COUNTERFACTUAL AUC FAIRNESS: LEARNING FAIR BIPARTITE RANKING UNDER INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal fairness and AUC fairness have largely developed separately: the former studies how predictions respond to interventions on sensitive attributes, whereas the latter compares ranking performance across observed groups. We bridge these perspectives by introducing *counterfactual AUC fairness* (CF-AUC), which asks whether ranking accuracy within each positive–negative pair population remains stable when both individuals' sensitive attributes are counterfactually intervened. By fixing the pair population, CF-AUC isolates intervention-induced ranking changes from changes in population composition. We decompose observational AUC disparities into causal and population components and show that observational parity can be achieved by introducing counterfactual ranking changes while leaving factual AUC unchanged. We further characterize CF-AUC's relation to its pairwise and population variants and distinguish it from fixed-threshold decision invariance and score invariance. We then formulate fair ranking as maximizing factual AUC subject to groupwise CF-AUC constraints, providing a differentiable surrogate formulation, and bound the discrepancy between true and estimated CF-AUC gaps in terms of counterfactual-generation error, scorer sensitivity, and pairwise margin concentration near the ranking boundary. Experiments on synthetic and real data demonstrate favorable fairness–utility trade-offs.

1 INTRODUCTION

Machine-learned risk scores increasingly shape decisions in health care (Daneshjou et al., 2021; Chinta et al., 2025), lending (Khandani et al., 2010), criminal justice (Brennan et al., 2009), education (Chinta et al.) and hiring (Hoffman et al., 2018). Fairness in these settings concerns both differences across observed groups and the causal role of sensitive attributes. Causal fairness addresses the latter through interventions and counterfactual comparisons, asking whether a sensitive attribute, or an impermissible pathway emanating from it, changes a prediction or decision (Kusner et al., 2017; Zhang & Bareinboim, 2018; Nabi & Shpitser, 2018; Chiappa, 2019; Zuo et al., 2023; Zhou et al., 2024; Grari et al., 2023; Khademi et al., 2019).

Many applications use scores to prioritize individuals rather than make decisions at one fixed threshold. Clinical scores order patients for limited care (Rajkomar et al., 2018); recidivism scores may prioritize supervision or support (Barabas et al., 2018); and credit scores prioritize manual review (Fuster et al., 2022). In these settings, performance depends on the relative ordering of individuals across a range of possible operating thresholds. Bipartite ranking captures this objective by ranking positive-label cases above negative-label cases. Its standard measure, the area under the receiver operating characteristic curve (AUC), is the probability of correctly ordering a randomly drawn positive–negative pair, with half credit for ties and is independent of the choice of any single deployment threshold (Hanley & McNeil, 1982; Clémençon et al., 2008).

An extensive literature studies fairness in AUC-based ranking. Within-group AUC compares a positive and a negative individual from the same protected group, whereas cross-group AUC (xAUC) compares individuals from different groups. Existing methods constrain disparities among these quantities or improve the worst-performing group-pair comparison (Kallus & Zhou, 2019; Narasimhan et al., 2020; Vogel et al., 2021; Yao et al., 2023; Yang et al., 2023; Borkan et al., 2019; Wu et al., 2025). These observational comparisons change both the factual pair population and its sensitive-attribute assignment. Consequently, equality of group-pair AUCs does not imply that the probability

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

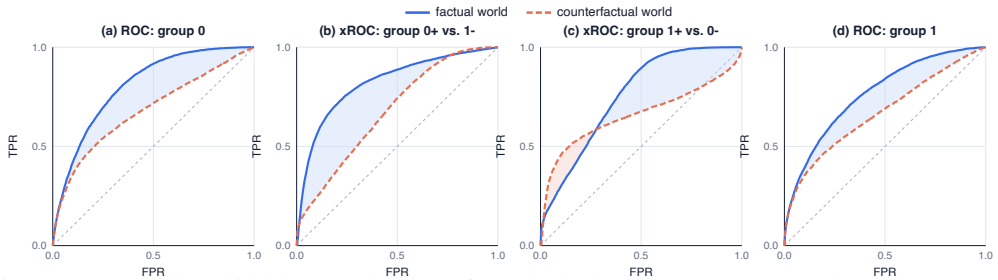


Figure 1: Factual (solid blue) and counterfactual (dashed orange) ROC and xROC curves for four protected-group pair populations. Factual labels and pair membership are held fixed while both individuals' sensitive attributes are intervened upon. Panels (a) and (d) show within-group comparisons, while Panels (b) and (c) show the two directional cross-group comparisons. Despite counterfactual invariance of all threshold-zero classifications, the factual and counterfactual AUCs differ, illustrating that counterfactual invariance of classification does not imply counterfactual invariance of ranking.

of correctly ranking a fixed factual positive–negative pair population remains unchanged under interventions on sensitive attributes.

This distinction has a concrete consequence: observational AUC parity can be attained by introducing counterfactual pairwise ranking changes, without changing overall factual AUC. We construct a shared scorer family in which closing the observational xAUC gap moves a counterfactually invariant scorer to one with a nonzero counterfactual AUC gap. A symmetric decomposition explains this phenomenon: causal responses can offset differences between factual pair populations. Conversely, a counterfactually pairwise stable ranker can have different AUCs on different populations. These observations motivate a different question:

> *Would the probability of correctly ranking the same positive–negative pair population change under interventions on their sensitive attributes?*

This question cannot be resolved by directly extending causal classification criteria to scoring rules. Invariance of a thresholded decision is too weak for ranking: two individuals can remain on the same side of the threshold while their relative order changes. Conversely, invariance of the numerical score is too strong, since strictly increasing transformations can change score values while preserving all pairwise orderings and hence the AUC. The natural object for causal AUC fairness is therefore the pairwise ordering, or its probability over a positive–negative pair population.

Counterfactual invariance of classification does not imply counterfactual invariance of bipartite ranking. Figure 1 makes this distinction concrete. The same positive–negative pair populations are compared before and after interventions on the sensitive attributes, with solid blue curves representing factual rankings and dashed orange curves their counterfactual counterparts. Panels (a) and (d) show within-group ROC curves, while Panels (b) and (c) show the two directional cross-group xROC curves. Although every individual retains the same threshold-zero classification after intervention, the pairwise rankings change substantially: the largest group-pair AUC gap is 0.143, and the population AUC decreases from 0.787 to 0.672. Thus, counterfactual fairness at a fixed classification threshold does not ensure causal fairness in ranking—a discrepancy that neither existing causal classification criteria nor observational AUC fairness criteria can capture. The structural equations and simulation protocol are given in Appendix A.

To bridge this gap, we introduce *counterfactual AUC fairness* (CF-AUC), a causal framework for fair bipartite ranking. CF-AUC distinguishes two aspects of fairness that observational comparisons cannot separate: the population defining a pair and the scoring world in which that pair is evaluated. Factual outcomes and protected-group membership select each positive–negative pair population; interventions then change the scoring inputs for those same individuals. This construction attributes changes in pairwise ranking accuracy to the intervention-induced response of the scorer, rather than to a change in the composition of the evaluation population. In particular, when an intervention also changes the outcome, redefining the pair population using counterfactual labels would evaluate a different prediction task. CF-AUC instead preserves the factual priority benchmark and asks whether the scorer recognizes it consistently across counterfactual worlds.

Our main contributions are as follows:

- We introduce *counterfactual AUC fairness (CF-AUC)*, which compares factual and counterfactual ordering probabilities over the same positive–negative pair populations under interventions on the sensitive attributes.
- We show that observational AUC parity and CF-AUC fairness do not imply one another. A symmetric decomposition of within-group and cross-group AUC disparities, together with causal examples, explains how population differences and causal responses yield different fairness conclusions.
- We formulate constrained AUC learning using estimated counterfactuals and establish a boundary-sensitive generation-error bound. Synthetic and real-data experiments examine fairness–utility trade-offs and the effects of counterfactual approximation.

2 Background

We review observational AUC fairness and counterfactual fairness. Let $A \in \{0,1\}$ denote a sensitive attribute, $Y \in \{0,1\}$ a binary outcome, and $\mathcal{X}$ the remaining observed covariates. For a scoring rule s_θ, write $S_{\theta,i} = s_\theta(A_i, \mathcal{X}_i) \in \mathbb{R}$ for individual i's predicted score. Throughout, $\kappa(a) = 1 - a$ denotes the binary sensitive-attribute flip.

2.1 Fair bipartite ranking

Bipartite ranking evaluates whether positive examples are ranked above negative examples. Define the tie-aware comparison $h(z) = \mathbb{I}\{z > 0\} + \frac{1}{2}\mathbb{I}\{z = 0\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function. For independently sampled units i and j, the factual AUC is $\mathrm{AUC}^{\mathrm{F}}(\theta) = \mathbb{E}\big[h(S_{\theta,i} - S_{\theta,j}) \mid Y_i = 1, Y_j = 0\big]$. It measures positive–negative ranking accuracy, assigning half credit to ties. The superscript F denotes evaluation in the factual world.

Conditioning additionally on group membership gives the observational group-pair AUC,

$$\mathrm{AUC}^{\mathrm{F}}_{ab}(\theta) = \mathbb{E}\big[h(S_{\theta,i} - S_{\theta,j}) \mid Y_i = 1, A_i = a;\, Y_j = 0, A_j = b\big], \tag{1}$$

for $a, b \in \{0,1\}$ with $\mathbb{P}(Y = 1, A = a) > 0$ and $\mathbb{P}(Y = 0, A = b) > 0$. Suppressing θ, within-group AUC parity requires $\mathrm{AUC}^{\mathrm{F}}_{00} = \mathrm{AUC}^{\mathrm{F}}_{11}$, whereas cross-group AUC (xAUC) parity requires $\mathrm{AUC}^{\mathrm{F}}_{01} = \mathrm{AUC}^{\mathrm{F}}_{10}$ (Kallus & Zhou, 2019; Narasimhan et al., 2020). Similar within-group AUCs can coexist with substantial cross-group disparities (Kallus & Zhou, 2019). These criteria compare different factual pair populations. They do not assess how interventions on A change ranking accuracy within the same pair population.

2.2 Counterfactual inference and fairness

Counterfactual inference. A structural causal model (SCM) $\mathcal{M} = (U, V, \mathcal{F}, P_U)$ consists of exogenous variables U with distribution P_U, endogenous variables V, and structural equations $\mathcal{F}$ specifying each endogenous variable as a function of its parents and exogenous inputs (Pearl, 2000). For a unit with exogenous state u, the counterfactual covariates $\mathcal{X}_{A \leftarrow a}(u)$ are obtained by replacing the structural equation for A with $A = a$, while preserving u and the remaining equations. Given factual evidence, counterfactual inference proceeds through *abduction* (infer the posterior distribution of U), *action* (replace the equation for A), and *prediction* (propagate this posterior through the modified model).

Counterfactual fairness. Counterfactual fairness requires a predictor's conditional distribution to remain unchanged under interventions on the sensitive attribute (Kusner et al., 2017). For a real-valued predictor $\hat{Y}$, this condition can be written as

$$\mathbb{P}\big(\hat{Y}_{A \leftarrow a}(U) \le z \mid \mathcal{X} = \mathbf{x}, A = a\big) = \mathbb{P}\big(\hat{Y}_{A \leftarrow \kappa(a)}(U) \le z \mid \mathcal{X} = \mathbf{x}, A = a\big), \tag{2}$$

for every $z \in \mathbb{R}$ and almost every factual context $(\mathbf{x}, a)$. Both sides condition on the same factual evidence; only the intervention changes. This formulation covers both continuous scores and discrete predictions.

In practice, SCMs are rarely fully known. Existing approaches construct predictors from inferred latent variables and observed non-descendants of A (Kusner et al., 2017), or use estimated counterfactual covariates to learn fair representations or predictors (Pfohl et al., 2019; Kim et al., 2021; Zhou et al., 2024; Grari et al., 2023; Garg et al., 2019). Path-specific approaches further distinguish permissible from impermissible effects of A (Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019).

3 Counterfactual AUC Fairness: Definition and Relations

We define CF-AUC on factual pair populations, explain its relationship to observational AUC disparities, and extend the criterion to selected causal pathways.

3.1 Counterfactual AUC Fairness

We first evaluate the joint intervention of both pair members' sensitive attributes, retaining factual labels and pair membership. We work within an SCM in which the relevant counterfactual covariates are uniquely defined and measurable, and interventions on one unit do not affect other units. For a measurable scorer, define $S_\theta(a;u) = s_\theta(a, \mathcal{X}_{A\leftarrow a}(u))$. Consistency assumption gives $\mathcal{X}_{A\leftarrow a} = \mathcal{X}$ on $\{A = a\}$.

For independent draws i and j from the factual population, let $\mathcal{P}_{ab} = \mathcal{P}\big(U_i, U_j \mid Y_i = 1, A_i = a;\, Y_j = 0, A_j = b\big)$. Let $\mathcal{G}$ be the nonempty set of supported group pairs, for which both conditioning strata have positive probability. Using tie-aware comparison h from Section. 2.1, define

$$H^{\mathrm{F}}_{ij}(\theta) = h\big(S_\theta(A_i;U_i) - S_\theta(A_j;U_j)\big), \quad H^{\kappa}_{ij}(\theta) = h\big(S_\theta(\kappa(A_i);U_i) - S_\theta(\kappa(A_j);U_j)\big). \tag{3}$$

Definition 3.1 (Counterfactual AUC)**.** *The counterfactual AUC of the pair population $(a,b) \in \mathcal{G}$ under the intervention policy κ is defined as* $\mathrm{AUC}^{\kappa}_{ab}(\theta) = \mathbb{E}_{\mathcal{P}_{ab}}\big[H^{\kappa}_{ij}(\theta)\big]$.

By consistency, the factual counterpart satisfies $\mathrm{AUC}^{\mathrm{F}}_{ab}(\theta) = \mathbb{E}_{\mathcal{P}_{ab}}\big[H^{\mathrm{F}}_{ij}(\theta)\big]$, and agrees with Eq. 1. The two AUCs evaluate the same $\mathcal{P}_{ab}$ under intervention (a,b) and $(\kappa(a),\kappa(b))$. Define

$$\delta_{ab}(\theta) = \mathrm{AUC}^{\mathrm{F}}_{ab}(\theta) - \mathrm{AUC}^{\kappa}_{ab}(\theta), \quad \Delta_{ab}(\theta) = |\delta_{ab}(\theta)|. \tag{4}$$

Definition 3.2 (CF-AUC fairness)**.** *For $\varepsilon \in [0,1]$, a scoring rule is ε-CF-AUC fair if*

$$\Delta_{\max}(\theta) := \max_{(a,b)\in\mathcal{G}} \Delta_{ab}(\theta) \le \varepsilon. \tag{5}$$

CF-AUC thus requires approximately unchanged ranking accuracy within each pair population.

Pairwise ranking stability. The same-unit coupling defines the stronger diagnostic $R_{ab}(\theta) = \mathbb{E}_{\mathcal{P}_{ab}}\big[\big|H^{\mathrm{F}}_{ij}(\theta) - H^{\kappa}_{ij}(\theta)\big|\big]$. If ties have probability zero in both worlds, R_{ab} is the probability of an order reversal. Since $\Delta_{ab} \le R_{ab}$, bounding every R_{ab} by ε implies ε-CF-AUC fairness. The converse fails because oppositely directed changes can cancel in the AUC contrast.

Population aggregation. Let $\pi^+_a = \mathbb{P}(A = a \mid Y = 1)$ and $\pi^-_b = \mathbb{P}(A = b \mid Y = 0)$. Independent pair sampling gives the counterfactual AUC over the whole population as $\mathrm{AUC}^{\kappa}(\theta) = \mathbb{E}\big[H^{\kappa}_{ij}(\theta) \mid Y_i = 1, Y_j = 0\big] = \sum_{(a,b)\in\mathcal{G}} \pi^+_a \pi^-_b\, \mathrm{AUC}^{\kappa}_{ab}(\theta)$. Let $\mathrm{AUC}^{\mathrm{F}}$ be the factual population AUC. The disparity of facatual and counterfactual population AUC is $\mathrm{AUC}^{\mathrm{F}}(\theta) - \mathrm{AUC}^{\kappa}(\theta) = \sum_{(a,b)\in\mathcal{G}} \pi^+_a \pi^-_b \delta_{ab}(\theta)$, Consequently, $|\,\mathrm{AUC}^{\mathrm{F}} - \mathrm{AUC}^{\kappa}\,| \le \Delta_{\max}$, which means groupwise CF-AUC fairness implies population level. However, population AUC stability does not imply groupwise CF-AUC: the aggregate assigns little weight to small pair populations and allows signed groupwise changes to cancel.

Identification. The SCM defines the target but does not establish its identification from observational data. Under independent pair sampling, identifying $\mathcal{P}(S_\theta(\kappa(a);U) \mid A = a, Y = y)$ for

every supported factual stratum suffices to identify CF-AUC for a fixed scorer. Identifying the corresponding conditional counterfactual covariate distributions is also sufficient. Neither requires recovering individual exogenous states or individual counterfactual realizations. Because selection uses factual A and Y, marginal interventional distributions alone are generally insufficient (Shpitser & Pearl, 2008; Nasr-Esfahany et al., 2023). Appendix B.1 gives an example.

3.2 Relation to observational AUC fairness

Observational AUC parity asks whether different groups have equal ranking accuracy. CF-AUC asks whether ranking accuracy changes for the same pair population when their sensitive attributes are counterfactual value. These criteria need not agree: observational parity can conceal counterfactual pairwise ranking changes, while counterfactual stability can coexist with observed disparities.

To make this distinction precise, consider supported group pairs (a, b) and $(\kappa(a), \kappa(b))$. Suppressing θ, define

$$D_{ab}^{\text{obs}} = \text{AUC}_{ab}^{\text{F}} - \text{AUC}_{\kappa(a)\kappa(b)}^{\text{F}}. \tag{6}$$

This covers xAUC disparity when $a \neq b$ and within-group AUC disparity when $a = b$.

Proposition 3.1 (Observational parity and CF-AUC fairness)**.** *The observational disparity decomposes as*

$$D_{ab}^{\text{obs}} = D_{ab}^{\text{causal}} + D_{ab}^{\text{pop}}. \tag{7}$$

The causal component averages AUC differences between assignments within each group pair; the population component averages differences between group pairs under each fixed assignment:

$$D_{ab}^{\text{causal}} = \tfrac{1}{2}\big(\delta_{ab} - \delta_{\kappa(a)\kappa(b)}\big), \tag{8}$$

$$D_{ab}^{\text{pop}} = \tfrac{1}{2}\Big[\ \text{AUC}_{ab}^{\text{F}} - \text{AUC}_{\kappa(a)\kappa(b)}^{\kappa} + \text{AUC}_{ab}^{\kappa} - \text{AUC}_{\kappa(a)\kappa(b)}^{\text{F}}\ \Big].$$

In general, observational parity $D_{ab}^{\text{obs}} = 0$ and exact CF-AUC fairness on these two group pairs, $\Delta_{ab} = \Delta_{\kappa(a)\kappa(b)} = 0$, do not imply one another.

Thus, equal observed AUCs can result from cancellation between causal responses and population differences. Conversely, when both CF-AUC gaps vanish, the observed disparity equals the population component, which may remain nonzero. Proofs and an additional bound appear in Appendix B. Figure 2 illustrates two causal mechanisms behind this distinction.

Equal AUCs can conceal counterfactual changes. Positive and negative cases from different groups can have different covariate distributions, even when changing an individual's A would leave their covariates unchanged. Figure 2(a) illustrates this situation: a background factor U_A affects both group membership A and the covariates $\mathcal{X}$. A scorer using only $\mathcal{X}$ can therefore have unequal group-pair AUCs despite being counterfactually invariant. Adding a group-dependent adjustment, $s_\theta(a, x) = x + \theta a$, can close the observational xAUC gap. However, this adjustment changes when A is intervened and can alter the rankings of the same pair. Appendix C shows that observational parity can be achieved this way while factual AUC remains unchanged.

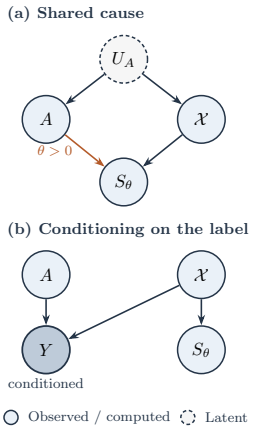


Figure 2: Relevant causal subgraphs. Dashed nodes are latent; shaded Y denotes conditioning.

Unequal AUCs can persist without counterfactual changes. Consider the scorer $s_\theta(a, x) = x$, and suppose $Y = 1$ when $\mathcal{X}$ exceeds a group-dependent threshold. Even if the groups initially have identical covariate distributions, a lower threshold admits lower-scoring individuals into the positive class, whereas a higher threshold retains higher-scoring individuals in the negative class. Comparing these cases can produce unequal directional xAUCs. Yet intervention on A leaves every pair's score unchanged, so $\Delta_{\max} = 0$. Figure 2(b) depicts this mechanism; Appendix D gives the explicit construction.

Remark 3.1. *The population component depends on the scorer and need not represent permissible differences. Moreover, $D_{ab}^{\text{causal}} = 0$ does not imply CF-AUC fairness: equal, nonzero signed contrasts can cancel within this component. The absolute groupwise gaps remain the fairness targets.*

3.3 Relation to score and decision invariance

Pointwise score invariance, $S_\theta(a;u) = S_\theta(\kappa(a);u)$, implies $R_{ab} = \Delta_{ab} = 0$ for every supported pair group. The converse fails. For a given pair population, applying the same strictly increasing transformation to both scores preserves their ordering without requiring score equality. Fixed-threshold decision invariance and CF-AUC fairness are also incomparable: unchanged decisions can conceal order reversals, whereas rankings can remain unchanged while decisions change. Figure 1 illustrates the former; Appendix B gives both counterexamples.

3.4 Path-dependent CF-AUC

Not every causal pathway from A to a score is necessarily discriminatory. An application may permit some effects while requiring invariance to others (Nabi & Shpitser, 2018; Chiappa, 2019). We give a concrete extension using the intervention in path-dependent counterfactual fairness (Kusner et al., 2017; Zuo et al., 2023). In an acyclic SCM, fix the impermissible paths in the graph augmented with the scoring rule, and let Z collect covariates outside these paths. Assume that every permissible path from A to the score passes through Z; any direct path to the score is treated as impermissible. Holding Z at its factual value then blocks changes along permissible paths without blocking the selected impermissible paths. Define $S^{\kappa,Z}_{\theta,i} = s_\theta\big(\kappa(A_i), \mathcal{X}_{A\leftarrow\kappa(A_i),\, Z\leftarrow Z_i}(U_i)\big)$ and use these scores in the counterfactual pair comparison. Then path-dependent CF-AUC can be defined accordingly, with the same factual AUCs and pair populations. Setting $Z = \varnothing$ recovers the total-flip criterion.

The path selection and conditional counterfactual distributions given factual A, Y require substantive justification. When no such separating set Z exists, more general path-specific interventions require an explicit nested counterfactual construction and its identification assumptions (Avin et al., 2005; Chiappa, 2019). Appendix F gives a mediator example and explains this distinction. The learning objective below uses the resulting scores directly. The decomposition above and our experiments concern the full intervention on A.

4 Learning with Estimated Counterfactuals

We learn a scorer that preserves factual AUC while controlling groupwise CF-AUC gaps. Generated counterfactual covariates provide one way to estimate the conditional score distributions required by the criterion. We first formulate this learning objective, then analyze how approximation error affects its fairness target.

4.1 Learning objective

For unit i, let $\mathcal{X}^\kappa_i = \mathcal{X}_{A\leftarrow\kappa(A_i)}(U_i)$ denote the true counterfactual covariates. A fitted generator $\hat{G}$ produces estimated covariates $\widetilde{\mathcal{X}}^\kappa_i$ from factual records, keeping non-descendants of A unchanged. We fix $\hat{G}$ and cache training counterfactuals before optimizing θ. Factual covariates, labels, and pair membership remain unchanged. The corresponding estimated counterfactual score is $\widetilde{S}^\kappa_{\theta,i} = s_\theta(\kappa(A_i), \widetilde{\mathcal{X}}^\kappa_i)$. For a pair drawn from $\mathcal{P}_{ab}$, define $M^{\mathrm{F}} = S_\theta(a;U_i) - S_\theta(b;U_j)$, $M^\kappa = S_\theta(\kappa(a);U_i) - S_\theta(\kappa(b);U_j)$, $\widetilde{M}^\kappa = \widetilde{S}^\kappa_{\theta,i} - \widetilde{S}^\kappa_{\theta,j}$. Set $\tilde{\delta}_{ab}(\theta) = \mathbb{E}_{\mathcal{P}_{ab}}\Big[h(M^{\mathrm{F}}) - h(\widetilde{M}^\kappa)\Big]$, $\widetilde{\Delta}_{ab}(\theta) = |\tilde{\delta}_{ab}(\theta)|$. Tildes denote evaluation using estimated counterfactuals. Expectations also include generation randomness, independently sampled across units. The true-SCM target remains $\Delta_{ab}(\theta)$.

Let $\widehat{\delta}_{ab}$ be the training pair average estimating $\tilde{\delta}_{ab}$, with every required factual stratum represented in the sample. We target

$$\begin{aligned} \max_{\theta\in\Theta} \quad & \widehat{\mathrm{AUC}}^{\mathrm{F}}(\theta) \\ \text{s.t.} \quad & |\widehat{\delta}_{ab}(\theta)| \le \varepsilon, \qquad (a,b) \in \mathcal{G}. \end{aligned} \tag{9}$$

The objective averages all factual positive–negative pairs, retaining their empirical group proportions. For optimization, we use factual pairwise logistic loss and replace h in the constraints by $h_\tau(m) = \big(1 + e^{-m/\tau}\big)^{-1}$, $\tau > 0$. This formulation permits different constrained AUC optimization methods (Narasimhan et al., 2020; Yao et al., 2023). Final evaluation uses h. Appendix G.1 gives the implementation and the smooth-to-hard approximation.

Path-dependent counterfactual generator. For Section 3.4, a generator supporting the joint intervention $A \leftarrow \kappa(A_i)$, $Z \leftarrow Z_i$ can be used. After abduction, fix Z to its observed value, flip A, and propagate the remaining structural equations using the same exogenous draw. Cache these covariates, train and evaluate the scores in Eq. 3.4. Then replace $\widetilde{S}^{\kappa}_{\theta,i}$ in the same pair averages and constraints.

4.2 The ranking boundary and generation error

Causal movement and the ranking boundary. Changes in score values affect AUC only when they change a pair comparison. The following bound makes this mechanism explicit. For θ and $(a, b) \in \mathcal{G}$, and use the same exogenous states in M^{F} and M^{κ}. Let $\Xi_{ab} = \mathbb{E}_{\mathcal{P}_{ab}}|M^{\kappa} - M^{\mathrm{F}}|$.

Theorem 4.1 (Causal movement and the ranking boundary)**.** *If $\Xi_{ab} < \infty$, then for every $r > 0$, $\Delta_{ab} \le R_{ab} \le \mathbb{P}_{\mathcal{P}_{ab}}(|M^{\mathrm{F}}| \le r) + \frac{\Xi_{ab}}{r}$. The bound holds with the tie convention in Sec. 2.1.*

The two terms measure factual mass near the ranking boundary and large intervention-induced margin changes, respectively. The proof is in Appendix E. The same argument controls error from counterfactual approximation.

Generation-error transfer. For the fitted generator, scorer θ, and $(a, b) \in \mathcal{G}$. Couple true and estimated covariates for each unit, preserving their respective conditional distributions given its factual record $(A_i, \mathcal{X}_i, Y_i)$, independently across units Assume s_θ is L_θ-Lipschitz in the covariates that may change, uniformly over A and the unchanged covariates. Define $e_{ab}(\theta) = L_\theta \mathbb{E}_{\mathcal{P}_{ab}}\Big[\, \|\widetilde{\mathcal{X}}^{\kappa}_i - \mathcal{X}^{\kappa}_i\| + \|\widetilde{\mathcal{X}}^{\kappa}_j - \mathcal{X}^{\kappa}_j\| \,\Big]$. Then the mean margin error is bounded by $\mathbb{E}_{\mathcal{P}_{ab}}|\widetilde{M}^{\kappa} - M^{\kappa}| \le e_{ab}(\theta)$.

Proposition 4.1 (Boundary-adaptive generation-error transfer)**.** *Under the stated assumptions, if $e_{ab}(\theta) < \infty$, then for every $r > 0$,*

$$|\Delta_{ab}(\theta) - \widetilde{\Delta}_{ab}(\theta)| \le \mathbb{E}_{\mathcal{P}_{ab}}\big|h(M^{\kappa}) - h(\widetilde{M}^{\kappa})\big| \le \mathbb{P}_{\mathcal{P}_{ab}}(|\widetilde{M}^{\kappa}| \le r) + \frac{e_{ab}(\theta)}{r}. \tag{10}$$

If $e_{ab}(\theta) = 0$, the comparisons agree almost surely and the fairness gaps coincide, including at ties.

A comparison can change only if the estimated margin is near zero or the margin error is large. Generation quality therefore affects CF-AUC through both scorer sensitivity and the distribution of pair margins. Oppositely directed changes can cancel, making the gap discrepancy smaller than the mean comparison discrepancy. Appendix G.2 proves the result and illustrates why covariate error alone cannot control CF-AUC error.

5 Experiments

In this section, we evaluate the fairness–utility trade-off of CF-AUC and examine how counterfactual generation error affects learning and fairness evaluation.

Metrics. We report test factual AUC and the maximum absolute CF-AUC gap across the four factual group-pair populations. We apply κ to both members, retain factual labels and pair membership, and use h for all AUC evaluations.

Methods. We compare **Unconstrained AUC** with **xAUC parity** and **Within-group AUC parity**, which constrain the corresponding observational disparities; **CF classification invariance**, which constrains decision changes at a fixed threshold; **CF score invariance**, which constrains the root-mean-square change in normalized scores; **Population CF-AUC**, which constrains the aggregate AUC change; and **CF-AUC (ours)**, which constrains each group-pair change. We vary the tolerance of each constrained method and compare achieved test utility and fairness. Appendix H.2 specifies the constraints and implementation.

5.1 SYNTHETIC DATASET

We consider a linear-logit SCM with heterogeneous effects of A on its descendants and a nonlinear SCM with a latent-risk measurement and an outcome mediator. We use a linear scorer and a network with one hidden $\tanh$ layer, respectively. For each setting, we conduct ten runs, each with 6,000 training, 8,000 validation, and 20,000 test units. Counterfactual covariates are cached before scorer training. Appendix H.1 gives the structural equations.

Fairness–utility trade-offs with true counterfactuals. CF-AUC substantially reduces groupwise gaps with little loss in factual AUC (Figure 3). In both linear and nonlinear setting, it retains higher AUC than baseline methods at comparable maximum CF-AUC gaps. While population CF-AUC and classification invariance can closely satisfy their own criteria, they leave groupwise ranking changes. The observational cross-group and within-group AUC criteria cannot achieve a small CF-AUC gap. These results illustrate why aggregate, decision-level and observational AUC constraints do not ensure groupwise CF-AUC fairness.

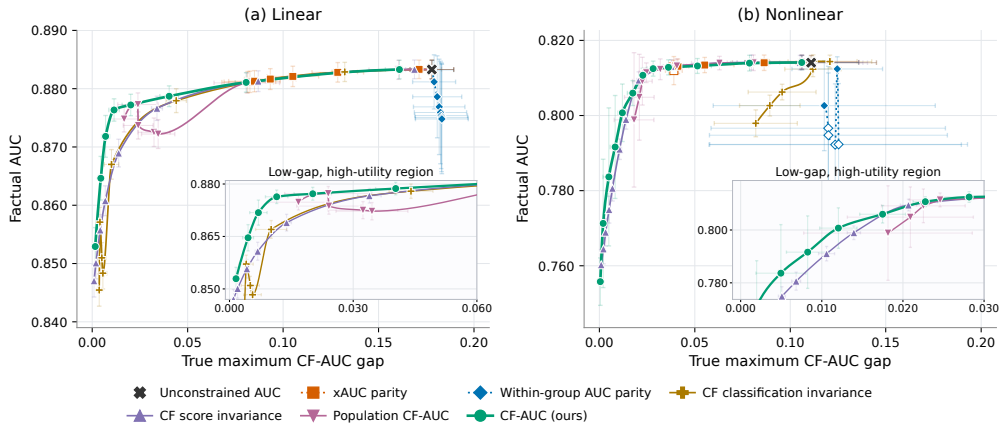


Figure 3: Fairness–utility trade-offs with ground-truth counterfactuals. Points show means over ten runs; error bars show one standard deviation. Insets enlarge the low-gap region. Hollow markers in (b) indicate a failed feasibility check in at least one run.

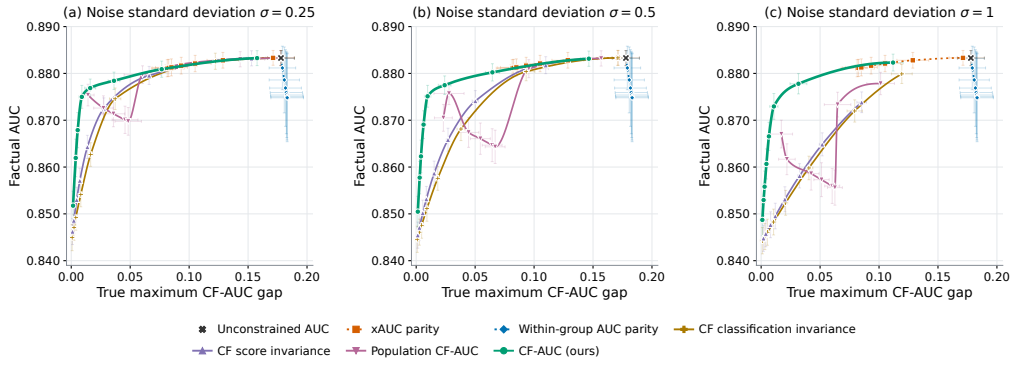


Figure 4: Fairness–utility trade-offs under controlled generation error in the linear SCM. They correspond to noise levels $\sigma = 0.25, 0.5, 1$, respectively. Fairness gaps are evaluated using ground-truth counterfactuals.

Performance under controlled generation error. In the linear SCM, we add zero-mean Gaussian noise to counterfactual covariates, with standard deviation σ measured in factual training standard-deviation units. At each noise level, we retrain all counterfactual methods across their constraint grids using the approximate generator $\hat{G}_\sigma$. Figure 4 shows that CF-AUC retains a favorable utility–fairness

trade-off at $\sigma = 0.25, 0.5, 1$: it achieves higher factual AUC than competing counterfactual constraints at comparable low true CF-AUC gaps. Appendix H.3 provides the noise model, retraining protocol, and additional results for more levels of σ.

Understanding generation error transfer. Using the CF-AUC scorers and noise perturbations on the counterfactuals in the linear SCM, we compare evaluation under true and generated counterfactuals. At comparable margin error, larger estimated-margin boundary mass is associated with more pairwise comparison disagreement (Figure 5). Gap discrepancies show weaker separation, consistent with cancellation between oppositely directed changes. These observations illustrate the mechanism in Proposition 4.1. Appendix H.4 gives the metric definitions and further comparisons.

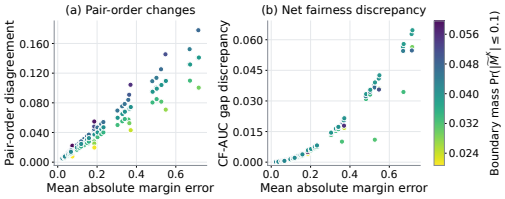


Figure 5: Generation-error transfer for CF-AUC: (a) pairwise comparison disagreement and (b) CF-AUC gap discrepancy between true and generated counterfactuals, plotted against mean absolute margin error. Darker colors indicate a larger fraction of generated margins within $r = 0.1$ of zero.

5.2 Real Data

Adult. We evaluate on Adult (Becker & Kohavi, 1996), with sex as the sensitive attribute A and income above \$50K as the positive label, $Y = \mathbb{I}\{\text{income} > \$50\text{K}\}$. We use the label-free DCE-VAE implementation of Zuo et al. (2023), based on Kim et al. (2021), and an MLP scorer following Zhou et al. (2024). Fairness comparisons are conditional on $\hat{G}$. Implementation details appear in Appendix I.

Figure 6 shows that CF-AUC retains higher utility than score invariance at comparable low gaps. Classification invariance attains similar utility but leaves larger groupwise gaps. Observational parity and population CF-AUC constraints also leave substantial groupwise changes. These results support directly constraining groupwise ranking accuracy under the modeled counterfactuals.

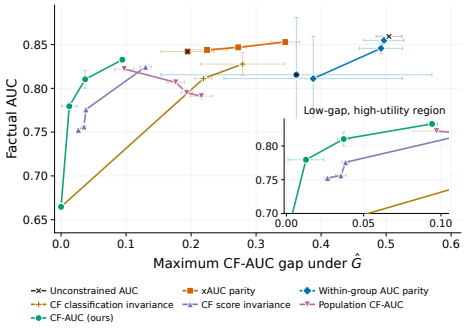


Figure 6: Fairness–utility trade-offs on Adult, evaluated using $\hat{G}$.

COMPAS. We also do the experiments on the ProPublica COMPAS dataset (ProPublica, 2016), using race as the protected attribute and two-year recidivism as the outcome. The experimental results and details are in Appendix J.

6 Conclusion and Discussion

We introduced CF-AUC fairness to connect causal fairness with bipartite ranking by comparing ordering probabilities across counterfactual worlds on specific pair populations. Counterfactual classification fairness at a fixed threshold does not imply CF-AUC fairness, while pointwise score invariance is stronger than required. Our decomposition further separates causal ranking changes from population differences in observational AUC disparities. We formulated constrained AUC learning, derived a boundary-sensitive generation-error bound, and demonstrated favorable fairness–utility trade-offs on synthetic and real data.

Limitations. CF-AUC controls groupwise averages and can conceal compensating pairwise changes. Besides, it depends on the estimated counterfactuals, which is still an unavoidable challenge in the community of Counterfactual Fairness. The path-dependent extension accommodates permissible effects through justified pathway choices. Future work includes empirical evaluation of this extension.

AI USE STATEMENT

The authors defined the research questions, developed the core theoretical and conceptual framework, and made the final methodological decisions. Generative AI tools assisted with intermediate mathematical derivations, suggested ingredients for proofs, and supported the development and checking of illustrative examples and proof arguments. These tools also provided feedback on research methodology and experimental design, suggested experimental parameters, and assisted with the interpretation and presentation of experimental results.

Additional uses included brainstorming, summarizing and discussing relevant literature, creating or editing auxiliary software code, drafting and revising manuscript text, improving its organization and readability, and formatting references. We did not use generative AI tools to generate synthetic datasets, implement the proposed method, assist with translation, or clean or reformat datasets. Qualitative and thematic data analysis was not applicable to this work.

All AI-assisted material was critically reviewed and revised by the authors. We independently checked the mathematical derivations and proofs, assessed experimental choices and interpretations against the results, and reviewed and tested AI-assisted code before use. The authors formulated and verified the final mathematical claims and experimental conclusions. We take full responsibility for the final submission, including all text, claims, code, figures, and other artifacts produced with generative AI assistance.

ETHICS STATEMENT

This work studies fairness in risk ranking using synthetic data and the publicly available Adult and Compas dataset. Fairness conclusions depend on causal assumptions and the factual-label benchmark, which may reflect existing biases. CF-AUC controls groupwise accuracy and does not guarantee fair treatment of every individual. Applications therefore require careful justification of the causal assumptions, evaluation population, and effects considered permissible.

REPRODUCIBILITY STATEMENT

The main text specifies the fairness criteria, learning objective, and theoretical assumptions, with proofs provided in the appendix. Implementation details appear in Appendix G.1. Experimental protocols, synthetic SCMs, Adult and Compas preprocessing are documented in Appendices H.2, H.1, I and J respectively. We report results across repeated runs and include variability estimates to support the interpretation of empirical comparisons.

REFERENCES

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 357–363, 2005.

Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency*, pp. 62–76. PMLR, 2018.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. Dataset.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 491–500, 2019. doi: 10.1145/3308560.3317593.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40, 2009.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019. doi: 10.1609/aaai.v33i01.33017801.

Sribala Vidyadhari Chinta, Zichong Wang, Zhipeng Yin, Nhat Hoang, Matthew Gonzalez, and Wenbin Zhang. Fairaied: Navigating fairness, bias, and ethics in educational ai applications.

Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif, Monique Antoinette Smith, Jun Liu, and Wenbin Zhang. Ai-driven healthcare: A review on ensuring fairness and mitigating bias. *PLOS Digital Health*, 4(5):e0000864, 2025.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. doi: 10.1214/009052607000000910.

Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A Novoa, Melissa Jenkins, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai: Assessments using diverse clinical images. *arXiv preprint arXiv:2111.08006*, 2021.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *Machine Learning*, 112(3):741–763, 2023. doi: 10.1007/s10994-022-06206-8.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3433–3443, 2019.

Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pp. 2907–2914, 2019.

Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8128–8136, 2021. doi: 10.1609/aaai.v35i9.16990.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11553.

Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5248–5255, 2020. doi: 10.1609/aaai.v34i04.5970.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25733–25754. PMLR, 2023.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Stephen R Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. Counterfactual reasoning for fair clinical risk prediction. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pp. 325–358. PMLR, 2019.

Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.

Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110(4):1–35, 2024. doi: 10.18637/jss.v110.i04.

ProPublica. compas-analysis. GitHub repository, 2016. URL https://github.com/propublica/compas-analysis. Data and analysis for "Machine Bias"; includes compas-scores-two-years.csv.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12): 866–872, 2018.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(64):1941–1979, 2008.

Robin Vogel, Aurélien Bellet, and Stéphan Clémençon. Learning fair scoring functions: Bipartite ranking under ROC-based fairness constraints. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 784–792. PMLR, 2021.

Mingyang Wu, Li Lin, Wenbin Zhang, Xin Wang, Zhenhuan Yang, and Shu Hu. Preserving AUC fairness in learning with noisy protected groups. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67427–67448. PMLR, 2025.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zhenhuan Yang, Yan Lok Ko, Kush R Varshney, and Yiming Ying. Minimax AUC fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11909–11917, 2023. doi: 10.1609/aaai.v37i10.26405.

Yao Yao, Qihang Lin, and Tianbao Yang. Stochastic methods for AUC optimization subject to AUC-based fairness constraints. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10324–10342. PMLR, 2023.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zeyu Zhou, Tianci Liu, Ruqi Bai, Jing Gao, Murat Kocaoglu, and David I. Inouye. Counterfactual fairness by combining factual and counterfactual predictions. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1517.

Zhiqun Zuo, Mahdi Khalili, and Xueru Zhang. Counterfactually fair representation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 12124–12140, 2023. doi: 10.52202/075280-0531.