# Causal Conceptions of Fairness and their Consequences

**Hamed Nilforoshan** [1] **Johann Gaebler** [1] **Ravi Shroff** [2] **Sharad Goel** [3]

## Abstract

Recent work highlights the role of causality in designing equitable decision-making algorithms. It is not immediately clear, however, how competing causal conceptions of fairness relate to one another, nor what the consequences are of using these definitions as design principles. Here we first assemble and categorize several existing causal definitions of algorithmic fairness into two broad families: (1) those that constrain the effects of decisions on counterfactual disparities; and (2) those that constrain the effects of protected characteristics—like race and gender—on decisions. We then show, analytically and empirically, that both families of definitions typically result in (strongly) Pareto dominated decision policies, meaning there is an alternative, unconstrained policy favored by every stakeholder with preferences in a large class. For example, in the case of algorithms to aid college admissions decisions, policies constrained to satisfy causal fairness definitions can simultaneously yield less student-body diversity and lower graduation rates compared to policies that explicitly optimize for desired outcomes. Indeed, under some definitions of causal fairness, we prove the resulting policies require admitting all students with the same probability, regardless of academic qualifications or group membership.

## 1. Introduction

Imagine designing an algorithm to guide decisions for college admissions. To help ensure algorithms such as this are broadly equitable, a plethora of formal fairness criteria have been proposed in the machine learning community (Barocas et al., 2017; Berk et al., 2021; Chouldechova, 2017; Chouldechova & Roth, 2020; Cleary, 1968; Corbett-Davies et al., 2017; Darlington, 1971; Dwork et al., 2012; Hardt et al., 2016; Kleinberg et al., 2016; Woodworth et al., 2017; Zafar et al., 2017a;b). For example, under the principle that fair algorithms should have comparable performance across demographic groups (Hardt et al., 2016), one might check that among applicants who were ultimately academically "successful" (e.g., who eventually earned a college degree, either at the institution in question or elsewhere), the algorithm would recommend admission for an equal proportion of candidates across race groups. Alternatively, following the principle that decisions should be agnostic to legally protected attributes like race and gender (Dwork et al., 2012), one might mandate that these features not be provided to the algorithm.

Recent scholarship has argued for extending equitable algorithm design by adopting a causal perspective, leading to myriad additional formal criteria for fairness (Chiappa, 2019; Coston et al., 2020; Imai & Jiang, 2020; Imai et al., 2020; Kilbertus et al., 2017; Kusner et al., 2017; Loftus et al., 2018; Mhasawade & Chunara, 2021; Nabi & Shpitser, 2018; Wang et al., 2019; Wu et al., 2019; Zhang & Bareinboim, 2018; Zhang et al., 2016). Here we synthesize and critically examine the statistical properties of popular causal approaches to fairness.

We begin, in Section 2, by proposing a two-part taxonomy for causal conceptions of fairness that mirrors the illustrative, non-causal fairness principles described above. Our first category of definitions encompasses those that consider the effect of decisions on counterfactual disparities. For example, recognizing the causal effect of college admission on later success, one might demand that among applicants who would be academically successful *if admitted* to a particular college, the algorithm would recommend admission for an equal proportion of candidates across race groups. The second category of definitions encompasses those that seek to limit both the direct and indirect effects of one's group membership on decisions. For example, because one's race might impact earlier educational opportunities, and hence test scores, one might require that admissions decisions are robust to the effect of race along such causal paths.

We show, in Section 3, that when the distribution of causal effects is known (or can be estimated), one can efficiently

[1] Stanford University, Stanford, CA [2] New York University, New York, NY [3] Harvard University, Cambridge, MA. Correspondence to: Hamed Nilforoshan <hamedn@cs.stanford.edu>, Johann Gaebler <jgaeb@stanford.edu>, Ravi Shroff <ravi.shroff@nyu.edu>, Sharad Goel <sgoel@hks.harvard.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2022. Copyright 2022 by the author(s).

compute utility-maximizing decision policies constrained to satisfy each of the causal fairness criteria we consider. However, for natural families of utility functions—for example, those that prefer both higher college graduation rates and more student-body diversity—we prove in Section 4 that causal fairness constraints typically lead to strongly Pareto dominated decision policies. In particular, in our running college admissions example, adhering to any of the common conceptions of causal fairness would simultaneously result in lower graduation rates and lower student-body diversity, relative to what one could achieve by explicitly tailoring admissions policies to achieve desired outcomes. In fact, under some definitions of causal fairness, we show that the induced policies require simply admitting all applicants with equal probability, irrespective of one's academic qualifications or group membership. These results, we hope, elucidate the structure—and limitations—of current causal approaches to equitable decision making.

## 2. Causal approaches to fair decision making

We describe two broad classes of causal notions of fairness: (1) those that consider outcomes when *decisions* are counterfactually altered; and (2) those that consider outcomes when *protected attributes* are counterfactually altered. We illustrate these definitions in the context of a running example of college admissions decisions.

### 2.1. Problem setup

Consider a population of individuals with observed covariates $X$, drawn i.i.d from a set $\mathcal{X} \subseteq \mathbb{R}^n$ with distribution $\mathcal{D}_X$. Further suppose that $A \in \mathcal{A}$ describes one or more discrete protected attributes, such as race or gender, which can be derived from $X$ (i.e., $A = a(X)$ for some measurable function $a$). Each individual is subject to a binary decision $D \in \{0, 1\}$, determined by a (randomized) rule $d(x) \in [0, 1]$, where $d(x) = \Pr(D = 1 \mid X = x)$ is the probability of receiving a positive decision. Given a budget $b \leq 1$, we require the decision rule to satisfy $\mathbb{E}[D] \leq b$, limiting the expected proportion of positive decisions.

In our running example, we imagine a population of applicants to some university, where $d$ denotes an admissions rule and $D$ indicates the binary admissions decision. In our setting, the covariates $X$ consist of an applicant's test score, interview performance, and race $A \in \{a_0, a_1\}$, where, for simplicity, we consider only two race groups. The budget $b < 1$ bounds the expected proportion of admitted applicants.

Assuming there is no interference between units (Imbens & Rubin, 2015), we write $Y(1)$ and $Y(0)$ for real-valued potential outcomes of interest under each of the two possible binary decisions, where $Y = Y(D)$ is the realized outcome. In our admissions example, $Y$ indicates graduation, with $Y(1)$ and $Y(0)$ describing, respectively, whether an applicant would graduate if admitted to or if rejected from that university.[1]

Given this setup, our goal is to construct decision policies $d$ that are broadly equitable, formalized in part by the causal notions of fairness described below. We assume that decisions are made algorithmically, using historical data of applicants, the admissions committee decisions, and subsequent outcomes.

### 2.2. Limiting the effect of decisions on disparities

A popular class of non-causal fairness definitions requires that error rates (e.g., false positive rates and false negative rates) are equal across protected groups (Corbett-Davies & Goel, 2018; Hardt et al., 2016). Causal analogues of these definitions have recently been proposed (Coston et al., 2020; Imai & Jiang, 2020; Imai et al., 2020), which require various conditional independence conditions to hold between the potential outcomes, protected attributes, and decisions.[2] Below we list three representative examples of this class of fairness definitions.

**Definition 1.** *Counterfactual predictive parity* holds when

$$Y(1) \perp\!\!\!\perp A \mid D = 0. \tag{1}$$

In our college admissions example, counterfactual predictive parity means that among rejected applicants, the proportion who would have graduated, had they been accepted, is equal across race groups.

**Definition 2.** *Counterfactual equalized odds* holds when

$$D \perp\!\!\!\perp A \mid Y(1). \tag{2}$$

In our running example, counterfactual equalized odds is satisfied when two conditions hold: (1) among applicants who would graduate if admitted (i.e., $Y(1) = 1$), students are admitted at the same rate across race groups; and (2) among applicants who would not graduate if admitted (i.e., $Y(1) = 0$), students are again admitted at the same rate across race groups.

**Definition 3.** *Conditional principal fairness* holds when

$$D \perp\!\!\!\perp A \mid Y(0), Y(1), W, \tag{3}$$

[1] Note that $Y(0)$ is not necessarily zero, as a rejected applicant may attend—and graduate from—a different university.

[2] In the literature on causal fairness, there is at times ambiguity between "predictions" $\hat{Y} \in \{0, 1\}$ of $Y$ and "decisions" $D \in \{0, 1\}$. Following past work (e.g., Corbett-Davies et al., 2017; Kusner et al., 2017; Wang et al., 2019), here we focus exclusively on decisions, with predictions implicitly impacting decisions but not explicitly appearing in our definitions.

where $W = w(X)$ describes a reduced set of the covariates $X$. When $W$ is constant (or, equivalently, when we do not condition on $W$), this condition is called *principal fairness*.

In our example, conditional principal fairness means that "similar" applicants—where similarity is defined by the potential outcomes and covariates $W$—are admitted at the same rate across race groups.

### 2.3. Limiting the effect of attributes on decisions

An alternative causal framework for understanding fairness considers the effects of protected attributes on decisions (Kilbertus et al., 2017; Kusner et al., 2017; Mhasawade & Chunara, 2021; Nabi & Shpitser, 2018; Wang et al., 2019; Wu et al., 2019; Zhang & Bareinboim, 2018; Zhang et al., 2016). This approach, which can be understood as codifying the legal notion of disparate treatment (Goel et al., 2017; Zafar et al., 2017a), considers a decision rule to be fair if, at a high level, decisions for individuals are the same in "(a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group" (Kusner et al., 2017).[3]

In contrast to "fairness through unawareness"—in which race and other protected attributes are barred from being an explicit input to a decision rule (cf. Barocas et al., 2017; Corbett-Davies & Goel, 2018; Dwork et al., 2012)—the causal versions of this idea consider both the direct and indirect effects of protected attributes on decisions. For example, even if decisions only directly depend on test scores, race may indirectly impact decisions through its effects on educational opportunities. This idea can be formalized by requiring that decisions remain the same in expectation even if one's protected characteristics are counterfactually altered.

**Definition 4.** *Counterfactual fairness* holds when

$$\mathbb{E}[D(a') \mid X] = \mathbb{E}[D \mid X]. \tag{4}$$

where $D(a')$ denotes the decision when one's protected attributes are counterfactually altered to be $a' \in \mathcal{A}$.

In our running example, this means that for each group of observationally identical applicants (i.e., those with the same values of $X$, meaning identical race, test scores, and interview quality), the proportion of students who are actually admitted is the same as the proportion who would be admitted if their race were counterfactually altered.

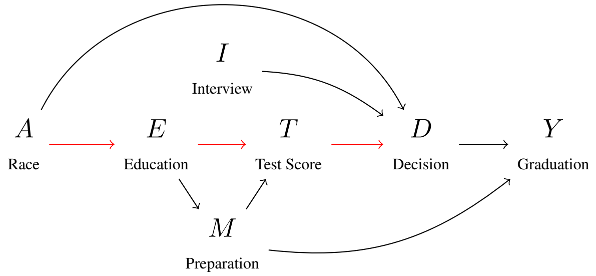

*Figure 1.* A causal DAG illustrating a hypothetical process for college admissions. Under path-specific fairness, one may require, for example, that race does not affect decisions along the path highlighted in red.

Counterfactual fairness aims to limit all direct and indirect effects of protected traits on decisions. In a generalization of this criterion—termed path-specific fairness (Chiappa, 2019; Nabi & Shpitser, 2018; Wu et al., 2019; Zhang et al., 2016)—one allows protected traits to influence decisions along certain causal paths but not others. For example, one may wish to allow the direct consideration of race by an admissions committee to implement an affirmative action policy, while also guarding against any indirect influence of race on admissions decisions that may stem from cultural biases in standardized tests (Williams, 1983).

The formal definition of path-specific fairness requires specifying a causal DAG describing relationships between attributes (both observed covariates and latent variables), decisions, and outcomes. In our running example of college admissions, we imagine that each individual's observed covariates are the result of the process illustrated by the causal DAG in Figure 1. In this graph, an applicant's race $A$ influences the educational opportunities $E$ available to them prior to college; and educational opportunities in turn influence an applicant's level of college preparation, $M$, as well as their score on a standardized admissions test, $T$, such as the SAT. We assume the admissions committee only observes an applicant's race, test scores, and interview quality, so that $X = (A, T, I)$, and makes their decision $D$ based on these attributes. Finally, whether or not an admitted student subsequently graduates (from any college), $Y$, is a function of both their preparation and whether they were admitted.[4]

[3] Conceptualizing a general causal effect of an immutable characteristic such as race or gender is rife with challenges, the greatest of which is expressed by the mantra, "no causation without manipulation" (Holland, 1986). In particular, analyzing race as a causal treatment requires one to specify what exactly is meant by "changing an individual's race" from, for example, white to Black (Gaebler et al., 2020). Such difficulties can sometimes be addressed by considering a change in the *perception* of race by a decision maker (Greiner & Rubin, 2011)—for instance, by changing the name listed on an employment application (Bertrand & Mullainathan, 2004), or by masking an individual's appearance (Chohlas-Wood et al., 2021b; Goldin & Rouse, 2000; Grogger & Ridgeway, 2006; Pierson et al., 2020).

[4] In practice, the racial composition of an admitted class may itself influence graduation rates, if, for example, diversity provides a net benefit to students (Page, 2007). Here, for simplicity, we avoid consideration of such peer effects.

To define path-specific fairness, we start by defining, for the decision $D$, path-specific counterfactuals, a general concept in causal DAGs (cf. Pearl, 2001). Suppose $\mathcal{G} = (\mathcal{V}, \mathcal{U}, \mathcal{F})$ is a causal model with nodes $\mathcal{V}$, exogenous variables $\mathcal{U}$, and structural equations $\mathcal{F}$ that define the value at each node $V_j$ as a function of its parents $\wp(V_j)$ and its associated exogenous variable $U_j$. (See, for example, Pearl (2009) for further details on causal DAGs.) Let $V_1, \ldots, V_m$ be a topological ordering of the nodes, meaning that $\wp(V_j) \subseteq \{V_1, \ldots, V_{j-1}\}$ (i.e., the parents of each node appear in the ordering before the node itself). Let $\Pi$ denote a collection of paths from node $A$ to $D$. Now, for two possible values $a$ and $a'$ for the variable $A$, the path-specific counterfactuals $D_{\Pi,a,a'}$ for the decision $D$ are generated by traversing the list of nodes in topological order, propagating counterfactual values obtained by setting $A = a'$ along paths in $\Pi$, and otherwise propagating values obtained by setting $A = a$.

Algorithm 1 describes the formal construction of path-specific counterfactuals, $Z_{\Pi,a,a'}$, for an arbitrary variable $Z$ (or collection of variables) in the DAG. To generate a sample $Z^*_{\Pi,a,a'}$ from the distribution of $Z_{\Pi,a,a'}$, we first sample values $U^*_j$ for the exogenous variables. Then, in the first loop, we traverse the DAG in topological order, setting $A$ to $a$ and iteratively computing values $V^*_j$ of the other nodes based on the structural equations in the usual fashion. In the second loop, we set $A$ to $a'$, and then iteratively compute values $\overline{V_j}^*$ for each node. $\overline{V_j}^*$ is computed using the structural equation at that node, with value $\overline{V_\ell}^*$ for each of its parents that are connected to it along a path in $\Pi$, and the value $V^*_\ell$ for all its other parents. Finally, we set $Z^*_{\Pi,a,a'}$ to $\overline{Z}^*$.

To see this definition in action, we work out an illustrative example, computing path-specific counterfactuals for the decision $D$ along the single path $\Pi = \{A \to E \to T \to D\}$ linking race to the admissions committee's decision through test preparation, highlighted in red in Figure 1. The quantity $D^*_{\Pi,a,a'} = \overline{D}^*$ is calculated below, where the first column corresponds to the first loop in Algorithm 1 and the second column corresponds to the second loop (for notational simplicity, we suppress the exogenous variables $U^*_j$ in the structural equations):

$$
\begin{aligned}
A^* &= a, & \overline{A}^* &= a', \\
E^* &= f_E(A^*), & \overline{E}^* &= f_E(\overline{A}^*), \\
M^* &= f_M(E^*), & \overline{M}^* &= f_M(E^*), \\
T^* &= f_T(E^*, M^*), & \overline{T}^* &= f_T(\overline{E}^*, M^*), \\
I^* &= I, & \overline{I}^* &= I, \\
D^* &= f_D(A^*, T^*, I^*), & \overline{D}^* &= f_D(A^*, \overline{T}^*, I^*).
\end{aligned}
$$

In particular, the value for the test score, $\overline{T}^*$, is computed using the value of $\overline{E}^*$ (since the edge $E \to T$ is on the specified path) and the value of $M^*$ (since the edge $M \to T$ is not on the path).

**Algorithm 1:** Path-specific counterfactuals

```
Data: $\mathcal{G}$ (topologically ordered), $\Pi$, $a$, and $a'$
Result: A sample $Z^*_{\Pi,a,a'}$ from $Z_{\Pi,a,a'}$

Sample values $\{U^*_j\}$ for the exogenous variables
/* Compute counterfactuals by
   setting $A$ to $a$                              */
for $j = 1, \ldots, m$ do
    if $V_j = A$ then
        $V^*_j \leftarrow a$
    else
        $\wp(V_j)^* \leftarrow \{V^*_\ell \mid V_\ell \in \wp(V_j)\}$
        $V^*_j \leftarrow f_{V_j}(\wp(V_j)^*, U^*_j)$
    end
end
/* Compute counterfactuals by
   setting $A$ to $a'$ and propagating
   values along paths in $\Pi$                     */
for $j = 1, \ldots, m$ do
    if $V_j = A$ then
        $\overline{V}^*_j \leftarrow a'$
    else
        for $V_k \in \wp(V_j)$ do
            if edge $(V_k, V_j)$ lies on a path in $\Pi$ then
                $V^\dagger_k \leftarrow \overline{V}^*_k$
            else
                $V^\dagger_k \leftarrow V^*_k$
            end
        end
    end
    $\wp(V_j)^\dagger \leftarrow \{V^\dagger_\ell \mid V_\ell \in \wp(V_j)\}$
    $\overline{V}^*_j \leftarrow f_{V_j}(\wp(V_j)^\dagger, U^*_j)$
end
$Z^*_{\Pi,a,a'} \leftarrow \overline{Z}^*$
```

Path-specific fairness formalizes the intuition that the influence of a sensitive attribute on a downstream decision may, in some circumstances, be considered legitimate (i.e., it may be acceptable for the attribute to affect decisions along certain paths in the DAG). For instance, an admissions committee may believe that the effect of race $A$ on admissions decisions $D$ which passes through college preparation $M$ is legitimate, whereas the effect of race along the path $A \to E \to T \to D$, which may reflect access to test prep or cultural biases of the tests, rather than actual academic preparedness, is illegitimate. In that case, the admissions committee may seek to ensure that the proportion of applicants they admit from a certain race group remains unchanged if one were to counterfactually alter the race of those individuals along the path $\Pi = \{A \to E \to T \to D\}$.

**Definition 5.** Let $\Pi$ be a collection of paths, and let $W = w(X)$ describe a reduced set of the covariates $X$. *Path-specific fairness*, also called $\Pi$-*fairness*, holds when, for any $a' \in \mathcal{A}$,

$$\mathbb{E}[D_{\Pi,A,a'} \mid W] = \mathbb{E}[D \mid W]. \tag{5}$$

In the definition above, the $A$ in $D_{\Pi,A,a'}$ denotes an individual's actual (non-counterfactually altered) group membership (e.g., their actual race). If $\Pi$ is the set of all paths from $A$ to $D$, then $D_{\Pi,A,a'} = D(a')$, in which case, for $W = X$, path-specific fairness is the same as counterfactual fairness.

## 3. Constructing causally fair policies

The definitions of causal fairness above constrain the set of decision policies one might adopt, but, in general, they do not yield a unique policy. For instance, a policy in which applicants are admitted randomly and independently with probability $b$—where $b$ is the specified budget—satisfies counterfactual equalized odds (Def. 2), conditional principal fairness (Def. 3), counterfactual fairness (Def. 4), and path-specific fairness (Def. 5)[5]. However, such a randomized policy may be sub-optimal in the eyes of a decision-makers aiming to maximize outcomes such as a diverse class or high graduation rates. Past work has described multiple approaches to selecting a single policy from among those satisfying any given fairness definition, including maximizing concordance of the decision with the outcome variable (Chiappa, 2019; Nabi & Shpitser, 2018) or with an existing policy (Wang et al., 2019) (e.g., in terms of binary accuracy or KL-divergence). Here, as we are primarily interested in the downstream consequences of various causal fairness definitions, we consider causally fair policies that maximize social welfare (Cai et al., 2020; Corbett-Davies et al., 2017; Kasy & Abebe, 2021; Liu et al., 2018).

Suppose $u(x)$ denotes the utility of assigning a positive decision to individuals with observed covariate values $x$, relative to assigning them negative decisions. In our running example, we set

$$u(x) = \mathbb{E}[Y(1) \mid X = x] + \lambda \cdot \mathbb{1}(a(x) = a_1), \tag{6}$$

where $\mathbb{E}[Y(1) \mid X = x]$ denotes the likelihood the applicant would graduate if admitted, $\mathbb{1}(a(x) = a_1)$ indicates whether the applicant identifies as belonging to race group $a_1$ (e.g., $a_1$ may denote a group historically underrepresented in higher education), and $\lambda \geq 0$ is an arbitrary constant that balances preferences for both high graduation rates and racial diversity.

We seek decision policies that maximize expected utility, subject to satisfying a given definition of causal fairness, as well as the budget constraint. Specifically, letting $\mathcal{C}$ denote the family of all decision policies that satisfy one of the causal fairness definitions listed above, a utility-maximizing policy $d^*$ is given by

$$\begin{aligned} d^* \in \arg\max_{d \in \mathcal{C}} \quad & \mathbb{E}[d(X)u(X)] \\ \text{s.t.} \quad & \mathbb{E}[d(X)] \leq b. \end{aligned} \tag{7}$$

Constructing optimal policies poses both statistical and computational challenges. One must, in general, estimate the joint distribution of covariates and potential outcomes—and, even more dauntingly, causal effects along designated paths for path-specific definitions of fairness. In some settings, it may be possible to obtain these estimates from observational analyses of historical data or randomized controlled trials, though both approaches typically involve substantial hurdles in practice.

If, however, one has this statistical information, Theorem 1 shows how to efficiently compute causally fair utility-maximizing policies by solving either a single linear program or a series of linear programs. In the case of counterfactual equalized odds, conditional principal fairness, counterfactual fairness, and path specific fairness, we show that the definitions induce linear constraints. For counterfactual predictive parity, the defining independence condition yields a quadratic constraint, which can be expressed as a linear constraint by further conditioning on one of the decision variables, and in turn solved through a series of linear programs. The proof of the theorem, and explicit construction of the LPs, is in the Appendix.

**Theorem 1.** Consider the optimization problem in Eq. (7).

1. If $\mathcal{C}$ is the class of policies that satisfies counterfactual equalized odds or conditional principal fairness, and the distribution of $(X, Y(0), Y(1))$ is known and supported on a finite set of size $n$, then a utility-maximizing policy constrained to lie in $\mathcal{C}$ can be constructed via a linear program with $O(n)$ variables and constraints.

2. If $\mathcal{C}$ is the class of policies that satisfies path-specific fairness (including counterfactual fairness), and the distribution of $(X, D_{\Pi,A,a})$ is known and supported on a finite set of size $n$, then a utility-maximizing policy constrained to lie in $\mathcal{C}$ can be constructed via a linear program with $O(n)$ variables and constraints.

3. Suppose $\mathcal{C}$ is the class of policies that satisfies counterfactual predictive parity, that the distribution of

[5] A policy satisfying counterfactual predictive parity (Def. 1) is not guaranteed to exist. For example, if $b = 0$—in which case $D = 0$ a.s.—and $\mathbb{E}[Y(1) \mid A = a_1] \neq \mathbb{E}[Y(1) \mid A = a_2]$, then Eq. (1) cannot hold. Similar counterexamples can be constructed for $b \ll 1$.

$(X, Y(1))$ is known and supported on a finite set of size $n$, and that the optimization problem in Eq. (7) has a feasible solution. Further suppose $Y(1)$ is supported on $m$ points, and let $\Delta^{m-1} = \{p \in \mathbb{R}^m \mid p_i \geq 0 \text{ and } \sum_{i=1}^m p_i = 1\}$ be the unit $(m-1)$-simplex. Then one can construct a set of linear programs $\mathcal{L} = \{L(v)\}_{v \in \Delta^m}$, with each having $O(n)$ variables and constraints, such that the solution to one of the LPs in $\mathcal{L}$ is a utility-maximizing policy constrained to lie in $\mathcal{C}$.

## 4. The structure of causally fair policies

Above, for each definition of causal fairness, we showed how to construct utility-maximizing policies that satisfy the corresponding constraints. Now we explore the structural properties of causally fair policies. We show—both empirically and analytically, under relatively mild distributional assumptions—that policies constrained to be causally fair are disfavored by every individual in a natural class of decision makers with varying preferences for diversity. To formalize these results, we start by introducing some notation and then defining the concept of (strong) Pareto dominance.

For a real-valued utility function $u$ and decision policy $d$, we write $u(d) = \mathbb{E}[d(X)u(X)]$ to denote the utility of $d$ under $u$.

**Definition 6.** For a budget $b$, we say a decision policy $d$ is *feasible* if $\mathbb{E}[d(X)] \leq b$.

Given a collection of utility functions encoding the preferences of different individuals, we say a decision policy $d$ is Pareto dominated if there exists a feasible alternative $d'$ such that none of the decision makers prefers $d$ over $d'$, and at least one decision maker strictly prefers $d'$ over $d$, a property formalized in Definition 7.

**Definition 7.** Suppose $\mathcal{U}$ is a collection of utility functions. A decision policy $d$ is *Pareto dominated* if there exists a feasible alternative $d'$ such that $u(d') \geq u(d)$ for all $u \in \mathcal{U}$, and there exists $u' \in \mathcal{U}$ such that $u'(d') > u(d')$. A policy $d$ is *strongly Pareto dominated* if there exists a feasible alternative $d'$ such that $u(d') > u(d)$ for all $u \in \mathcal{U}$. A policy $d$ is *Pareto efficient* if it is feasible and not Pareto dominated, and the *Pareto frontier* is the set of Pareto efficient policies.

To develop intuition about the structure of causally fair decision policies, we continue working through our illustrative example of college admissions. We consider a collection of decision makers with utilities $\mathcal{U}$ of the form

$$u(x) = \mathbb{E}[Y(1) \mid X = x] + \lambda \cdot \mathbb{1}(a(x) = a_1), \tag{8}$$

for $\lambda \geq 0$, as in Eq. (6). In this example, decision makers differ in their preferences for diversity (as determined by $\lambda$), but otherwise have similar preferences. We call such a collection of utilities *consistent modulo* $a$.

**Definition 8.** We say that a set of utilities $\mathcal{U}$ is *consistent modulo* $a$ if, for any $u, u' \in \mathcal{U}$:

1. For any $x$, $\text{sign}(u(x)) = \text{sign}(u'(x))$;
2. For any $x_1$ and $x_2$ such that $a(x_1) = a(x_2)$, $u(x_1) > u(x_2)$ if and only if $u'(x_1) > u'(x_2)$.

For consistent utilities, the Pareto frontier takes a particularly simple form, represented by (a subset of) group-specific threshold policies.

**Proposition 1.** Suppose $\mathcal{U}$ is a set of utilities this is consistent modulo $a$, and $d$ is a Pareto efficient decision policy. Then there exists a group-specific threshold policy $d'$ with $u(d) = u(d')$ for all $u \in \mathcal{U}$. Specifically, there exist group-specific constants $t_{a_i}$ and $p_{a_i}$, and a utility function $u' \in \mathcal{U}$ such that $d'$ takes the following form:

$$d'(x) = \begin{cases} 1 & \text{if } u'(x) > t_{a(x)}, \\ p_{a(x)} & \text{if } u'(x) = t_{a(x)}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

In Proposition 1, the threshold policy $d'$ is constructed with respect to a specific utility function $u' \in \mathcal{U}$. We note, however, that because we assume $\mathcal{U}$ is consistent modulo $a$, $d'$ is in fact a threshold rule with respect to every $u \in \mathcal{U}$, where only the specific values for the thresholds change depending on $u$.

With these preliminaries in place, we now empirically explore the structure of causally fair decision policies in the context of our stylized example of college admissions, given by the causal DAG in Figure 1. We specifically consider an example in which the exogenous variables $\mathcal{U} = \{u_R, u_E, u_M, u_T, u_I\}$ in the DAG are independently distributed as follows:

$$U_R, U_D \sim \text{UNIF}(0, 1), \quad U_E, U_M, U_T, U_I \sim \text{N}(0, \sigma^2).$$

For fixed constants $\mu_R$, $\beta_{E,0}$, $\beta_{E,R}$, $\beta_{M,0}$, $\beta_{M,E}$, $\beta_{T,0}$, $\beta_{T,E}$, $\beta_{T,M}$, $\beta_{T,B}$, and $\mu_I$, we define the endogenous variables $\mathcal{V} = \{R, E, M, T, I, D, Y\}$ in the DAG by the fol-

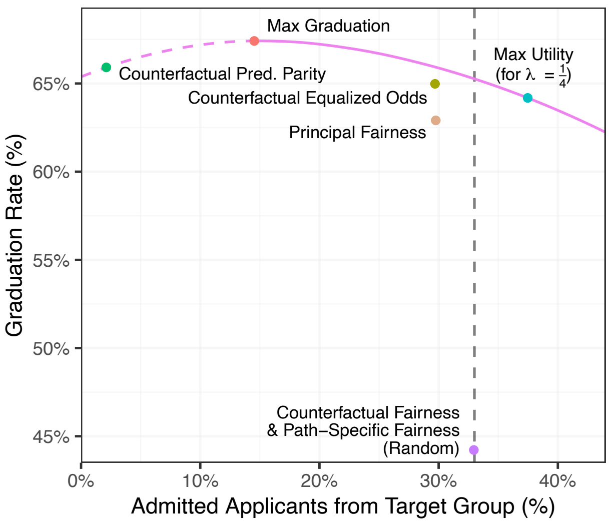

*Figure 2.* Utility maximizing policies for various definitions of causal fairness in an illustrative example of college admissions, with the Pareto frontier depicted by the solid purple curve. For path-specific fairness, we set $\Pi$ equal to the single path $A \to E \to T \to D$, and set $W = X$. In all cases, policies restricted to satisfy any of our above definitions of causal fairness are strongly Pareto dominated. In particular, in each case, there are alternative policies that simultaneously result in greater student-body diversity and higher graduation rates than under the restricted policies.

lowing structural equations:[6]

$$
\begin{aligned}
f_R(u_R) &= \begin{cases} a_1 & \text{if } u_R < \mu_R \\ a_0 & \text{otherwise} \end{cases}, \\
f_E(r, u_E) &= \beta_{E,0} + \beta_{E,R} \cdot \mathbb{1}(r = w) + u_E, \\
f_M(e, u_M) &= \beta_{M,0} + \beta_{M,E} \cdot e + u_M, \\
f_T(e, m, u_T) &= \beta_{T,0} + \beta_{T,E} \cdot e \\
&\qquad + \beta_{T,M} \cdot m + \beta_{T,B} \cdot e \cdot m + u_T, \\
f_I(u_i) &= \mathbb{1}(u_I < \mu_I), \\
f_Y(m, d) &= m \cdot d.
\end{aligned}
$$

In this hypothetical example, applicants in the target race group $a_1$ have, on average, fewer educational opportunities than those applicants in group $a_0$, which leads to lower average academic preparedness, as well as lower average test scores. The quality of the interview is the same in distribution across race groups.

Now, for the family of utilities defined by Eq. (8), we apply Theorem 1 to compute utility-maximizing policies for each of the above causal definitions of fairness. We plot the results in Figure 2, where, for each policy, the horizontal axis shows the expected proportion of admitted applicants from the target race group, and the vertical axis shows the expected graduation rate. The Pareto frontier is depicted by the solid purple curve, computed using Proposition 1.[7] (For reference, the dashed purple line corresponds to max-utility policies constrained to satisfy lower levels of student-body diversity, though those policies are not on the Pareto frontier.)

In every case, policies restricted to satisfy one of the above definitions of causal fairness are strongly Pareto dominated, meaning that there is an alternative feasible policy favored by all decision makers with preferences in $\mathcal{U}$. In particular, for each definition of causal fairness, there is an alternative feasible policy in which one simultaneously achieves more student-body diversity and higher graduate rates. In some instances, the efficiency gap is quite stark. Utility-maximizing policies constrained to satisfy either counterfactual fairness or path-specific fairness require one to admit each applicant independently with fixed probability $b$ (where $b$ is the budget), regardless of academic preparedness or group membership.[8] These results show that constraining decision-making algorithms to satisfy popular definitions of causal fairness can have unintended consequences, and may even harm the very groups they were ostensibly designed to protect.

The patterns illustrated in Figure 2 and discussed above are not idiosyncratic to our particular example, but rather hold quite generally. Indeed, Theorem 2 shows that for *almost every* joint distribution on $(X, Y(0), Y(1))$, any decision policy satisfying counterfactual equalized odds or conditional principal fairness is strongly Pareto dominated. Similarly, for almost every joint distribution on $(X, X_{\Pi,A,a})$, we show that policies satisfying path-specific fairness (including counterfactual fairness) are strongly Pareto dominated. (NB: The analogous statement for counterfactual predictive parity is not true, which we address in Theorem 3.)

The notion of *almost every* distribution that we use here was formalized by Christensen (1972), Hunt et al. (1992), Anderson et al. (2001), and others (cf. Ott & Yorke, 2005, for a review). Suppose, for a moment, that combinations of covariates and outcomes take values in a finite set of

[6] In our example, we use constants $\mu_R = 1/3$, $\beta_{E,0} = 1$, $\beta_{E,R} = -1$, $\beta_{M,0} = -1$, $\beta_{M,E} = 1$, $\beta_{T,0} = 50$, $\beta_{T,E} = 4$, $\beta_{T,M} = 4$, $\beta_{T,B} = 1$. We do not supply the structural equation for the decision policy, $f_D(r, t, i, u_D)$, since that depends on which notion of causal fairness—if any—one chooses to adopt.

[7] To trace out the purple curve in Figure 2, it is sufficient to sweep over group-specific threshold policies. To see this, recall that for any policy $d$ on the Pareto frontier, Proposition 1 shows that there exists a group-specific threshold policy $d'$ with $u(d) = u(d')$ for all $u \in \mathcal{U}$. Let $u_\lambda$ denote the utility function with the corresponding value of $\lambda$ in Eq. (8). Now, since $u_0(d)$ equals graduation rate, the graduation rate under $d$ and $d'$ are identical. Similarly, since $\lim_{\lambda \to \infty} u_\lambda(d)/\lambda$ is the proportion of admitted applicants who are Black, that proportion is also identical under $d$ and $d'$.

[8] For path-specific fairness, we set $\Pi$ equal to the single path $A \to E \to T \to D$, and set $W = X$ in this example.

size $m$. Then the space of joint distributions on covariates and outcomes can be represented by the unit $(m-1)$-simplex: $\Delta^{m-1} = \{p \in \mathbb{R}^m \mid p_i \geq 0 \text{ and } \sum_{i=1}^m p_i = 1\}$. Since $\Delta^{m-1}$ is an $(m-1)$-dimensional hyperplane in $\mathbb{R}^m$, it inherits the usual Lebesgue measure on $\mathbb{R}^{m-1}$. In this finite-dimensional setting, *almost every* distribution means a subset of distributions that has full Lebesgue measure on the simplex. Given a property that holds for almost every distribution in this sense, that property holds almost surely under any probability distribution on the space of distributions that is described by a density on the simplex. We use a generalization of this basic idea that extends to infinite-dimensional spaces, allowing us to consider distributions with arbitrary support. (See the Appendix for further details.)

**Theorem 2.** Suppose $\mathcal{U}$ is a set of utilities consistent modulo $a$. Then for almost every joint distribution of $X$, $Y(0)$, and $Y(1)$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$, any decision policy satisfying counterfactual equalized odds or conditional principal fairness is strongly Pareto dominated.

Likewise, for almost every joint distribution of $X$ and its path-specific counterfactuals $X_{\Pi,A,a}$ on $\mathcal{X} \times \mathcal{X}^{\mathcal{A}}$, any decision policy satisfying path-specific fairness—including the special case of counterfactual fairness—is strongly Pareto dominated.

Conditional predictive parity differs from the other causal fairness definitions in that can be one multiple threshold policy that satisfies it—though no more (see Proposition **??** in Appendix A). However, this policy can nevertheless fail to be Pareto efficient, since it can effectively encode a preference for the baseline group.

**Theorem 3.** Suppose that there exists a utility function $u$ such that the relevant (counterfactual) conditional distributions of $U = u(X)$ are logit-normal, i.e.,

$$\mathcal{D}(U \mid A = a) = \text{logit}^{-1}(\mathcal{N}(\mu_a, \sigma_a^2)).$$

Let $d$ be a feasible, utility-maximizing policy constrained to satisfy counterfactual predictive parity, and let $d^*$ be a feasible unconstrained utility-maximizing policy. If $\mathbb{E}[U \mid A = a] < \mathbb{E}[U]$, then $u(d) < u(d^*)$, and

$$\mathbb{E}[d(X) \mid A = a] < \mathbb{E}[d^*(X) \mid A = a].$$

Path-specifically fair policies typically impose much more restrictive constraints than the other causal fairness definitions. In particular, if it is possible for individuals with identical observables to have different observables counterfactually (i.e., if $Pr(X_{\Pi,A,a} = x' \mid X = x) < 1$), then it can easily happen that the only path-specifically fair policies are fully randomized. Fully (or even partially) randomized policies can incur substantial penalties compared to policies which can use covariate information to allocate resources more efficiently. For instance, an admissions policy under which one admits students uniformly at random through a lottery can reasonably be expected to result in a substantially lower graduation rate than a policy under which informative covariates, such as test score, are used to admit applicants.

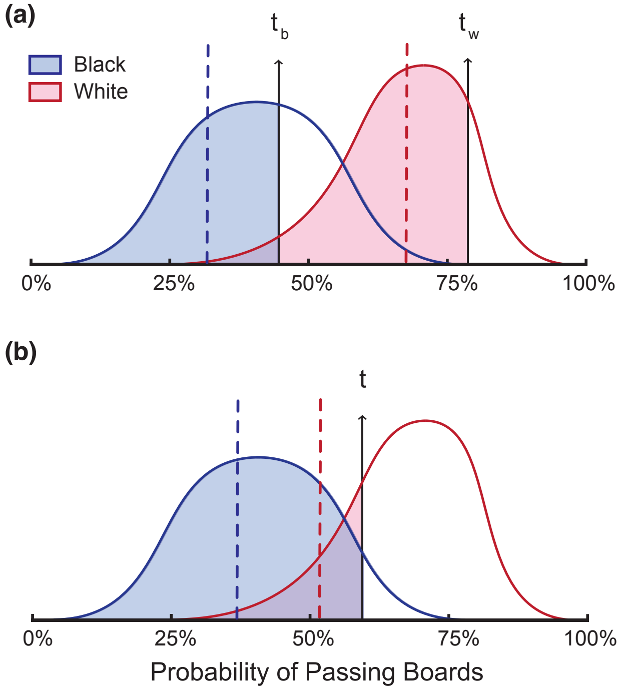

*Figure 3.* **An illustration of the problem of infra-marginality.** We show a hypothetical distribution of boards-passage probabilities. When outcome distributions differ between groups prior to making decisions, post-decision infra-marginal statistics (e.g. the false omission rate) will also differ regardless of whether a **(a)** double or **(b)** single threshold is applied. Solid vertical lines indicate test score thresholds policies; dashed vertical lines indicate mean boards passage rate among rejected applicants (i.e. false omission rate).

**Theorem 4.** Suppose $w(X) = X$ and $\Pr(X_{\Pi,A,a} = x' \mid X = x) > 0$ for all $x, x'$. Then, any $\Pi$-fair policy is fully randomized, i.e., $d(x) = b$ or $d(x) = 0$ for all $x \in \mathcal{X}$.

Theorem 4, is a consequence of the fact that $\Pi$-fairness, in effect, requires one to treat an actual individual (e.g., with covariates $X = x$) and any of their counterfactual dopplegänger's (e.g., with covariates $X_{\Pi,A,a} = x'$ the same. However, it likewise requires one to treat any *other* actual individual (e.g., with covariates $X = x''$) with a similar dopplegänger (e.g., with $X_{\Pi,A,a} = x'$) the same. Therefore, if any two individuals have a positive probability of having the same counterfactual dopplegänger, then *everyone* must be treated identically by the decisionmaker, i.e., the decisonmaker must use a lottery to make their decision. (Note that, in fact, the hypotheses of Theorem 4 can be weakened; see Appendix A for a generalization.)

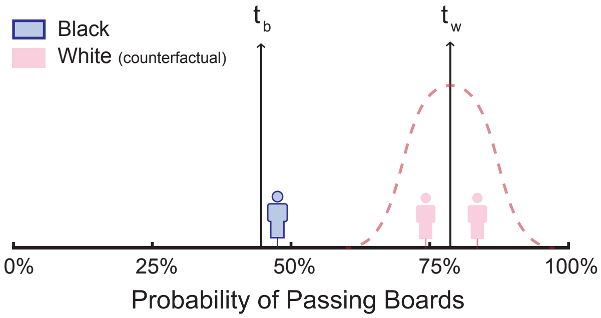

*Figure 4.* **An illustration of the problem of counterfactual non-determinism.** When an individual in the real world is above the admissions threshold, but their distribution of counterfactual outcomes falls above and below the threshold, policies which require eliminating the causal effect of race necessitate randomizing the decision in the real world to mirror the uncertainty of the counterfactual world.

## 5. Discussion

We have worked to collect, synthesize, and investigate several causal conceptions of fairness that recently have appeared in the machine learning literature. These definitions formalize intuitively desirable properties—for example, minimizing the direct and indirect effects of race on decisions. But, as we have shown both analytically and with a synthetic example, they can, perhaps surprisingly, lead to policies with unintended downstream outcomes. For instance, in our running example of college admissions, enforcing various causal fairness definitions can lead to a student body that is both less academically prepared and less diverse than what one could achieve under natural alternative policies, potentially harming the very groups these definitions were ostensibly designed to protect. Our results thus highlight a gap between the goals and potential consequences of popular causal approaches to fairness.

What, then, is the role of causal reasoning in designing equitable algorithms? Under a consequentialist perspective to algorithm design (Cai et al., 2020; Chohlas-Wood et al., 2021a), one aims to construct polices with the most desirable expected outcomes, a task that inherently demands causal reasoning. Formally, this approach corresponds to solving the unconstrained optimization problem in Eq. (7), where preferences for diversity may be directly encoded in the utility function itself, rather than by constraining the class of policies, mitigating potentially problematic consequences. While conceptually appealing, this consequentialist approach still faces considerable practical challenges, including estimating the expected effects of decisions, and eliciting preferences over outcomes.

Our analysis demonstrates some of the limitations of mathematical formalizations of fairness, and reinforces the need to explicitly consider the consequences of actions, particularly when decisions are automated and carried out at scale. Looking forward, we hope our work clarifies the ways in which causal reasoning can aid the equitable design of algorithms.

## Acknowledgements

We thank Guillaume Basse, Jennifer Hill, and Ravi Sojitra for helpful conversations. H.N. was supported by an NSF Graduate Fellowship and Stanford Knight Hennessy Scholarship. J.G was supported by a Stanford Knight Hennessy Scholarship.

## References

Anderson, R. M., Zame, W. R., et al. Genericity with infinitely many parameters. *Advances in Theoretical Economics*, 1(1):1–62, 2001.

Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Bertrand, M. and Mullainathan, S. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

Cai, W., Gaebler, J., Garg, N., and Goel, S. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 22–28, 2020.

Chiappa, S. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Chohlas-Wood, A., Coots, M., Brunskill, E., and Goel, S. Learning to be fair: A consequentialist approach to equitable decision-making. *arXiv preprint arXiv:2109.08792*, 2021a.

Chohlas-Wood, A., Nudell, J., Yao, K., Lin, Z., Nyarko, J., and Goel, S. Blind justice: Algorithmically masking race in charging decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 35–45, 2021b.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Chouldechova, A. and Roth, A. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

Christensen, J. P. R. On sets of haar measure zero in abelian polish groups. *Israel Journal of Mathematics*, 13(3-4): 255–260, 1972.

Cleary, T. A. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124, 1968.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 582–593, 2020.

Darlington, R. B. Another look at “cultural fairness” 1. *Journal of educational measurement*, 8(2):71–82, 1971.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Gaebler, J., Cai, W., Basse, G., Shroff, R., Goel, S., and Hill, J. A causal framework for observational studies of discrimination. *arXiv preprint arXiv:2006.12460*, 2020.

Goel, S., Perelman, M., Shroff, R., and Sklansky, D. A. Combatting police discrimination in the age of big data. *New Criminal Law Review: An International and Interdisciplinary Journal*, 20(2):181–232, 2017.

Goldin, C. and Rouse, C. Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, 90(4):715–741, 2000.

Greiner, D. J. and Rubin, D. B. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785, 2011.

Grogger, J. and Ridgeway, G. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475):878–887, 2006.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Hunt, B. R., Sauer, T., and Yorke, J. A. Prevalence: a translation-invariant “almost every” on infinite-dimensional spaces. *Bulletin of the American mathematical society*, 27(2):217–238, 1992.

Imai, K. and Jiang, Z. Principal fairness for human and algorithmic decision-making. *arXiv preprint arXiv:2005.10400*, 2020.

Imai, K., Jiang, Z., Greiner, J., Halen, R., and Shin, S. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *arXiv preprint arXiv:2012.02845*, 2020.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Kasy, M. and Abebe, R. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 576–586, 2021.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

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

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Mhasawade, V. and Chunara, R. Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 784–794, 2021.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ott, W. and Yorke, J. Prevalence. *Bulletin of the American Mathematical Society*, 42(3):263–290, 2005.

Page, S. E. Making the difference: Applying a logic of diversity. *Academy of Management Perspectives*, 21(4): 6–20, 2007.

Pearl, J. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, 2001*, pp. 411–420. Morgan Kaufman, 2001.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., et al. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, 4(7):736–745, 2020.

Wang, Y., Sridhar, D., and Blei, D. M. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*, 2019.

Williams, T. S. Some issues in the standardized testing of minority students. *Journal of Education*, pp. 192–208, 1983.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pp. 1920–1953. PMLR, 2017.

Wu, Y., Zhang, L., Wu, X., and Tong, H. Pc-fairness: A unified framework for measuring causality-based fairness. *arXiv preprint arXiv:1910.12586*, 2019.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P., and Weller, A. From parity to preference-based notions of fairness in classification. *arXiv preprint arXiv:1707.00010*, 2017b.

Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Zhang, L., Wu, Y., and Wu, X. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

# A. Proofs

*Proof of Theorem 1.* Define decision variables $d_x \in [0, 1]$ corresponding to each possible value of $X$, where $d_x$ denotes the probability of taking a positive action for individuals with covariate value $x$.

Letting $p(x) = \Pr(X = x)$ denote the density of $X$, note that the objective function $\sum_x d_x u(x) p(x)$ and the budget constraint $\sum_x d_x p(x) \leq b$ are both linear in the decision variables.

We now show that each of the causal fairness definitions can be enforced via linear constraints.

*Counterfactual Predictive Parity* Unlike subsequent definitions of fairness, we find the utility-maximizing solution that satisfies counterfactual predictive parity via a *series* of linear programs. We start with the definition $Y(1) \perp\!\!\!\perp A = 0$ which is true if and only if $Pr(Y(1) = y \mid A = a, D = 0) = Pr(Y(1) = y \mid D = 0)$. We represent $Pr(Y(1) = y \mid D = 0)$ as $C_y$ and re-write the LHS as:

$$
\begin{aligned}
&Pr(Y(1) = y \mid A = a, D = 0))\\
&= \frac{Pr(Y(1) = y, A = a, D = 0)}{Pr(A = a, D = 0)}\\
&= \frac{\sum_x Pr(X = x)(1 - d_x)\mathbb{1}_{a(x)=a} Pr(Y(1) = y \mid X = x)}{\sum_x Pr(X = x)\mathbb{1}_{a(x)=a}(1 - d_x)}\\
&= C_y
\end{aligned}
$$

The above results in the linear constraint $\sum_x Pr(X = x)(1 - d_x)\mathbb{1}_{a(x)=a} Pr(Y(1) = y \mid X = x) = C_y(\sum_x Pr(X = x)\mathbb{1}_{a(x)=a}(1 - d_x))$, where $C_y$ is implemented as a parameter in our linear program.

*Counterfactual Equalized Odds* For each $y \in \{0, 1\}$, compute the distribution of co-variates for members of the first group ($a_0$), conditional on $Y(1) = y$, and form the matrix $P^{a_0}$ where $P^{a_0}_{y,x} = Pr(X = x \mid Y(1) = y, A = a_0)$. Similarly, form a matrix for members of the second group $P^{a_1}$ where $P^{a_1}_{y,x} = Pr(X = x \mid Y(1) = y, A = a_1)$.

Counterfactual equalized odds constraints can thus be written in the form $P^{a_0} d = P^{a_1} d$. This constraint holds if and only if $D \perp\!\!\!\perp A \mid Y(1) \iff \mathbb{E}[D \mid A = a_0, Y(1)] = \mathbb{E}[D \mid A = a_1, Y(1)]$ if D is a binary decision, because for any value of $y$ and $a$:

$$
\begin{aligned}
&\mathbb{E}[D \mid A = a, Y(1) = y] = \mathbb{E}[d(X) \mid A = a, Y(1) = y]\\
&= \sum_{x \in \mathcal{X}} (\mathbb{E}[d(x) \mid X = x, Y(1) = y, A = a]\\
&\cdot \Pr(X = x \mid Y(1) = y, A = a))\\
&= \sum_x P^a_{y,x} d_x = (P^a d)_y
\end{aligned}
$$

*Conditional Principal Fairness* For each conditional principal stratum $s = (y_0, y_1, w)$ where $y_0 \in \{0, 1\}$, $y_1 \in \{0, 1\}$, and $w \in \mathcal{W}$, compute the distribution of covariates for members of the first group ($a_0$), conditional on $Y(0) = y_0, Y(1) = y_1, W = w$, and form the matrix $P^{a_0}$ where $P^{a_0}_{s,x} = Pr(X = x \mid Y(0) = y_0, Y(1) = y_1, W = w, A = a_0)$. Similarly, form a matrix for members of the second group $P^{a_1}$ where $P^{a_1}_{s,x} = Pr(X = x \mid Y(0) = y_0, Y(1) = y_1, W = w, A = a_1)$.

Conditional principal fairness constraints can thus be written in the form $P^{a_0} d = P^{a_1} d$. This constraint holds if and only if $D \perp\!\!\!\perp A \mid Y(0), Y(1), W \iff \mathbb{E}[D \mid A = a_0, Y(0), Y(1), W] = \mathbb{E}[D \mid A = a_1, Y(0), Y(1), W]$ if D is a binary decision, because for any principal stratum $s = (y_0, y_1, w)$:

$$
\begin{aligned}
&\mathbb{E}[D \mid A = a, Y(0) = y_0, Y(1) = y_1, W = w]\\
&= \mathbb{E}[d(X) \mid A = a, Y(0) = y_0, Y(1) = y_1, W = w]\\
&= \sum_{x \in \mathcal{X}} (\mathbb{E}[d(x) \mid X = x, Y(0) = y_0, Y(1) = y_1, W = w, A = a]\\
&\cdot \Pr(X = x \mid Y(0) = y_0, Y(1) = y_1, W = w, A = a))\\
&= \sum_x P^a_{s,x} d_x = (P^a d)_s
\end{aligned}
$$

*Counterfactual Fairness* For each $x \in \mathcal{X}$, compute the distribution of counterfactual covariates conditional on $X = x$, and form the stochastic matrix $P$ where $P_{x,x'} = Pr(X(a_i) = x' \mid X = x)$, where $X(a_i)$ is $X$ after counterfactually intervening on protected attribute.

Counterfactual fairness constraints can thus be written in the form $Pd = d$. This constraint holds if and only if $\mathbb{E}[D(a_i) \mid X] = \mathbb{E}[D \mid X]$, because for any covariate value $x$: $\mathbb{E}[D \mid X = x] = d_x$, and

$$
\begin{aligned}
&\mathbb{E}[D(a_i) \mid X = x] = \mathbb{E}[d(X(a_i)) \mid X = x] =\\
&= \sum_{x' \in \mathcal{X}} \mathbb{E}[d(x') \mid X = x, X(a_i) = x'] \cdot \Pr(X(a_i) = x' \mid X = x)\\
&= \sum_{x'} P_{x,x'} d_{x'} = (Pd)_x
\end{aligned}
$$

*Path-Specific Fairness* For each $w \in \mathcal{W}$, compute the distribution of counterfactual covariates conditional on $W = w$, and form the stochastic matrix $P^1$ where $P^1_{w,x'} = Pr(X_{\Pi,A,a} = x' \mid W = w)$, where $X_{\Pi,A,a}$ is $X$ after counterfactually intervening on protected attribute along the set of unfair pathways $\Pi$. Similarly, form a matrix $P^2$ where $P^2_{w,x} = Pr(X = x \mid W = w)$.

Path-specific fairness constraints can thus be written in the form $P^1 d = P^2 d$. This constraint holds if and only if $\mathbb{E}[D_{\Pi,A,a_i} \mid W] = \mathbb{E}[D \mid W]$, because for any value of reduced covariates $w$:

$$\begin{aligned}&\mathbb{E}[D \mid W=w] = \mathbb{E}[d(X) \mid W=w] = \\ &= \sum_{x\in\mathcal{X}} \mathbb{E}[d(x) \mid W=w, X=x] \cdot \Pr(X=x \mid W=w) \\ &= \sum_{x} P^2_{w,x} d_x = (P^2 d)_w\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}[D_{\Pi,A,a_i} \mid W=w] = \mathbb{E}[d(X_{\Pi,A,a_i}) \mid W=w] = \\ &= \sum_{x'\in\mathcal{X}} \mathbb{E}[d(x') \mid W=w, X_{\Pi,A,a_i}=x'] \\ &\qquad \cdot \Pr(X_{\Pi,A,a_i}=x' \mid W=w) \\ &= \sum_{x'} P^1_{w,x'} d_{x'} = (P^1 d)_w\end{aligned}$$

□

*Proof of Theorem 2.* Let $\mathbb{K}$ denote the set of bounded total variation Borel measures on $\mathcal{K} = \mathcal{X}\times\mathcal{Y}\times\mathcal{Y}$ or, respectively, $\mathcal{K} = \mathcal{X}\times\mathcal{X}$, as appropriate. Then $\mathcal{K}$ is a Banach space under the total variation norm, and the set $\mathbb{P}$ of probability measures is trivially a convex subset. If $p, p' \in \mathbb{P}$, then: (1) $\lambda \cdot p[E] + (1-\lambda)\cdot p'[E] \geq 0$ for all Borel sets $E$ and $\lambda \in [0,1]$, since $p(E), p'(E) \geq 0$; and (2) $\lambda \cdot p(\mathcal{K}) + (1-\lambda)\cdot p'[\mathcal{K}] = 1$, since $p[\mathcal{K}] = p'[\mathcal{K}] = 1$.

Let $\mathbf{D} \subset \mathbb{P}$ denote the set of measures (i.e., distributions) such that there exists a decision policy satsifying counterfactual equalized odds (resp., conditional principal fairness or path-specific fairness) that is Pareto efficient. By Theorem 4.6 in (Ott & Yorke, 2005), it suffices to show the following:

1. The set $\mathbb{P}$ is universally measurable;
2. There exists $k \in \mathbb{N}$, a $k$-dimensional subspace $V$ of $\mathbb{K}$, and a constant $x_0 \in \mathbb{K}$ such that $(\mathbb{P} + x_0) \cap V$ has positive measure in the $k$-dimensional Lebesgue measure on $V$;
3. For all $x \in \mathbb{K}$, $(\mathbf{D} + x) \cap V$ is a null set in the $k$-dimensional probability measure on $V$.

To see the first claim, note that all Borel sets are universally measurable, and all closed sets are Borel sets. Therefore it suffices to show that $\mathbb{P}$ is closed in $\mathbb{K}$. For, assume $p_n \to p$ is a convergent sequence in $\mathbb{K}$, where $p_n \in \mathbb{P}$ for all $n \in \mathbb{N}$. Then, (1) for all Borel sets $E \subset \mathcal{K}$, $0 \leq \lim_{n\to\infty} p_n[E] = p[E]$; and (2) $1 = \lim_{n\to\infty} p_n[\mathcal{K}] = p[\mathcal{K}]$. Therefore $p \in \mathbb{P}$.

Next, to see the second claim, let $\{x_1, \ldots, x_k\}$ be any collection of distinct points in $\mathcal{K}$, and consider measure $m$ where

$$m[E] = \frac{|\{1 \leq i \leq k : x_k \in E\}|}{k}.$$

Now, let $m_i$ for $i = 1, \ldots, k$ be the measure defined by

$$m_i[E] = \mathbf{1}(x_i \in E) + \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}(x_i \in E)$$

and let $V = \textsc{Span}(m_1, \ldots, m_k)$. Then, for any $(\lambda_1, \ldots, \lambda_k) \in \Delta^k$ in the $k$-simplex, we trivially have that

$$m + \sum_{i=1}^{k} \lambda_i \cdot m_i \in \mathbb{P}.$$

Since the $k$-simplex has Lebesgue measure $\frac{1}{k!}$ in $\mathbb{R}^k$, it follows that $\lambda_V(\mathbb{P} - m) \geq \frac{1}{k!}$.

Next, we show the third claim. For simplicity, assume $\mathbf{D}$ is the set of distributions under which there is a Pareto efficient decision policy satisfying counterfactual equalized odds; the proof in the other two cases is virtually identical. Choose distinct groups $a_0, a_1 \in \mathcal{A}$ and distinct $y_0, y_1 \in \mathcal{Y}$. For each $a \in \{a_0, a_1\}$, $y \in \{y_0, y_1\}$, choose $v_{a,y,0}$ and $v_{a,y,1}$ such that the $A$-value of $v_{a,y,i}$ is $a$, the $Y(1)$ value is $y$, and the $X$-values satisfy $u(x_{a,y,0}) < u(x_{a,y,1})$ for all $u \in \mathcal{U}$. Let $V = \textsc{Span}(v_{a_0,y_0,0}, \ldots, v_{a_1,y_1,1})$. Then, as shown above, there exists $m \in \mathbb{P}$ such that $\lambda_V(\mathbf{D} - m) > 0$.

Let $p \in \mathbb{K}$ be arbitrary. □

We partition $\mathcal{X}$ so that $\mathcal{X}_a = \{x \in \mathcal{X} : a(x) = a\}$. We first prove Theorem 4 in the case when $\mathcal{X}$ is discrete to illustrate the intuition, before giving the more technical general proof.

*Proof of Theorem 4.* By Eq. (5), we have that, for all $a \in \mathcal{A}$,

$$d(x) = \sum_{x'\in\mathcal{X}_a} d(x') \cdot \Pr(X_{\Pi,A,a} = x' \mid X = x).$$

Now, let $d_{\max} = \max_{x\in\mathcal{X}} d(x)$ and let $x_{\max}$ be such that $d(x_{\max}) = d_{\max}$. Let $a_0 = a(x_{\max})$. Now, suppose for some $x \in \mathcal{X}_{a_0}$, $d(x) < d_{\max}$. Then, it follows that

$$\begin{aligned}d_{\max} &= d(x_{\max}) \\ &= \sum_{x\in\mathcal{X}_{a_0}} d(x) \cdot \Pr(X_{\Pi,A,a_0} = x \mid X = x_{\max}) \\ &< \sum_{x\in\mathcal{X}_{a_0}} d_{\max} \cdot \Pr(X_{\Pi,A,a_0} = x \mid X = x_{\max}) \\ &= d_{\max},\end{aligned}$$

where the inequality follows from the fact that $\Pr(X_{\Pi,A,a_0} = x \mid X = x_{\max}) > 0$ for all $x \in \mathcal{X}_{a_0}$ and the fact that for some $x \in \mathcal{X}_{a_0}$, $d(x) < d_{\max}$. This is a contradiction, and therefore $d(x) = d_{\max}$ for all $x \in \mathcal{X}_{a_0}$.

Now, let $x \in \mathcal{X}$ be arbitrary. It follows immediately that

$$\begin{aligned} d(x) &= \sum_{x' \in \mathcal{X}_{a_0}} d(x') \cdot \Pr(X_{\Pi,A,a_0} = x' \mid X = x) \\ &= \sum_{x' \in \mathcal{X}_{a_0}} d_{\max} \cdot \Pr(X_{\Pi,A,a_0} = x' \mid X = x) \\ &= d_{\max}. \end{aligned}$$

Therefore $d(x) = d_{\max}$ for all $x \in \mathcal{X}$. □

The following is a general analogue of Theorem 4.

**Theorem 5.** Suppose $w(X) = X$ and that

1. For all $a \in \mathcal{A}$ and $\epsilon > 0$ there exists $\delta > 0$ such that for any event $S$, if $\Pr(X \in S \mid A = a) < \delta$ then $\Pr(X_{\Pi,A,a} \in S \mid X) < \epsilon$ a.s.;
2. For all $a \in \mathcal{A}$, if $\Pr(X_{\Pi,A,a} \in S) > 0$, then $\Pr(X \in S \mid A = a) > 0$.

Then any $\Pi$-fair policy is fully randomized, i.e., $d(X) = b$ or $d(X) = 0$ a.s.

The technical hypotheses of Theorem **??** simply ensure that the conditional probability measures $\Pr(E \mid X)$ are "sufficiently" mutually non-singular distributions on $\mathcal{X}$ with respect to the distribution of $X$. (E.g., the conditional distribution of $X_{\Pi,A,a} \mid X$ cannot have atoms that $X$ itself does not have, and *vice versa*.)

*Proof.* Let $d_{\max} = \|d(x)\|_\infty$. Note that there must be some $a_0 \in \mathcal{A}$ such that $\Pr(d_{\max} - d(X) > \epsilon \mid A = a_0) > 0$ for all $\epsilon$. Assume that $\Pr(d(X) = d_{\max} \mid A = a_0) = 0$. By Markov's inequality, for any $\epsilon > 0$, a.s.,

$$\begin{aligned} \Pr(d_{\max} - d(X_{\Pi,A,a_0}) \geq \rho \mid X) &\leq \frac{\mathbb{E}[d_{\max} - d(X_{\Pi,A,a_0}) \mid X]}{\rho} \\ &= \frac{d_{\max} - d(X)}{\rho}, \end{aligned}$$

where the equality follows from Eq. (5). Rearranging, it follows that, a.s.,

$$\Pr(d_{\max} - d(X_{\Pi,A,a}) < \rho \mid X) \geq 1 - \frac{d_{\max} - d(X)}{\rho}. \tag{10}$$

Now, choose $\delta$ sufficiently small that for all $S$ such that $\Pr(X \in S \mid A = a_0) < \delta$, $\Pr(X_{\Pi,A,a_0} \in S \mid X) < \frac{1}{2}$ a.s. Now, by hypothesis, there exists some $\epsilon$ such that $\Pr(d_{\max} - d(X) < \epsilon \mid A = a_0) < \delta$. Then, we have that $\Pr(d_{\max} - d(X_{\Pi,A,a_0}) < \epsilon \mid X) < \frac{1}{2}$ a.s. However, by hypothesis, $\Pr(d_{\max} - d(X) < \frac{\epsilon}{2} \mid A = a_0) > 0$. Therefore, with positive probability, we have that

$$\begin{aligned} 1 - \frac{d_{\max} - d(X)}{\epsilon} &> 1 - \frac{\frac{\epsilon}{2}}{\epsilon} \\ &= \frac{1}{2} \\ &> \Pr(d_{\max} - d(X_{\Pi,A,a}) \mid X). \end{aligned}$$

This contradicts Eq. (10), and so $\Pr(d(X) = d_{\max} \mid A = a_0) > 0$.

Now, we show that $\Pr(d(X) = d_{\max} \mid A = a_0) = 1$. For, suppose $\Pr(d(X) < d_{\max} \mid A = a_0) > 0$. Then, by hypothesis, there exists some $\epsilon$ such that $\Pr(d(X_{\Pi,A,a}) < d_{\max} \mid X) > \epsilon$ a.s., and so $\mathbb{E}[d(X_{\Pi,A,a}) \mid X] < d_{\max}$ a.s. Then, note that

$$\begin{aligned} d_{\max} &= \mathbb{E}[d(X) \mid d(X) = d_{\max}, A = a_0] \\ &= \mathbb{E}[\mathbb{E}[d(X_{\Pi,A,a_0}) \mid X] \mid d(X) = d_{\max}, A = a_0] \\ &< \mathbb{E}[d_{\max} \mid d(X) = d_{\max}, A = a_0] \\ &= d_{\max}. \end{aligned}$$

This is a contradiction, and so $\Pr(d(X) = d_{\max} \mid A = a_0) = 1$.

It follows immediately that $\Pr(d(X_{\Pi,A,a}) = d_{\max} \mid X, A = a_0) = 1$ a.s. Therefore, it follows from the second hypothesis that $\Pr(d(X_{\Pi,a,a}) = d_{\max} \mid X) = 1$ a.s., and so $d(X) = d_{\max}$ a.s. □