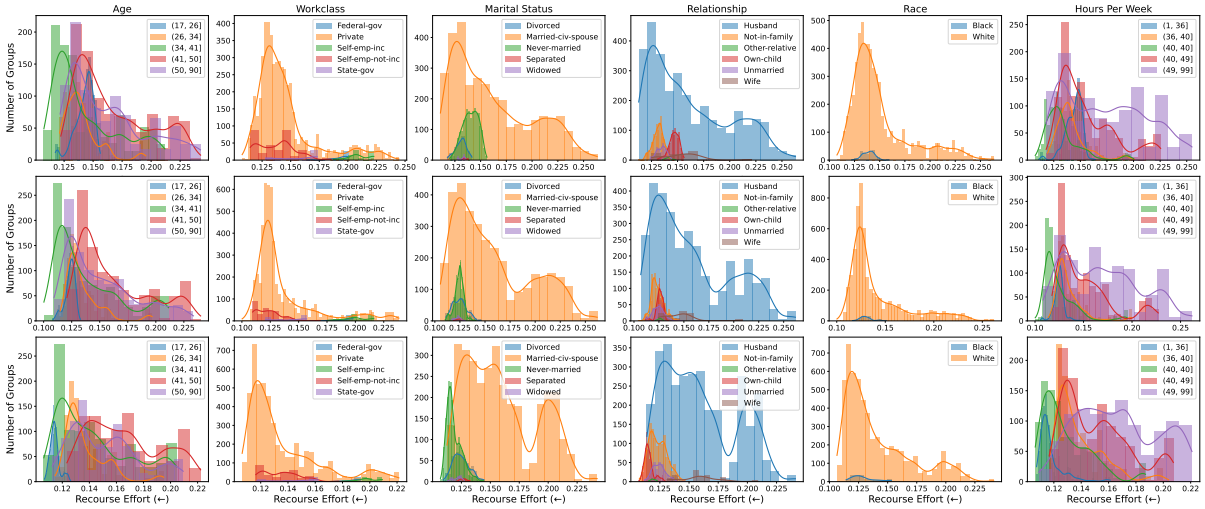

Fig. 11. Distribution of group recourse effort for the Adult dataset trained on LR (top), NN (middle), and RF (bottom).

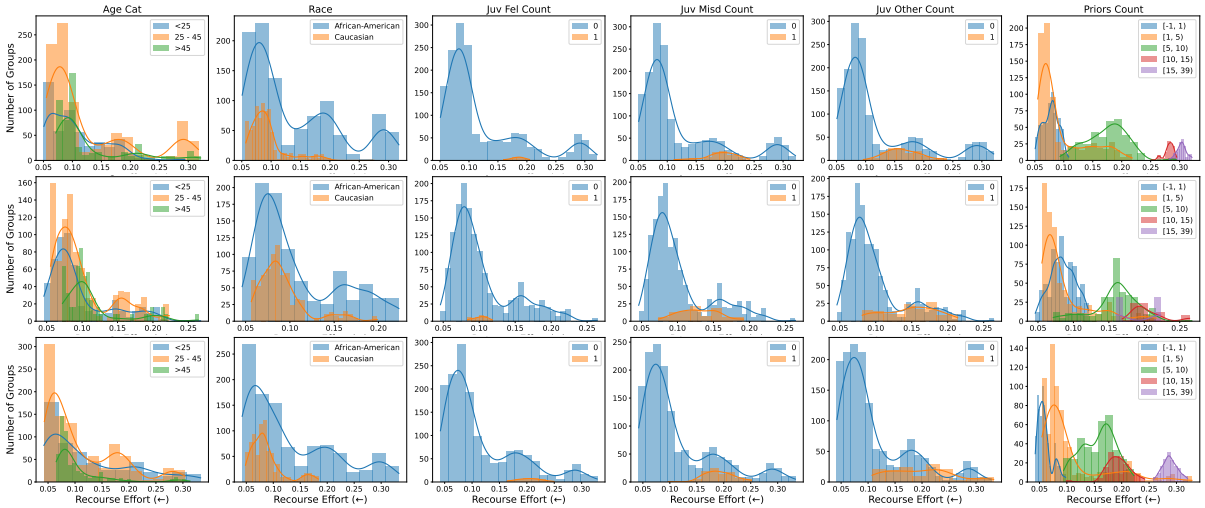

Fig. 12. Distribution of group recourse effort for the compas dataset trained on LR (top), NN (middle), and RF (bottom).

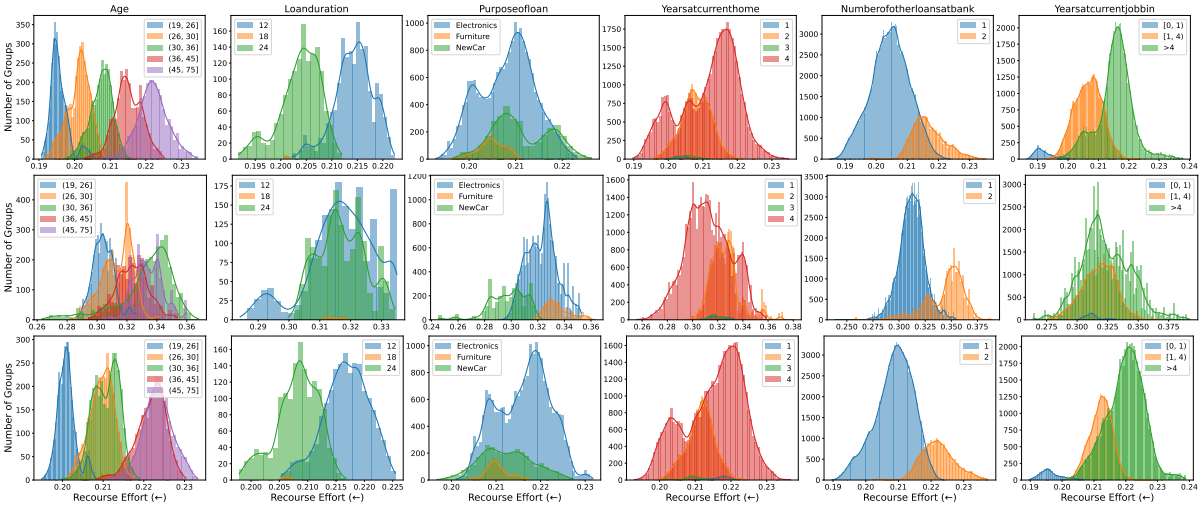

Fig. 13. Distribution of group recourse effort for the German dataset trained on LR (top), NN (middle), and RF (bottom).

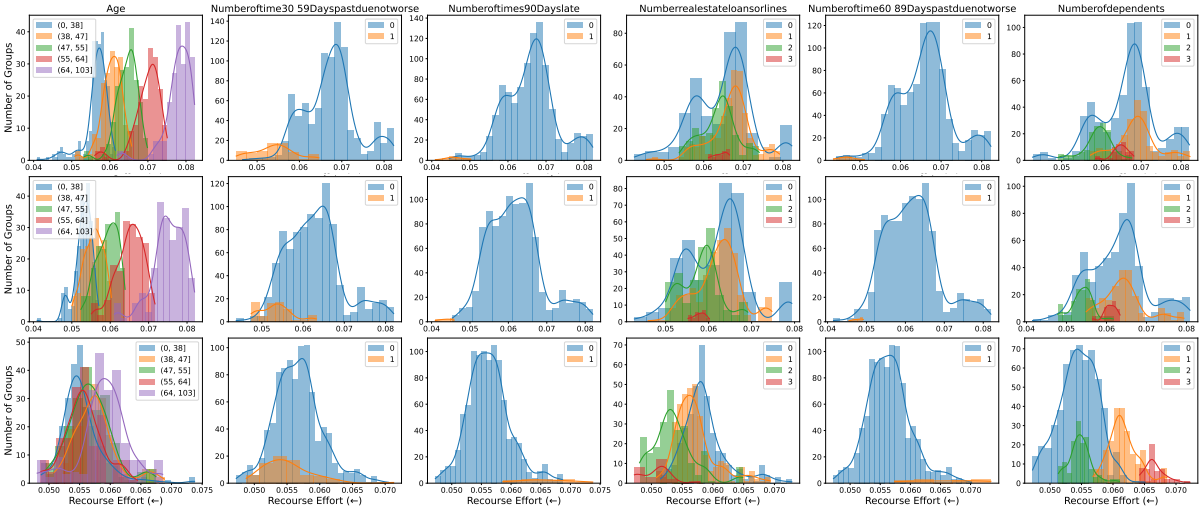

Fig. 14. Distribution of group recourse effort for the GiveMe dataset trained on LR (top), NN (middle), and RF (bottom).

A.4 Varying the Number of Counterfactuals

In Sec 6, we evaluate recourse fairness using $k = 3$ counterfactual explanations for each instance. To determine if the number of counterfactuals generated has an impact on recourse fairness, we sample a group of 20 individuals per sensitive feature intersection (e.g., 20 black men, 20 white men, etc), and then explain each individual multiple times with the Heuristic explainer while varying the number of explanations (k). Figs. 15-18 show the results of this process for the Adult, Compas, German, and GiveMe datasets on LR, NN, and RF models. We vary k 3-10 in increments of 1 and k 15-20 in increments of 5. Note the x-axis uses log scaling.

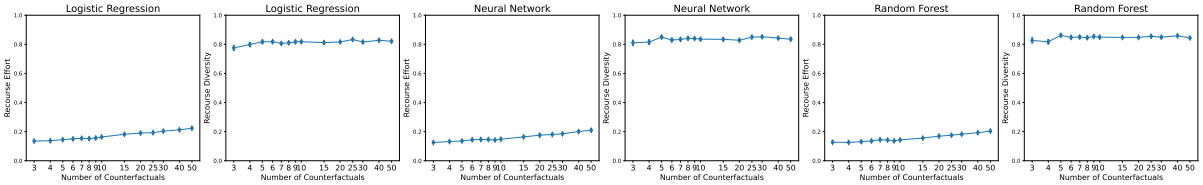

Fig. 15. Effect of varying the number of explanations on recourse effort and diversity for explanations of individuals in the Adult dataset trained on three model types. Sample contains 20 individuals per intersection of race/sex (80 total), explanations generated with the Heuristic method.

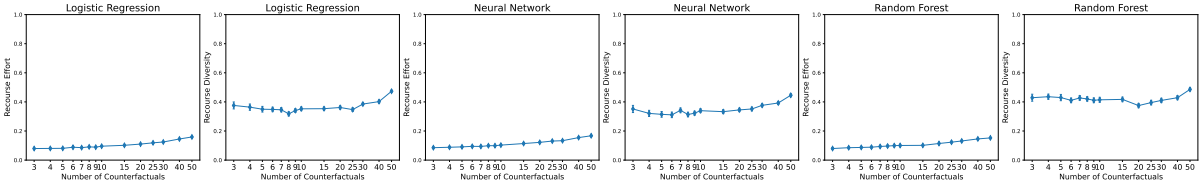

Fig. 16. Effect of varying the number of explanations on recourse effort and diversity for explanations of individuals in the Compas dataset trained on three model types. Sample contains 20 individuals per intersection of race/sex (80 total), explanations generated with the Heuristic method.

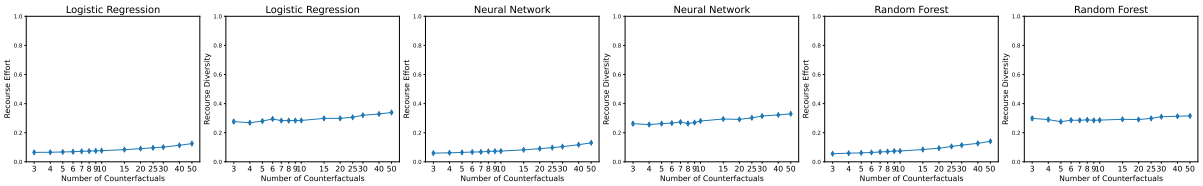

Fig. 17. Effect of varying the number of explanations on recourse effort and diversity for explanations of individuals in the German dataset trained on three model types. Sample contains 20 individuals per value of sex (40 total), explanations generated with the Heuristic method.

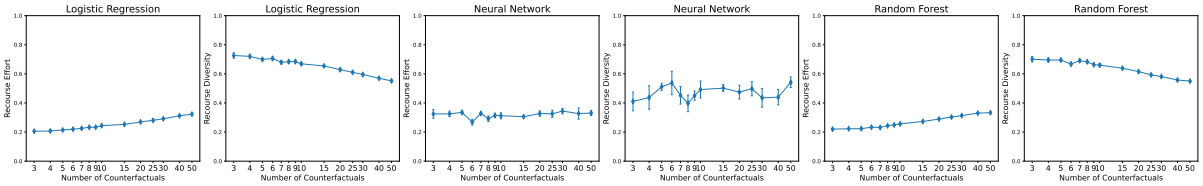

Fig. 18. Effect of varying the number of explanations on recourse effort and diversity for explanations of individuals in the GiveMe dataset trained on three model types. Sample contains 20 individuals per age bracket (100 total), explanations generated with the Heuristic method.

Considering these figures we find that varying the number of counterfactuals does not have a clear nor consistent impact on recourse diversity. This is likely because our definition of diversity (Sec 4) normalizes to the number of actions. Further, the Heuristic explainer (Sec 6) uses a random uniform sample of vectors centered on x as its starting directions, then walks along those vectors and selects the k vectors with the shortest walk to the

counterfactual class. This means that Heuristic explainer does not explore every possible direction between these vectors and thus cannot create many near-identical explanations clustered around a single direction. This is not necessarily true for all explanation techniques. For example, explanation methods that use integer programming optimization with multiple random initializations may generate multiple explanations that converge to the same local or global minima and thus find that increasing k results in decreasing diversity.

We encourage future works to benchmark the recourse fairness of different counterfactual explanation techniques under varying circumstances, including studying the effect of varying the number of explanations. That said, in practice using a large value of k is likely to result in the user being overwhelmed by the large amount of information. Thus, we postulate that small values of k (say $k < 10$) are most important to study for measuring realistic impacts on users in practice.