CausalFairnessInAction: An Open Source Python Library for Causal Fairness Analysis

Kriti Mahajan
Cross Border Science
Amazon
kritimhj@amazon.com

Abstract

As machine learning (ML) systems are increasingly deployed in high-stakes domains, the need for robust methods to assess fairness has become more critical. While statistical fairness metrics are widely used due to their simplicity, they are limited in their ability to explain why disparities occur, as they rely on associative relationships in the data. In contrast, causal fairness metrics aim to uncover the underlying data-generating mechanisms that lead to observed disparities, enabling a deeper understanding of the influence of sensitive attributes and their proxies. Despite their promise, causal fairness metrics have seen limited adoption due to their technical and computational complexity. To address this gap, we present CausalFairnessInAction, the first open-source Python package designed to compute a diverse set of causal fairness metrics at both the group and individual levels. The metrics implemented are broadly applicable across classification and regression tasks (with easy extensions for intersectional analysis) and were selected for their significance in the fairness literature. We also demonstrate how standard statistical fairness metrics can be decomposed into their causal components, providing a complementary view of fairness grounded in causal reasoning.

1 Introduction

Statistical fairness metrics are easy to compute but only capture associations (i.e. conditional probabilities), not causality — limiting their ability to identify whether observed statistical disparities are truly caused by protected attributes or not. Causal fairness metrics, based on Structural Causal Models (SCMs) [6], overcome this by attributing observed disparities to specific sources (protected attributes, mediators or confounders). They also enable causal decompositions of statistical fairness metrics, thus delivering deeper insights into fairness audits. Despite their value, causal metrics are rarely used in practice due to technical challenges: they are harder to compute, require do-interventions, and face identifiability constraints. Each causal fairness metric often needs a custom architecture, and disagreement over the causal graph adds complexity. To address this, we introduce **CausalFairnessInAction**—the first open-source Python package to implement generalizable algorithms for key, established causal fairness metrics in the literature, including *Counterfactual Effects* [7]/[12], *Counterfactual Equalized Odds* [11], and *Counterfactual Fairness* [3] (see Table 1 and Table 2 for an overview). The package is designed to work with minimal identifiability assumptions[1], does not (necessarily) require a fully specified SCM and supports both group and individual-level fairness

[1] This is not equivalent to us making the claim that we address the challenge of identifiability or choice of causal graph. Instead, the paper focuses on implementing metrics for which identifiability constraints are not very strong and thus can be implemented for a wide variety of problems and domains without running into problems of identifiability; For causal model discovery, see packages like `CausalNex`, `DoWhy`, or `CausalML`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Expo.

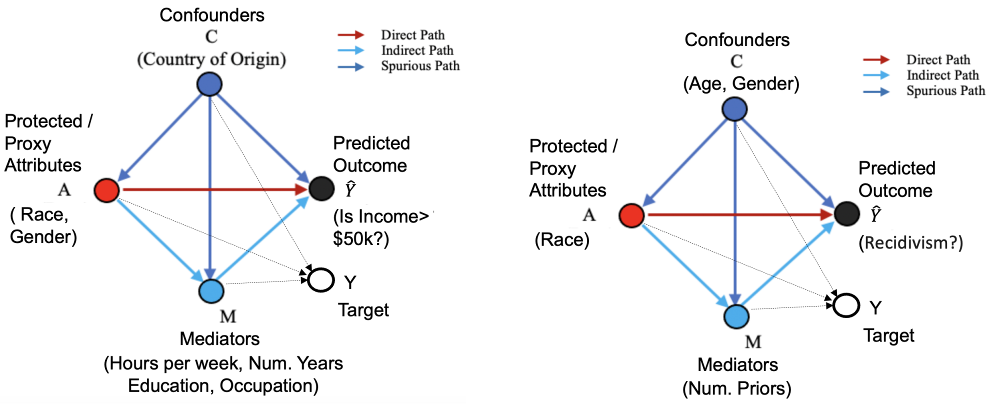
metrics. We demonstrate CausalFairnessInAction on three datasets—*Adult Income*, *COMPAS*, and *LSAC*—and provide code for replication.[2].

The paper positions itself as being a novel contribution to the causal fairness literature by implementing novel computational algorithms for computing three existing causal fairness metrics in the CausalFairnessInAction package, thus filling a critical gap in enabling practical use of causal fairness metrics[3]. This work contributes to the counterfactual measurement branch of causal fairness literature [14].

1.1 A Brief Literature Review

There has been considerable interest in the use of causal mechanisms to better understand black-box machine learning systems, and literature on causal fairness situates itself within the same. The causal fairness literature has three primary approaches for aiding algorithmic fairness assessment [14]: 1) Counterfactual measurement: helps answering what-if cause-effect questions without running randomized control trials. For instance, ceteris paribus, if a woman's gender was changed to male, would her expected income be higher?; 2)Sensitivity analysis: how sensitive a model is to latent / confounding variables (which is often the status of protected attributes). For instance, sensitivity analysis can be used to "explore how sensitive our estimate of the causal link between legal representation and guilty verdict is to different levels of jury racism" and give recommendations for altering jury selection to minimize bias [14] and 3) Impact evaluation: to measure the long term consequences of automated decision making systems through the use of interventions. Following the principle of what gets measured gets managed, we recognize that causal identification of discrimination is crucial before moving on to remedial actions and impact analysis. Thus, this paper - and package - focus on addressing the gap in practical, broad adoption of causal (counterfactual) fairness metrics by providing implementations of [12]/[7], [11] and [3].

2 Methodology



(a) Adult Income Dataset (Source: [7])

(b) COMPAS Dataset (Source: [11])

Figure 1: Standard Fairness Model

Notation and Preliminaries : Causal fairness is typically formalized using Structural Causal Models (SCMs)[6], where a Directed Acyclic Graph (DAG) represents observed variables (nodes) and their causal relationships (edges). Y is the true outcome, $\hat{Y}$ the predicted outcome, and y the favorable outcome (e.g., $y = 1$ in the Adult Income dataset). A is the set of protected attributes (e.g., race, gender), X contains all features excluding A, and a_0, a_1 denote advantaged and disadvantaged groups. The causal effect of an intervention $do(X = x)$ is expressed via the counterfactual distribution $P(Y_x = y)$, where Y_x is the outcome had X been set to x. Often, $P(y|x) = P(Y = y|X = x)$ is used interchangeably. The *Standard Fairness Model*[12] (see Figure 1 for examples) outlines three

[2] Forthcoming at: https://github.com/amazon-science/causal-fairness-in-action

[3] The implementation has intentionally been optimized so as to not require any special hardware requirements

causal pathways from A to Y and $\hat{Y}$: **Direct Path** ($A \rightarrow Y$): Captures direct discrimination (e.g., gender or race directly influencing income or recidivism), interpreted as *disparate treatment*; **Indirect Path** ($A \rightarrow M \rightarrow Y$): Effects mediated by variables like education or prior offenses, indicating *disparate impact*; **Spurious Path** ($A \leftarrow C \rightarrow Y$): Non-causal associations due to confounders (C), such as country of residence or age/gender, also contributing to *disparate impact*. Discrimination is assessed on the DAG using counterfactuals. By applying different do-interventions, specific paths can be isolated. The key idea is *ceteris paribus*, how does changing the protected attribute A affect Y or $\hat{Y}$?

2.1 CausalFairnessInAction : Definitions and Computational Algorithms

Table 1: Overview of Metrics Implemented in CausalFairnessInAction

Metric	Query Addressed	Level	Supported Metric Decomposition	Counterfactual Estimation Procedure
Counterfactual Effects (for Statistical Parity)	What would the disadvantaged (advantaged) group's **acceptance rate** be if they had the identity (A), mediating characteristics (M), or confounding characteristics (C) of the advantaged (disadvantaged) group?	Aggregate	Direct, Indirect, Spurious	Conditional probabilities (computed or estimated using GMM)
Counterfactual Equalized Odds	What would the disadvantaged (advantaged) group's **error rate** be if they had the identity (A), mediating characteristics (M), or confounding characteristics (C) of the advantaged (disadvantaged) group?	Aggregate	Direct, Spurious	Conditional probabilities (computed or estimated using GMM)
Counterfactual Fairness	What would the disadvantaged (advantaged) individual's **predicted Y** be if they had the identity (A), mediating characteristics (M), and confounding characteristics (C) of the advantaged (disadvantaged) group?	Individual	N/A	Predictions from functional relationships of fitted SCM

The core of the *CausalFairnessInAction* package is the `CausalFairnessDecomposition` class (see Table 2), built on the standard fairness model [12]. It accepts $X, Y, \hat{Y}$, lists for A, M, optionally C, derived from an SCM (algorithmically discovered or expert-curated) or a DAG, and a task-type flag (regression/classification). The class provides three main methods: `analyse_mean_difference` – causal decomposition of statistical parity; `analyse_equalized_odds` – causal decomposition of error rates; `analyse_counterfactual_fairness` – individual-level counterfactual fairness analysis. Each method compares the outcome (acceptance rate, error rates, or predicted outcome $\hat{Y}_i$) in two counterfactual worlds: one with observed A, and one with counterfactual A. The first two rely on (estimated) conditional probabilities using Gaussian Mixture Models for scalability; the third requires a DAG and fits a graphical causal model.

Table 2: Pseudo-Algorithm for Causal Fairness Metrics

A. Counterfactual Effects (Mean Difference)	**B. Counterfactual Equalized Odds (EO)**	**C. Counterfactual Fairness**
Inputs: D, A, M, C, a_0, a_1, y	**Inputs:** $D, A, C, a_0, a_1, y, \hat{f}$	**Inputs:** A, M, C, a_0, a_1, DAG
1. For each $(m, c) \in D$: - Compute: $\mathbb{E}(Y = y \mid a_0, m, c)$ - Compute: $\mathbb{E}(Y = y \mid a_1, m, c)$ **2.** Estimate via GMM: $P(m \mid a_0, c)$, $P(m \mid a_1, c)$ $P(c \mid a_0)$, $P(c \mid a_1)$ **3.** Combine expectations and probabilities to compute the counterfactual effects	**1.** For each $c_j \in D$: - Predict: $\hat{f}(c_j, a_0)$, $\hat{f}(c_j, a_1)$ - Obtain: $P(\hat{y}_{a_0,c_j})$, $P(\hat{y}_{a_1,c_j})$ **2.** Estimate via GMM: $P(c \mid a_0)$, $P(c \mid a_1)$ **3.** Combine predictions and probabilities to compute the Ctf-EO	**1.** Fit SCM using DAG and dataset D **2.** For each individual $i \in D$: - Get A_{obs} (observed) and A_{cf} (counterfactual) - Sample from SCM under: $do(A = A_{obs}) \Rightarrow D_{obs}$ $do(A = A_{cf}) \Rightarrow D_{cf}$ - Predict: $\hat{f}(D_{obs})$, $\hat{f}(D_{cf})$ - Check: $Y_{obs} \neq Y_{cf}$

2.2 Counterfactual Effects : How does the protected attribute affect the predicted outcome? Calculating Disparate Treatment, Disparate Impact and Explaining the Causal Mechanism Behind Observed Statistical Parity

Counterfactual effects [12] is a family of three causal measures of discrimination related to statistical parity, namely: **Counterfactual Direct Effect (Ctf-DE)**: measures direct discrimination along $A \rightarrow \hat{Y}$ by holding M and C constant, isolating the effect of A on $\hat{Y}$. [7] define symmetric Ctf-DE as: $\text{DE}_a^{\text{sym}}(y|a) = \frac{1}{2}\left(\text{DE}_{a_0,a_1}(y|a) - \text{DE}_{a_1,a_0}(y|a)\right)$ i.e. the net treatment , which is the difference between the positive and negative effect of protected group membership. Direct discrimination exists if $\text{DE}_a^{\text{sym}}(y|a) > 0$ i.e. the negative effect is greater than the positive effect; **Counterfactual Indirect Effect (Ctf-IE)**: measures indirect discrimination along $A \rightarrow M \rightarrow \hat{Y}$ by holding A and C fixed, capturing the effect of M on $\hat{Y}$. [7] define symmetric Ctf-IE as: $\text{IE}_a^{\text{sym}}(y|a) = \frac{1}{2}\left(\text{IE}_{a_0,a_1}(y|a) - \text{IE}_{a_1,a_0}(y|a)\right)$.Indirect discrimination exists if $\text{IE}_a^{\text{sym}}(y|a) > 0$; **Counterfactual Spurious Effect (Ctf-SE)**: measures confounding impact along $A \leftarrow C \rightarrow \hat{Y}$, varying C while fixing A and M[4]. It is given by: $SE_{a_0,a_1}(y) = P(y_{a_0}|a_1) - P(y|a_0)$. As shown in [7] **disparate treatment** (direct discrimination) exists if the symmetric difference due to A, $\text{DE}_a^{\text{sym}}(y|a)$, differs from zero. **Disparate impact** (indirect discrimination) exists if either the symmetric indirect effect, $\text{IE}_a^{\text{sym}}(y|a)$, or the spurious effect, $SE_{a_0,a_1}(y|a)$, is non-zero. Statistical disparity decomposes as: Mean $\text{Difference}_{a_0,a_1}(y) = \text{DE}_a^{\text{sym}}(y|a) + \text{IE}_a^{\text{sym}}(y|a) + SE_{a_0,a_1}(y|a)$. Without confounders, Ctf-DE and Ctf-IE reduce to the natural direct and indirect effects, respectively.

Algorithmic Procedure: [12] provide empirical formulas to estimate these effects from observed data using conditional probabilities, avoiding the need for a fully specified SCM , thus aiding ease of application (see Appendix A.1 for the empirical formulations). For each combination $m \in$ M and each combination $c \in$ C, we get a subset of D defined by (m, c). For each subset (m, c) : we calculate condition probability / expectation of the outcome of interest Y_y for a_0 and a_1 i.e. $\mathbb{E}(y|a_0, m, c)$ and $\mathbb{E}(y|a_1, m, c)$ respectively. These are the differences in the realisation of Y_y when M and C are the same but A is different. Then for each m, we get the probability of m given c for a_0 and a_1 i.e.$P(m|a_0, c)$ and $P(m|a_1, c)$ i.e. the difference in probability of m when c is the same but A is varied. Then lastly, for each c, we get it's probability for a_0 and a_1 i.e. $P(c|a_0)$ and $P(c|a_1)$ respectively. Each of these computed quantities is then combined as per (1) to (11) (see Appendix A.1) to get $\text{DE}_a^{\text{sym}}(y|a)$, $\text{IE}_a^{\text{sym}}(y|a)$, $SE_{a_0,a_1}(y|a)$ and Mean $\text{Difference}_{a_0,a_1}(y)$.

2.3 Counterfactual Equalised Odds (Ctf-EO) : How does the protected attribute affect the model error rate?

Like counterfactual effects,[11] use the standard fairness model to define three causal counterfactual metrics based on equalized error rates: **Counterfactual Direct Error Rate** ($ER^d_{a_0,a_1}(\hat{y} \mid a, y) = P(\hat{y}_{a_1,y}, (\hat{P}A \backslash X)_{a_0,y} \mid a, y) - P(\hat{y}_{a_0,y} \mid a, y)$), **Counterfactual Indirect Error Rate** ($ER^i_{a_0,a_1}(\hat{y} \mid a, y) = P(\hat{y}_{a_0,y}, (\hat{P}A \setminus X)_{a_1,y} \mid a, y) - P(\hat{y}_{a_0,y} \mid a, y)$), and **Counterfactual Spurious Error Rate** ($ER^s_{a_0,a_1}(\hat{y} \mid y) = P(\hat{y}_{a_0,y} \mid a_1, y) - P(\hat{y}_{a_0,y} \mid a_0, y)$). These measure how error rates would change if the disadvantaged (advantaged) group had the identity, mediators, or confounders of the advantaged (disadvantaged) group. They conclude that error rates driven by ER^d indicate bias, while those due to ER^i or ER^s are not discriminatory. [5]. Using these three counterfactual error metrics, [11] show that equalized odds can be broken down into direct, indirect and spurious components as follows: $ER_{a_0,a_1}(\hat{y} \mid y) = ER^d_{a_0,a_1}(\hat{y} \mid a_0, y) - ER^i_{a_1,a_0}(\hat{y} \mid a_0, y) - ER^s_{a_1,a_0}(\hat{y} \mid y)$.

Algorithmic Procedure: Unlike counterfactual effects, these metrics face a key limitation: Ctf-EO cannot reliably estimate direct, indirect, and spurious effects in the presence of mediators due to identifiability issues from conditioning on both Y and $\hat{Y}$ (whereas Counterfactual Effects condition only on $\hat{Y}$). The common fix is excluding M from features, using only protected attributes A and confounders C, enabling accurate estimation of ER^d and ER^s. This solution allows for the accurate identification and estimation of Counterfactual Direct Error Rate and Counterfactual Spurious Error Rate. However, this remedial strategy is undesirable in real world applications because the exclusion of M is likely to negatively impact the predictive performance of the model. Thus, to ensure that the metric is used correctly, we remove M from consideration, and modify [11] to estimate the simplified counterfactual error rates as follows: $ER^d_{a_0,a_1}(\hat{y} \mid a, y) = \sum_c \left(P(\hat{y}_{a_1,c}) - P(\hat{y}_{a_0,c})\right) P(c \mid a, y)$ and $ER^s_{a_0,a_1}(\hat{y} \mid y) = \sum_c P(\hat{y}_{a_1,c}) \left(P(c \mid a_1, y) - P(c \mid a_0, y)\right)$. To compute these from the observed data D, we use the following procedure: For each $c \in$ C we use the fitted estimator $\hat{f}$ as $\hat{f}(a_1, c)$ and $\hat{f}(a_0, c)$ to get $P(\hat{y}_{a_1,c})$ and $P(\hat{y}_{a_0,c})$ respectively. These quantities are the differences in the realisation of Y_y when C is the same but A is different. Then for each $c \in$ C , we get it's probability for a_0 and a_1 i.e. $P(c|a_0)$ and $P(c|a_1)$ respectively. Each of these computed quantities is then combined as per (12) to (14) (see Appendix A.2) to get $ER^d_{a_0,a_1}(\hat{y} \mid a, y)$, $ER^s_{a_0,a_1}(\hat{y} \mid y)$ and $ER_{a_0,a_1}(\hat{y} \mid y)$.

[4] Ctf-SE has no symmetric form since confounders C are non-descendants of A and remain unchanged under interventions.

[5] Definitions overlap with Ctf-DE, Ctf-IE, and Ctf-SE from Section 2.2, differing by focusing on error rates instead of mean difference.

2.4 Counterfactual Fairness

Unlike Counterfactual Equalised Odds and Counterfactual Effects, Counterfactual fairness[3] is an individual level causal fairness metric which is achieved if changing an individual i's protected attributes doesn't change the predicted outcome $\hat{Y}_i$ i.e. $P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$

Algorithmic Procedure: To empirically test for counterfactual fairness first, a fully specified SCM must be defined using a specified Directed Acyclic Graph (DAG) and dataset D. For each instance i in D, we retrieve the observed value A_{obs} and compute a counterfactual value A_{cf}. We then generate samples from the SCM through two do-interventions using the standard "abduction,action,prediction"[6] procedure. First, we perform an intervention do$(A = A_{\text{obs}})$ to produce samples for the observed state $D_{do=\text{observed}}$. Second, we perform an intervention do$(A = A_{\text{counterfactual}})$ to produce samples for the counterfactual state $D_{do=\text{cf}}$. Using these samples, we fit functions $\hat{f}(D_{do=\text{obs}})$ and $\hat{f}(D_{do=\text{cf}})$ to obtain predicted outcomes $Y_{D_{do=\text{obs}}}$ and $Y_{D_{do=\text{cf}}}$. To assess counterfactual fairness, we compare the observed and counterfactual predictions. If $Y_{D_{do=\text{obs}}} \neq Y_{D_{do=\text{cf}}}$, then the prediction function $\hat{f}$ is not counterfactually fair.

2.5 Scalability : To address scalability, the algorithms include the following optimizations: **A. GMMs for Conditional Probability Estimation.** For Counterfactual Effects, estimating probabilities via conditional expectations on a 50,000-sample dataset takes approximately 2 seconds; using GMMs reduces this to 300–500 ms, depending on the number of features. For Counterfactual Equalized Odds, estimation takes approximately 1 second, with GMMs reducing latency to 300–500 ms, depending on feature count and model complexity. **B. Parallelization of Interventions.** For Counterfactual Fairness, computing for a single instance takes approximately 10 seconds without parallelization, and about 1 second with it. Actual times vary with the number of features, interventions, and the predictive model used.

3 Results: Application of CausalFairnessInAction to Benchmark Datasets

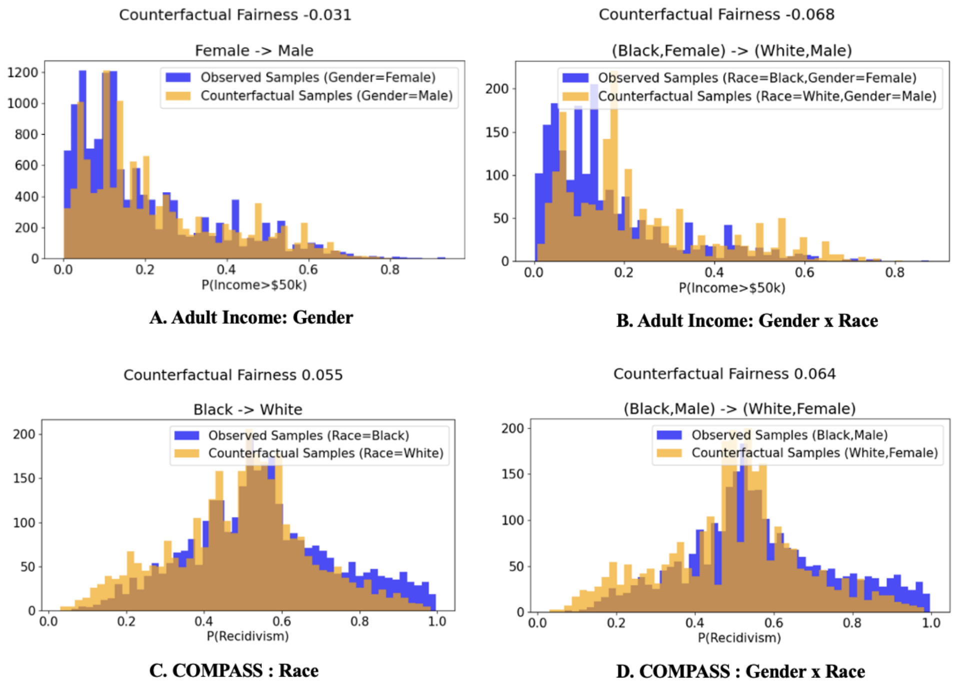
Dataset	Protected Attribute	Mean Difference	FNR	FPR	$\text{DE}_a^{\text{sym}}(y\|a)$	$\text{IE}_a^{\text{sym}}(y\|a)$	$\text{SE}_{a_0,a_1}(y\|a)$	ER^d	ER^i	ER^s	Counterfactual Fairness
Adult Income	Gender	0.203	0.410	-0.104	0.165	0.039	0.000	0.000	0.000	0.000	-0.031
Adult Income	Intersectional	0.221	0.445	-0.115	0.152	0.069	0.000	0.000	0.000	0.000	-0.068
COMPAS	Race (Black)	0.326	-0.310 (-42)	-0.253 (-0.41)	0.154	0.071	0.101	FPR: -0.297, FNR: -0.265	0	FPR: 0.113, FNR: 0.162	0.055
COMPAS	Intersectional	0.620	-0.620	-0.518	0.513	0.081	0.027	-	-	-	0.640
LSAC	Race (Black)	0.978	-	-	0.554	0.429	0.000	-	-	-	0.001
LSAC	Intersectional	0.990	-	-	0.531	0.458	0.000	-	-	-	-0.007

Table 3: Statistical and Causal Fairness Metrics

3.1 Adult Income Dataset: We fit a logistic regression using the structure and features in Fig.1.a[6] to predict $P(\text{Income} > \$50k)$. **Counterfactual Effects:** On average, women are 20.3% less likely than men to be predicted as earning above \$50k. Most of this disparity (16.5%) is due to *disparate treatment* ($\text{DE}_a^{\text{sym}}(y \mid a)$), meaning that simply having the social identity of a woman lowers $P(\text{Income} > \$50k)$. The remaining 3.9% is due to *disparate impact* ($\text{IE}_a^{\text{sym}}(y \mid a)$) via M (years of education and occupation associated with women); **Counterfactual Equalized Odds:** When refitting the model without M, the predictor always outputs 0, making the equalized odds and its decomposition uninformative. We thus cannot determine whether observed disparities stem from disparate treatment or impact; **Counterfactual Fairness:** As shown in Fig.2.A, the observed and counterfactual distributions do not overlap—changing a woman's gender to male shifts $\hat{Y}$ rightward, increasing $P(\text{Income} > \$50k)$. Hence, the fitted logistic regression is not counterfactually fair.

3.2 COMPAS Recidivism Dataset: We fit a logistic regression using the structure and features in Fig.1.b **Counterfactual Effects:** Black individuals are 32.6% more likely than white individuals to be predicted as high-risk for recidivism. Most of this is due to *disparate treatment* (15.4%), meaning that being Black alone increases $P(\text{Recidivism})$. *Disparate impact* comes from both M and C: confounders like age and gender raise risk by $\sim 10\%$ (spurious effect), and M contributes an additional 7.1%. **Counterfactual Equalized Odds:** Excluding M does not make the model naive, though it increases error rates. Decomposing FPR/FNR shows most of the disparity stems from direct discrimination: 29.7% of the 41% FPR and 26.5% of the 42% FNR. **Counterfactual Fairness:** The

[6] Country of residence is included as a cause of gender to replicate Zhang and Bareinboim, 2018



A. Adult Income: Gender

B. Adult Income: Gender x Race

C. COMPASS : Race

D. COMPASS : Gender x Race

Figure 2: Counterfactual Fairness

COMPAS model is not counterfactually fair (see Fig. 2.C)—changing race from Black to white shifts the distribution of $\hat{Y}$ leftward, decreasing $P(\text{Recidivism})$.

3.3 Law School Admission Council (LSAC) Dataset: We fit a Random Forest regressor on GPA, LSAT, Race, and Gender to predict average grade, where $A =$(Race, Gender), $M =$(GPA, LSAT), and no C. **Counterfactual Effects:** The predicted average grade for the white subgroup is 0.978 higher than for the Black subgroup. Both *disparate treatment* and *disparate impact* are significant, with most of the gap (0.55) due to direct discrimination. Since there are no confounders, the direct and indirect effects correspond to the natural direct and indirect effects. **Counterfactual Equalized Odds:** Not applicable since this is a regression task. **Counterfactual Fairness:** The model is counterfactually fair, illustrating that fairness can differ at the individual vs. group level.[7]

3.4 Intersectional Causal Fairness: This package supports intersectional analysis to detect potential "double" or higher-order discrimination. **Adult Income:** Black women are 22.1% less likely than white men to have $P(\text{Income} > \$50k)$—2% more than the non-intersectional gender gap—with direct effect being the largest contributor. The model is also more counterfactually unfair: changing a Black woman's identity to a white man increases $P(\text{Income} > \$50k)$ by 6% (vs. 2% non-intersectionally) (see Fig. 2.B). **COMPAS:** The mean difference in $P(\text{Recidivism})$ between Black men and white women (60%) exceeds the non-intersectional racial gap, with direct discrimination as the main driver. Counterfactual unfairness also rises by $\sim 11\%$: changing a Black man's identity to a white woman lowers predicted recidivism by 64% (vs. 55%) (see Fig. 2.D). **LSAC:** The mean difference between Black women and white men is slightly higher than the non-intersectional comparison (0.99 vs. 0.978), again mainly due to direct discrimination. As in the non-intersectional case, the model remains counterfactually fair.

[7] Our experiments also show that "fairness through unawareness"—excluding A from training—can worsen fairness. For example, excluding Race in the LSAC dataset leads to a counterfactual fairness score of -0.50, meaning changing a Black individual's race to white increases the predicted average grade by 0.50.

4 Limitations of Causal Fairness

Generally, 1) deciding the right causal model from competing models of bias or achieving causal fairness simultaneously across multiple competing models remains an active area of research and 2) defining a hypothetical intervention on protected attributes remains a fraught process. The example application to benchmark datasets highlighted how 1) lack of identifiability can limit analysis [4] and 2) lack of methods for falsifying DAGs in the presence of competing causal models can lead to disagreements about the validity of the conclusions. For example, in the Adult Income dataset, identifiability issues prevented the causal decomposition of equalized odds. For counterfactual effects, the DAG must be Markovian; otherwise, counterfactual probabilities cannot be empirically estimated[8]. Extending the three discussed metrics to path-specific discrimination [6] is also limited by stricter identifiability constraints. Hence, causal fairness metrics should be applied cautiously.

5 Conclusion

This paper introduced CausalFairnessInAction - the first open source generalizable implementation for calculating key causal fairness metrics and applied it to 3 fairness benchmarking datasets. The application to benchmark datasets demonstrated how CausalFairnessInAction provides practitioners with the actionable insight - for example, at the very least the Adult Income model must eliminate at least 16.5% difference in statistical parity, while the COMPAS model needs to address 15.4% disparity in statistical parity and 29.7-26.5% in error rates (all of which can be attributed to direct discrimination), but this varies intersectionally.

But now that the causal bias has been detected, what remedial steps can practitioners take to achieve causal fairness? The easiest way to achieve causal fairness is to train an estimator using only the observable non-descendants of A [3]. However, as most observable features are likely to be descendants of A, this strategy is unsuitable. [9] Latent variables which are non-descendants of A but affect X and Y can be used to train counterfactually fair estimators [3]. Counterfactual effects [12] or counterfactual error rates [11] can be used for feature and sample selection to minimize direct, indirect and spurious discrimination. Using counterfactual fairness , a multi-world causal fairness penalty can be created to achieve counterfactual fairness under competing SCMs [8].While addressing causal bias correction algorithms is out of scope for the current paper, this is an active area of research in the causal fairness literature which we aim to incorporate into forthcoming versions of the package along side sensitivity and impact analysis. Future work will expand the metrics available and extend the package to include methods for bias reduction in the causal fairness literature.

References

[1] Barocas, S., Hardt, M. &; Narayanan, A., Fairness and Machine Learning. Fairness and machine learning. Available at: https://fairmlbook.org/ [Accessed September 17, 2022]. Castelnovo, A. et al., 2022. A clarification of the nuances in the fairness metrics landscape. Nature News. Available at: https://www.nature.com/articles/s41598-022-07939-1 [Accessed September 17, 2022].

[2] Kilbertus, N. et al., 2018. Avoiding discrimination through causal reasoning. arXiv.org. Available at: https://arxiv.org/abs/1706.02744 [Accessed September 17, 2022].

[3] Kusner, M.J. et al., 2018. Counterfactual fairness. arXiv.org. Available at: https://arxiv.org/abs/1703.06856 [Accessed September 17, 2022].

[4] Makhlouf, K., Zhioua, S. &; Palamidessi, C., 2022. Survey on causal-based machine learning fairness notions. arXiv.org. Available at: https://arxiv.org/abs/2010.09553 [Accessed September 17, 2022].

[5] PARK, K.E.V.I.N.A. &; QUERCIA, R.O.B.E.R.T.O.G., Who lends beyond the Red Line? - university of Pennsylvania. Available at: https://penniur.upenn.edu/uploads/media/Park_Quercia.pdf [Accessed September 16, 2022].

[6] Pearl, J., Causality, 2nd edition, 2009. Available at: http://bayes.cs.ucla.edu/BOOK-2K/ [Accessed September 17, 2022].

[8] In the presence of unobserved confounding, counterfactual effects may be estimated using counterfactual randomization [12], which is not implemented here.

[9] Linear regression which includes the protected attributes is guaranteed to be counterfactually fair [3]

[7] Plecko, D. &; Bareinboim, E., Causal fairness analysis. Available at: https://causalai.net/r90.pdf [Accessed September 16, 2022].

[8] Russell, C. et al., 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. Advances in Neural Information Processing Systems. Available at: https://papers.nips.cc/paper/2017/hash/1271a7029c9df08643b631b02cf9e116-Abstract.html [Accessed September 17, 2022].

[9] Spielkamp, M., 2020. Inspecting algorithms for bias. MIT Technology Review. Available at: https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/ [Accessed September 17, 2022].

[10] Williams, D.R., Priest, N. &; Anderson, N.B., 2016. Understanding associations among race, socioeconomic status, and Health: Patterns and prospects. Health psychology : official journal of the Division of Health Psychology, American Psychological Association. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817358/ [Accessed September 17, 2022].

[11] Zhang, J. &; Bareinboim, E., 2018. Equality of opportunity in classification: A Causal approach. Advances in Neural Information Processing Systems. Available at: https://proceedings.neurips.cc/paper/2018/hash/ff1418e8cc993fe8abcfe3ce2003e5c5-Abstract.html [Accessed September 17, 2022].

[12] Zhang, J. &; Bareinboim, E., 2018. Fairness in decision-making - the causal explanation formula: Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence and thirtieth innovative applications of Artificial Intelligence Conference and eighth AAAI symposium on educational advances in artificial intelligence. Guide Proceedings. Available at: https://dl.acm.org/doi/abs/10.5555/3504035.3504283 [Accessed September 17, 2022].

[13] Zhang, L., Wu, Y. &; Wu, X., 2016. A causal framework for discovering and removing direct and indirect discrimination. arXiv.org. Available at: https://arxiv.org/abs/1611.07509 [Accessed September 17, 2022].

[14] Kusner, M.J. &; Loftus, J.R., 2020. The Long Road to fairer algorithms. Nature News. Available at: https://www.nature.com/articles/d41586-020-00274-3 [Accessed September 17, 2022].

A Appendix

A.1 Counterfactual effects: Empirical Formulations

Given that protected group membership can have positive and negative impacts, the direct disadvantage due to protected group membership is given by:

$$DE_{a_0,a_1}(y|a) = P(y_{a_1}, m_{a_0}, c_a|a) - P(y_{a_0}, m_{a_0}, c_a|a) \tag{1}$$

while direct advantage is given by

$$DE_{a_1,a_0}(y|a) = P(y_{a_0}, m_{a_1}, c_a|a) - P(y_{a_1}, m_{a_1}, c_a|a) \tag{2}$$

The disadvantage due to the impact of A mediated through M is given by:

$$IE_{a_0,a_1}(y|a) = P(y_{a_0}, m_{a_1}, c_a|a) - P(y_{a_0}, m_{a_0}, c_a|a) \tag{3}$$

while the advantage is given by

$$IE_{a_1,a_0}(y|a) = P(y_{a_1}, m_{a_0}, c_a|a) - P(y_{a_1}, m_{a_1}, c_a|a) \tag{4}$$

Lastly, spurious effect if given by:

$$SE_{a_0,a_1}(y) = P(y_{a_0}|a_1) - P(y|a_0) \tag{5}$$

The corresponding empirical formulations for (1) to (5) are as follows:

$$DE_{a_0,a_1}(y|a) = \sum_{c,m}(\mathbb{E}(y|a_1, m, c) - \mathbb{E}(y|a_0, m, c))P(m|a_0, c)P(c|a) \tag{6}$$

$$DE_{a_1,a_0}(y|a) = \sum_{c,m}(\mathbb{E}(y|a_0, m, c) - \mathbb{E}(y|a_1, m, c))P(m|a_1, c)P(c|a) \tag{7}$$

$$IE_{a_0,a_1}(y|a) = \sum_{c,m}\mathbb{E}(y|a_0, m, c)(P(m|a_1, c) - P(m|a_0, c))P(c|a) \tag{8}$$

$$IE_{a_1,a_0}(y|a) = \sum_{c,m}\mathbb{E}(y|a_1, m, c)(P(m|a_0, c) - P(m|a_1, c))P(c|a) \tag{9}$$

$$SE_{a_0,a_1}(y|a) = \sum_{c,m}\mathbb{E}(y|a_0, m, c)P(m|a_0, c)(P(c|a_1) - P(c|a_0)) \tag{10}$$

$$\text{Mean Difference}_{a_0,a_1}(y) = \text{DE}_a^{\text{sym}}(y|a) + \text{IE}_a^{\text{sym}}(y|a) + SE_{a_0,a_1}(y|a) \tag{11}$$

A.2 Counterfactual Equalized Odds: Empirical Formulations

$$ER^d_{a_0,a_1}(\hat{y} \mid a, y) = \sum_c \left(P(\hat{y}_{a_1,c}) - P(\hat{y}_{a_0,c})\right) P(c \mid a, y) \tag{12}$$

$$ER^s_{a_0,a_1}(\hat{y} \mid y) = \sum_c P(\hat{y}_{a_1,c}) \left(P(c \mid a_1, y) - P(c \mid a_0, y)\right) \tag{13}$$

Using these two counterfactual error metrics, equalized odds can be broken down into direct and spurious components:

$$ER_{a_0,a_1}(\hat{y} \mid y) = ER^d_{a_0,a_1}(\hat{y} \mid a_0, y) - ER^s_{a_1,a_0}(\hat{y} \mid y) \tag{14}$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract states that a novel causal fairness package has been introduced and the paper elaborates on that.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation section and the conclusion discuss the limitations of causal fairness metrics

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper implements algorithms whose theoretical grantees have been proven in their parent papers.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the algorithms and datasets required to reproduce the same.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The package was cleared by IP review very close to the submission date. If accepted, the paper will be updated with a link to the public package repo and relevant replication scripts for this paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 2 and 3 elaborate on this

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Because the implemented methods didn't include statistical tests in them.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the lack of specialized hardware for using this package is mentioned

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Reviewed and complied with NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The limitations mention the ill effects of using causal fairness when inappropriate.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Artifacts with high risk for misuse are not part of this publication.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All authors have been notified and mentioned in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The package was cleared by IP review very close to the submission date. If accepted, the paper will be updated with a link to the public package repo and relevant replication scripts for this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not have crowdsourcing experiments and research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not have study participants

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.