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 existing 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 popular 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 *almost always*—in a measure theoretic sense—result in strongly Pareto dominated decision policies, meaning there is an alternative, unconstrained policy favored by every stakeholder with preferences drawn from a large, natural class. For example, in the case of college admissions decisions, policies constrained to satisfy causal fairness definitions would be disfavored by every stakeholder with neutral or positive preferences for both academic preparedness and diversity. Indeed, under a prominent definition of causal fairness, we prove the resulting policies require admitting all students with the same probability, regardless of academic qualifications or group membership. Our results highlight formal limitations and potential adverse consequences of common mathematical notions of causal fairness.

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 (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., 2017; 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 (cf. Corbett-Davies & Goel, 2018; 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; Galhotra et al., 2022; 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., 2017). Less attention, however, has been given to understanding the potential downstream consequences of using these causal definitions of fairness as algorithmic design principles, leaving an important gap to fill if these criteria are to responsibly inform policy choices.

Here we synthesize and critically examine the statistical properties and concomitant consequences 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

[*]Equal contribution [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>.

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 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 degree attainment and more student-body diversity—we prove in Section 4 that causal fairness constraints *almost always* lead to strongly Pareto dominated decision policies. To establish this result, we use the theory of prevalence (Anderson & Zame, 2001; Christensen, 1972; Hunt et al., 1992; Ott & Yorke, 2005), which extends the notion of full-measure sets to infinite-dimensional vector spaces. In particular, in our running college admissions example, adhering to any of the common conceptions of causal fairness would simultaneously result in a lower number of degrees attained 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 prove 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., $A = \alpha(X)$ for some measurable function α). 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.[1] Given a budget b with $0 < b < 1$, 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 a particular college, where d denotes an admissions rule and D indicates a binary admissions decision. To simplify our exposition, we assume all admitted students attend the school. In our setting, the covariates X consist of an applicant's test score and race $A \in \{a_0, a_1\}$, where, for notational convenience, we consider two race groups. The budget $b \leq 1$ 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 is a binary variable that indicates college graduation (i.e., degree attainment), with $Y(1)$ and $Y(0)$ describing, respectively, whether an applicant would attain a college degree if admitted to or if rejected from the school we consider. Note that $Y(0)$ is not necessarily zero, as a rejected applicant may attend—and graduate from—a different university.

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 focus on decisions that are made algorithmically, informed by historical data on applicants 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 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; Mishler et al., 2021), 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: counterfactual predictive parity (Coston et al., 2020), counterfactual equalized odds (Coston et al., 2020; Mishler et al., 2021), and conditional principal fairness (Imai & Jiang, 2020).[3]

[1] That is, $D = \mathbb{1}_{U_D \leq d(X)}$, where U_D is an independent uniform random variable.

[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.

[3] Our subsequent analytical results extend in a straightforward manner to structurally similar variants of these definitions (e.g., requiring $Y(0) \perp\!\!\!\perp A \mid D = 1$ or $D \perp\!\!\!\perp A \mid Y(0)$, variants of counterfactual predictive parity and counterfactual equalized odds, respectively).

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 attained a college degree, 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

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

where, for a measurable function ω on $\mathcal{X}$, $W = \omega(X)$ 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., 2017). 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).[4]

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. 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, which in turn influence test scores. This idea can be formalized by requiring that decisions remain the same in expectation even if one's protected characteristics are counterfactually altered, a condition known as counterfactual fairness (Kusner et al., 2017).

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 any $a' \in \mathcal{A}$.

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 and test score), the proportion of students who are actually admitted is the same as the proportion who would be admitted if their race were counterfactually altered.

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., 2017)—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

[4] 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., 2022; Hu & Kohler-Hausmann, 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).

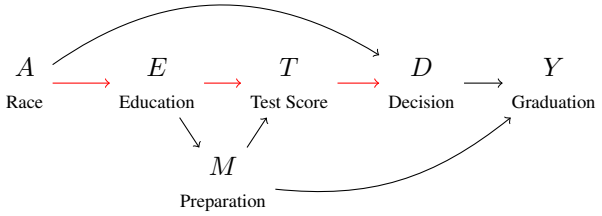

Figure 1. A causal DAG illustrating a hypothetical process for college admissions. Under path-specific fairness, one may require, for example, that race does not affect decisions along the path highlighted in red.

SAT. We assume the admissions committee only observes an applicant's race and test score so that $X = (A, T)$, 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.[5]

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 (2009a) 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 Π 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 Π, and otherwise propagating values obtained by setting $A = a$. (In Algorithm 1 in the Appendix, we formally define path-specific counterfactuals for an arbitrary node—or collection of nodes—in the DAG.)

To see this idea 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. In the system of equations below, the first column corresponds to draws V^* for each node V in the DAG, where we set A to a, and then propagate that value as usual. The second column corresponds to draws $\overline{V}^*$ of path-specific counterfactuals, where we set A to a', and then propagate the counterfactuals only along the path $A \to E \to T \to D$. 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). As a result of this process, we obtain a draw $\overline{D}^*$ from the distribution of $D_{\Pi,a,a'}$.

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

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 Π be a collection of paths, and, for a measurable function w on $\mathcal{X}$, let $W = \omega(X)$ describe a reduced set of the covariates X. *Path-specific fairness*, also called Π-*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, rather than a particular counterfactual level a, the baseline level of the path-specific effect is A, i.e., an individual's actual (non-counterfactually altered) group membership (e.g., their actual race). We have implicitly assumed that the decision variable D is a descendant of the covariates X. In particular, without loss of generality, we assume D is defined by the structural equation $f_D(x, u_D) = \mathbb{1}_{u_D \leq d(x)}$, where the exogenous variable $u_D \sim \text{UNIF}(0, 1)$, so that $\Pr(D = 1 \mid X = x) = d(x)$. If Π 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

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

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).[6] However, such a randomized policy may be sub-optimal in the eyes of decision-makers aiming to maximize outcomes such as class diversity or degree attainment. 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 utility (Cai et al., 2020; Chohlas-Wood et al., 2021a; 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

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

where $\mathbb{E}[Y(1) \mid X = x]$ denotes the likelihood the applicant would graduate if admitted, $\mathbb{1}_{\alpha(x)=a_1}$ 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 student graduation 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.

We prove that if one has this statistical information, it is possible to efficiently compute causally fair utility-maximizing policies by solving either a single linear program or a series of linear programs (Appendix, Theorem B.1). In the case of counterfactual equalized odds, conditional principal fairness, counterfactual fairness, and path-specific fairness, we show that the definitions can be translated to linear constraints. For counterfactual predictive parity, the defining independence condition yields a quadratic constraint, which we show can be expressed as a linear constraint by further conditioning on one of the decision variables, and the optimization problem in turn can be solved through a series of linear programs.

4. The structure of causally fair policies

Above, for each definition of causal fairness, we sketched 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.

4.1. Pareto dominance and consistent utilities

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.

[6] 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 $b \ll 1$.

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

Definition 8. We say that a set of utilities $\mathcal{U}$ is *consistent modulo* α 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 $\alpha(x_1) = \alpha(x_2)$, $u(x_1) > u(x_2)$ 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 that is consistent modulo α. Then any Pareto efficient decision policy d is a multiple threshold policy. That is, for any $u \in \mathcal{U}$, there exist group-specific constants $t_a \geq 0$ such that, a.s.:*

$$d(x) = \begin{cases} 1 & u(x) > t_{\alpha(x)}, \\ 0 & u(x) < t_{\alpha(x)}. \end{cases} \tag{8}$$

The proof of Proposition 1 is in the Appendix.[7]

4.2. An empirical example

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. In the hypothetical pool of 100,000 applicants we consider, 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. See Section C in the Appendix for additional details, including the specific structural equations we use.

For the utility function in Eq. (6) with $\lambda = \frac{1}{4}$, we apply Theorem B.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 number of admitted applicants from the target race group, and the vertical axis shows the expected number of college graduates. Additionally, for the family of utilities $\mathcal{U}$ given by Eq. (6) for $\lambda \geq 0$, we depict the Pareto frontier by the solid purple curve, computed via Proposition 1.[8] For reference, the dashed purple line corresponds to max-utility policies constrained to satisfy the level of diversity indicated on the x-axis, though these policies are not on the Pareto frontier, as they result in fewer college graduates and lower diversity than the policy that maximizes graduation alone (indicated by the "max graduation" point in Figure 2).

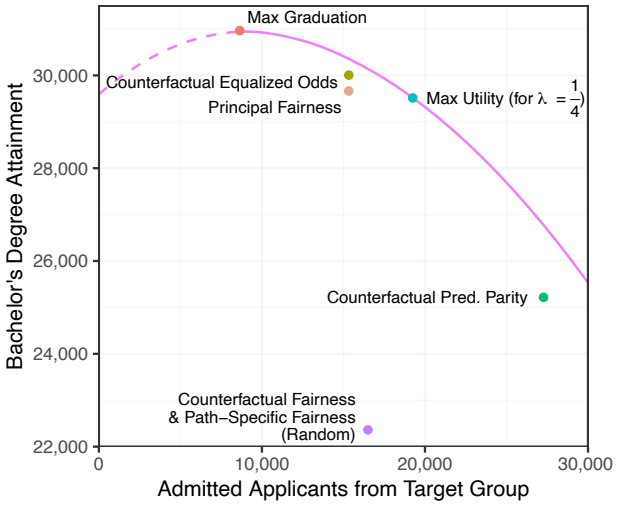

Figure 2. Utility-maximizing policies for various definitions of causal fairness in an illustrative example of college admissions, with the Pareto frontier depicted by the solid purple curve. For path-specific fairness, we set Π equal to the single path $A \to E \to T \to D$, and set $W = X$. In all 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 bachelor's degree attainment than under the restricted policies.

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 more college graduates. In

[7] In the statement of the proposition, we do not specify what happens at the thresholds $u(x) = t_{\alpha(x)}$ themselves, as one can typically ignore the exact manner in which decisions are made at the threshold. Specifically, given a threshold policy d, we can construct a standardized threshold policy d' that is constant within group at the threshold (i.e., $d'(x) = c_{\alpha(x)}$ when $u(x) = t_{\alpha(x)}$), and for which: (1) $\mathbb{E}[d'(X)|A] = \mathbb{E}[d(X)|A]$; and (2) $u(d') = u(d)$. In our running example, this means we can standardize threshold policies so that applicants at the threshold are admitted with the same group-specific probability.

[8] For all the cases we consider, the optimal policies admit the maximum proportion of students allowed under the budget b (i.e., $\Pr(D = 1) = b$). To compute the Pareto frontier in Figure 2, it is sufficient—by Proposition 1 and Footnote 7—to sweep over (standardized) group-specific threshold policies relative to the utility $u_0(x) = \mathbb{E}[Y(1)|X = x]$.

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.[9] 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.

4.3. The statistical structure of causally fair policies

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

The notion of *almost every* distribution that we use here was formalized by Christensen (1972), Hunt et al. (1992), Anderson & Zame (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 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 Δ^{m-1} is a subset of 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.)

To prove this result, we make relatively mild restrictions on the set of distributions and utilities we consider to exclude degenerate cases, as formalized by Definition 9 below.

Definition 9. Let $\mathcal{G}$ be a collection of functions from $\mathcal{Z}$ to $\mathbb{R}^d$ for some set $\mathcal{Z}$. We say that a distribution of Z on $\mathcal{Z}$ is *$\mathcal{G}$-fine* if $g(Z)$ has a density for all $g \in \mathcal{G}$.

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

In particular, $\mathcal{U}$-fineness ensures that the distribution of $u(X)$ has a density. In the absence of $\mathcal{U}$-fineness, corner cases can arise in which an especially large number of policies may be Pareto efficient, in particular when $u(X)$ has large atoms and X can be used to predict the potential outcomes $Y(0)$ and $Y(1)$ even after conditioning on $u(X)$. See Prop. E.7 for details. Our example of college admissions, where $\mathcal{U}$ is defined by Eq. (6), is $\mathcal{U}$-fine.

Theorem 1. *Suppose $\mathcal{U}$ is a set of utilities consistent modulo α, and that there exists a $\mathcal{U}$-fine distribution of X such that $\Pr(u(X) > 0, A = a) > 0$ for all $u \in \mathcal{U}$ and $a \in \mathcal{A}$, where $A = \alpha(X)$. Then,*

- *For almost every $\mathcal{U}$-fine distribution of X and $Y(1)$, any decision policy satisfying counterfactual equalized odds is strongly Pareto dominated.*
- *If $|\textsc{Img}(\omega)| < \infty$ and there exists a $\mathcal{U}$-fine distribution of X such that $\Pr(A = a, W = w) > 0$ for all $a \in \mathcal{A}$ and $w \in \textsc{Img}(\omega)$, where $W = \omega(X)$, then, for almost every $\mathcal{U}$-fine joint distribution of X, $Y(0)$, and $Y(1)$, any decision policy satisfying conditional principal fairness is strongly Pareto dominated.*
- *If $|\textsc{Img}(\omega)| < \infty$ and that there exists a $\mathcal{U}$-fine distribution of X such that $\Pr(A = a, W = w_i) > 0$ for all $a \in \mathcal{A}$ and some distinct $w_0, w_1 \in \textsc{Img}(\omega)$, then, for almost every $\mathcal{U}^{\mathcal{A}}$-fine joint distributions of A and the counterfactuals $X_{\Pi,A,a'}$, any decision policy satisfying path-specific fairness is strongly Pareto dominated.*[10]

The proof of Theorem 1 is given in the Appendix. At a high-level, the causal definitions considered in the theorem have too many constraints to lie on the Pareto frontier. In the finite-dimensional case, imagine starting with a randomly selected joint distribution of X and $Y(1)$. Counterfactual equalized odds constrains decision policies for each value of $y \in \mathcal{Y}$, requiring $D \perp\!\!\!\perp A \mid Y(1) = y$. For any specific y, there is typically a unique decision policy on the frontier that satisfies that accompanying constraint—which, by Proposition 1, is a group-specific threshold rule. However, since the decision policy is unique, there are no more degrees of freedom to adjust it, and so one must simply hope that the other constraints happen to be satisfied by the given distribution. Theorem 1 makes precise the idea that such a coincidental occurrence is a measure zero event. This phenomenon is similar to the problem of infra-marginality (Ayres, 2002; Simoiu et al., 2017), which likewise afflicts non-causal notions of fairness (Corbett-Davies & Goel, 2018; Corbett-Davies et al., 2017).

[10] Here, $u^{\mathcal{A}} : (x_a)_{a \in \mathcal{A}} \mapsto (u(x_a))_{a \in \mathcal{A}}$ and $\mathcal{U}^{\mathcal{A}}$ is the set of $u^{\mathcal{A}}$ for $u \in \mathcal{U}$. In other words, the requirement is that the joint distribution of the $u(X_{\Pi,A,a})$ has a density.

In some common settings, path-specific fairness with $W = X$ constrains decisions so severely that the only allowable policies are constant (i.e., $d(x_1) = d(x_2)$ for all $x_1, x_2 \in \mathcal{X}$). For instance, in our running example, path-specific fairness requires admitting all applicants with the same probability, irrespective of academic preparation or group membership. Thus, all applicants are admitted with probability b, where b is the budget, under the optimal policy constrained to satisfy path-specific fairness.

To build intuition for this result, we sketch the argument for a finite covariate space $\mathcal{X}$. Given a policy d that satisfies path-specific fairness, select $x^* \in \arg\max_{x \in \mathcal{X}} d(x)$. By the definition of path-specific fairness, for any $a \in \mathcal{A}$,

$$\begin{aligned} d(x^*) &= \mathbb{E}[D_{\Pi,A,a} \mid X = x^*] \\ &= \sum_{x \in \alpha^{-1}(a)} d(x) \cdot \Pr(X_{\Pi,A,a} = x \mid X = x^*). \end{aligned} \tag{9}$$

Now suppose there exists an $a' \in \mathcal{A}$ such that $\Pr(X_{\Pi,A,a'} = x \mid X = x^*) > 0$ for all $x \in \alpha^{-1}(a')$. In this case, because $d(x) \leq d(x^*)$ for all $x \in \mathcal{X}$, Eq. (9) shows that in fact $d(x) = d(x^*)$ for all $x \in \alpha^{-1}(a')$. Now, let x' be arbitrary. Again, by the definition of path-specific fairness, we have that

$$\begin{aligned} d(x') &= \mathbb{E}[D_{\Pi,A,a'} \mid X = x'] \\ &= \sum_{x \in \alpha^{-1}(a')} d(x) \cdot \Pr(X_{\Pi,A,a'} = x \mid X = x') \\ &= \sum_{x \in \alpha^{-1}(a')} d(x^*) \cdot \Pr(X_{\Pi,A,a'} = x \mid X = x^*) \\ &= d(x^*), \end{aligned}$$

where we use in the third equality the fact $d(x) = d(x^*)$ for all $x \in \alpha^{-1}(a')$, and in the final equality the fact that $X_{\Pi,A,a'}$ is supported on $\alpha^{-1}(a')$.

Theorem 2 formalizes and extends this argument to more general settings, where $\Pr(X_{\Pi,A,a'} = x \mid X = x^*)$ is not necessarily positive for all $x \in \alpha^{-1}(a')$. The proof of Theorem 2 is in the Appendix, along with extensions to continuous covariate spaces and a more complete characterization of Π-fair policies for finite $\mathcal{X}$.

Theorem 2. *Suppose $\mathcal{X}$ is finite and $\Pr(X = x) > 0$ for all $x \in \mathcal{X}$. Suppose $Z = \zeta(X)$ is a random variable such that:*

1. *$Z = Z_{\Pi,A,a'}$ for all $a' \in \mathcal{A}$,*
2. *$\Pr(X_{\Pi,A,a'} = x' \mid X = x) > 0$ for all $a' \in \mathcal{A}$ such that $\alpha(x) \neq a'$ and $x, x' \in \mathcal{X}$ such that $\zeta(x) = \zeta(x')$.*

Then, for any Π-fair policy d, with $W = X$, there exists a function f such that $d(X) = f(Z)$, i.e., d is constant across individuals having the same value of Z.

The first condition of Theorem 2 holds for any reduced set of covariates Z that is not causally affected by changes in A (e.g., Z is not a descendent of A). The second condition requires that among individuals with covariates x, a positive fraction have covariates x' in a counterfactual world in which they belonged to another group a'. Because $\zeta(x)$ is the same in the real and counterfactual worlds—since Z is unaffected by A, by the first condition—we only consider x' such that $\zeta(x') = \zeta(x)$ in the second condition.

In our running example, the only non-race covariate is test score, which is downstream of race. Further, among students with a given test score, a positive fraction achieve any other test score in the counterfactual world in which their race is altered. As such, the empty set of reduced covariates—formally encoded by setting ζ to a constant function—satisfies the conditions of Theorem 2. The theorem then implies that under any Π-fair policy, every applicant is admitted with equal probability.

Even when decisions are not perfectly uniform lotteries, as in our admissions example, Theorem 2 suggests that enforcing Π-fairness can lead to unexpected outcomes. For instance, suppose we modify our admissions example to additionally include age as a covariate that is causally unconnected to race—as some past work has done. In that case, Π-fair policies would admit students based on their age alone, irrespective of test score or race. Although in some cases such restrictive policies might be desirable, this strong structural constraint implied by Π-fairness appears to be a largely unintended consequence of the mathematical formalism.

The conditions of Theorem 2 are relatively mild, but do not hold in every setting. Suppose that in our admissions example it were the case that $T_{\Pi,A,a_0} = T_{\Pi,A,a_1} + c$ for some constant c—that is, suppose the effect of intervening on race is a constant change to an applicant's test score. Then the second condition of Theorem 2 would no longer hold for a constant ζ. Indeed, any multiple-threshold policy in which $t_{a_0} = t_{a_1} + c$ would be Π-fair. In practice, though, such deterministic counterfactuals would seem to be the exception rather than the rule. For example, it seems reasonable to expect that test scores would depend on race in complex ways that induce considerable heterogeneity.

Lastly, we note that $W \neq X$ in some variants of path-specific fairness (e.g., Nabi & Shpitser, 2018; Zhang & Bareinboim, 2018), in which case Theorem 2 does not apply. Although, in that case, policies are typically still Pareto dominated in accordance with Theorem 1.

We conclude our analysis by investigating counterfactual predictive parity, the least demanding of the causal notions of fairness we have considered, requiring only that $Y(1) \perp\!\!\!\perp A \mid D = 0$. As such, it is in general possible to have a policy

on the Pareto frontier that satisfies this condition. However, in Proposition 2, we show that this cannot happen in some common cases, including our example of college admissions. In that setting, when the target group has lower average graduation rates—a pattern that often motivates efforts to actively increase diversity—decision policies constrained to satisfy counterfactual predictive parity are Pareto dominated. The proof of the proposition is in the Appendix.

Proposition 2. *Suppose $\mathcal{A} = \{a_0, a_1\}$, and consider the family $\mathcal{U}$ of utility functions of the form*

$$u(x) = r(x) + \lambda \cdot \mathbb{1}_{\alpha(x)=a_1},$$

indexed by $\lambda \geq 0$, where $r(x) = \mathbb{E}[Y(1) \mid X = x]$. Suppose the conditional distributions of $r(X)$ given A are beta distributed, i.e.,

$$r(X) \mid A = a \sim \textsc{Beta}(\mu_a, v),$$

with $v > 2$ and $\mu_{a_0} > \mu_{a_1} > 1/v$.[11] Then any policy satisfying counterfactual predictive parity is strongly Pareto dominated.

In Proposition 2, we restrict to a family of beta distributions where $E[r(X)]$ is lower in the target group compared to the non-target group, a condition which, as discussed above, often holds in settings where one seeks to prioritize diversity.

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. In contrast to prior impossibility results (Chouldechova, 2017; Kleinberg et al., 2017), in which different formal notions of fairness are shown to be in conflict with each other, we demonstrate trade-offs between formal notions of fairness and resulting social welfare. 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.

[11] Here we parameterize the beta distribution in terms of its mean μ and sample size v. In terms of the common, alternative α-β parameterization, $\mu = \alpha/(\alpha + \beta)$ and $v = \alpha + \beta$.

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; Liang et al., 2021), one aims to construct policies 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 illustrates some of the limitations of mathematical formalizations of fairness, reinforcing 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 a Stanford Knight-Hennessy Scholarship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656518. J.G was supported by a Stanford Knight-Hennessy Scholarship. R.S. was supported by the NSF Program on Fairness in AI in Collaboration with Amazon under the award "FAI: End-to-End Fairness for Algorithm-in-the-Loop Decision Making in the Public Sector," no. IIS-2040898. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

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

Ayres, I. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.

Benji. The sum of an uncountable number of positive numbers. Mathematics Stack Exchange, 2020. URL https://math.stackexchange.com/q/20661. (version: 2020-05-29).

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.

Billingsley, P. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. ISBN 0-471-00710-2. A Wiley-Interscience Publication.

Brozius, H. Conditional expectation - $E[f(X,Y)|Y]$. Mathematics Stack Exchange, 2019. URL https://math.stackexchange.com/q/3247577. (Version: 2019-06-01).

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". *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. *Statistics and Public Policy*, 2022.

Galhotra, S., Shanmugam, K., Sattigeri, P., and Varshney, K. R. Causal feature selection for algorithmic fairness. *Proceedings of the 2022 International Conference on Management of Data (SIGMOD)*, 2022.

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.

Hu, L. and Kohler-Hausmann, I. What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*, 2020.

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.

Kemeny, J. G. and Snell, J. L. *Finite Markov chains*. Undergraduate Texts in Mathematics. Springer-Verlag, New York-Heidelberg, 1976. Reprinting of the 1960 original.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 656–666, 2017.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Kusner, M., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4069–4079, 2017.

Liang, A., Lu, J., and Mu, X. Algorithmic design: Fairness versus accuracy. *arXiv preprint arXiv:2112.09975*, 2021.

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.

Mishler, A., Kennedy, E. H., and Chouldechova, A. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 386–400, 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, 2009a.

Pearl, J. *Causality*. Cambridge University Press, 2 edition, 2009b. doi: 10.1017/CBO9780511803161.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., and Goel, S. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4(7):736–745, 2020.

Rao, M. M. *Conditional measures and applications*, volume 271 of *Pure and Applied Mathematics (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2005. ISBN 978-1-57444-593-0; 1-57444-593-6. doi: 10.1201/9781420027433. URL https://doi-org.stanford.idm.oclc.org/10.1201/9781420027433.

Rudin, W. *Real and Complex Analysis*. McGraw-Hill Book Co., New York, third edition, 1987. ISBN 0-07-054234-1.

Rudin, W. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition, 1991. ISBN 0-07-054236-8.

Silva, C. E. *Invitation to ergodic theory*, volume 42 of *Student Mathematical Library*. American Mathematical Society, Providence, RI, 2008. ISBN 978-0-8218-4420-5. doi: 10.1090/stml/042. URL https://doi-org.stanford.idm.oclc.org/10.1090/stml/042.

Simoiu, C., Corbett-Davies, S., and Goel, S. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

Steele, R. Space of vector measures equipped with the total variation norm is complete. Mathematics Stack Exchange, 2019. URL https://math.stackexchange.com/q/3197508. (Version: 2019-04-22).

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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 228–238, 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. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3929–3935, 2017.

A. Path-specific counterfactuals

Constructing policies which satisfy path-specific fairness requires computing path-specific counterfactual values of features. In Algorithm 1, we describe 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 Π, and the value V^*_ℓ for all its other parents. Finally, we set $Z^*_{\Pi,a,a'}$ to $\overline{Z}^*$.

Algorithm 1: Path-specific counterfactuals

Data: $\mathcal{G}$ (topologically ordered), Π, 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 Π */
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* Π **then**
        $V^\dagger_k \leftarrow \overline{V}^*_k$
      else
        $V^\dagger_k \leftarrow V^*_k$
      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
end
$Z^*_{\Pi,a,a'} \leftarrow \overline{Z}^*$

B. Constructing causally fair policies

In order to construct causally fair policies, we prove that the optimization problem in Eq. (7) can be efficiently solved as a single linear program—in the case of counterfactual equalized odds, conditional principal fairness, counterfactual fairness, and path-specific fairness—or as a series of linear programs in the case of counterfactual predictive parity.

Theorem B.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 $(X, Y(1))$ 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 k points, and let $\Delta^{k-1} = \{p \in \mathbb{R}^k \mid p_i \geq 0 \text{ and } \sum_{i=1}^k p_i = 1\}$ be the unit $(k-1)$-simplex. Then one can construct a set of linear programs $\mathcal{L} = \{L(v)\}_{v \in \Delta^k}$, 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 $\mathcal{C}$.*

Proof. Let $\mathcal{X} = \{x_1, \ldots, x_m\}$; then, we seek decision variables d_i, $i = 1, \ldots, m$, corresponding to the probability of making a positive decision for individuals with covariate value x_i. Therefore, we require that $0 \leq d_i \leq 1$.

Letting $p_i = \Pr(X = x_i)$ denote the mass of X at x_i, note that the objective function $\mathbb{E}[d(X) \cdot u(X)]$ equals $\sum_{i=1}^m d_i \cdot u(x_i) \cdot p_i$ and the budget constraint $\sum_{i=1}^m d_i \cdot p_i \leq b$ are both linear in the decision variables.

We now show that each of the causal fairness definitions can be enforced via linear constraints. We do so in three parts as listed in theorem.

Theorem B.1 Part 1 First, we consider counterfactual equalized odds. A decision policy satisfies counterfactual equalized odds when $D \perp\!\!\!\perp A \mid Y(1)$. Since D is binary, this condition is equivalent to the expression $\mathbb{E}[d(X) \mid A = a, Y(1) = y] = \mathbb{E}[d(X) \mid Y(1) = y]$ for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$ such that $\Pr(Y(1) = y) > 0$.

Expanding this expression and replacing $d(x_j)$ by the corresponding decision variable d_j, we obtain that

$$\sum_{i=1}^m d_i \cdot \Pr(X = x_i \mid A = a, Y(1) = y) = \sum_{i=1}^m d_i \cdot \Pr(X = x_i \mid Y(1) = y)$$

for each $a \in \mathcal{A}$ and each of the finitely many values $y \in \mathcal{Y}$ such that $\Pr(Y(1) = y) > 0$. These constraints are linear in the d_i by inspection.

Next, we consider conditional principal fairness. A decision policy satisfies conditional principal fairness when $D \perp\!\!\!\perp A \mid Y(0), Y(1), W$, where $W = \omega(X)$ denotes a reduced set of the covariates X. Again, since D is binary, this condition is equivalent to the expression $\mathbb{E}[d(X) \mid A = a, Y(0) = y_0, Y(1) = y_1, W = w] = \mathbb{E}[d(X) \mid Y(0) = y_0, Y(1) = y_1, W = w]$ for all y_0, y_1, and w satisfying $\Pr(Y(0) = y_0, Y(1) = y_1, W = w) > 0$. As above, expanding this expression and replacing $d(x_j)$ by the corresponding decision variable d_j yields linear constraints of the form

$$\sum_{i=1}^m d_i \cdot \Pr(X = x_i \mid A = a, S = s) = \sum_{j=1}^m d_i \cdot \Pr(X = x_i \mid S = s)$$

for each $a \in \mathcal{A}$ and each of the finitely many values of $S = (Y(0), Y(1), W)$ such that $s = (y_0, y_1, w) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{W}$ satisfies $\Pr(Y(0) = y_0, Y(1) = y_1, W = w) > 0$. Again, these constraints are linear by inspection.

Theorem B.1 Part 2 Suppose a decision policy satisfies path-specific fairness for a given collection of paths Π and a (possibly) reduced set of covariates $W = \omega(X)$, meaning that for every $a' \in \mathcal{A}$, $\mathbb{E}[D_{\Pi,A,a'} \mid W] = \mathbb{E}[D \mid W]$.

Recall from the definition of path-specific counterfactuals that $D_{\Pi,A,a'} = f_D(X_{\Pi,A,a'}, U_D) = \mathbb{1}_{U_D \leq d(X_{\Pi,A,a'})}$, where $U_D \perp\!\!\!\perp \{X_{\Pi,A,a}, X\}$. Since $W = \omega(X)$, $U_D \perp\!\!\!\perp$

$\{X_{\Pi,A,a}, W\}$, it follows that

$$\begin{aligned}
&\mathbb{E}[D_{\Pi,A,a'} \mid W = w] \\
&\quad = \sum_{i=1}^{m} \mathbb{E}[D_{\Pi,A,a'} \mid X_{\Pi,A,a} = x_i, W = w] \\
&\qquad \cdot \Pr(X_{\Pi,A,a} = x_i \mid W = w) \\
&\quad = \sum_{i=1}^{m} \mathbb{E}[\mathbb{1}_{U_D \le d(X_{\Pi,A,a'})} \mid X_{\Pi,A,a} = x_i, W = w] \\
&\qquad \cdot \Pr(X_{\Pi,A,a'} = x_i \mid W = w) \\
&\quad = \sum_{i=1}^{m} d(X_{\Pi,A,a'}) \cdot \Pr(X_{\Pi,A,a'} = x_i \mid W = w) \\
&\quad = \sum_{i=1}^{m} d_i \cdot \Pr(X_{\Pi,A,a'} = x_i \mid W = w).
\end{aligned}$$

An analogous calculation yields that $\mathbb{E}[D \mid W = w] = \sum_{i=1}^{m} d_i \cdot \Pr(X = x_i \mid W = w)$. Equating these expressions gives

$$\begin{aligned}
&\sum_{i=1}^{m} d_i \cdot \Pr(X = x_i \mid W = w) \\
&\qquad = \sum_{i=1}^{m} d_i \cdot \Pr(X_{\Pi,A,a'} = x_i \mid W = w)
\end{aligned}$$

for each $a' \in \mathcal{A}$ and each of the finitely many $w \in \mathcal{W}$ such that $\Pr(W = w) > 0$. Again, each of these constraints is linear by inspection.

Theorem B.1 Part 3 A decision policy satisfies counterfactual predictive parity if $Y(1) \perp\!\!\!\perp A \mid D = 0$, or equivalently, $\Pr(Y(1) = y \mid A = a, D = 0) = \Pr(Y(1) \mid D = 0)$ for all $a \in \mathcal{A}$. We may rewrite this expression to obtain:

$$\frac{\Pr(Y(1) = y, A = a, D = 0)}{\Pr(A = a, D = 0)} = C_y,$$

where $C_y = \Pr(Y(1) = y \mid D = 0)$.

Expanding the numerator on the left-hand side of the above equation yields

$$\begin{aligned}
&\Pr(Y(1) = y, A = a, D = 0) \\
&\quad = \sum_{i=1}^{m} [1 - d_i] \cdot \Pr(Y(1) = y, A = a, X = x_i)
\end{aligned}$$

Similarly, expanding the denominator yields

$$\begin{aligned}
&\Pr(Y(1) = y, D = 0) \\
&\qquad = \sum_{i=1}^{m} [1 - d_i] \cdot \Pr(Y(1) = y, X = x_i).
\end{aligned}$$

for each of the finitely many $y \in \mathcal{Y}$. Therefore, counterfactual predictive parity corresponds to

$$\begin{aligned}
&\sum_{i=1}^{m} [1 - d_i] \cdot \Pr(Y(1) = y, A = a, X = x_i) \\
&\qquad = C_y \cdot \sum_{i=1}^{m} [1 - d_i] \cdot \Pr(Y(1) = y, X = x_i),
\end{aligned} \tag{10}$$

for each $a \in \mathcal{A}$ and $y \in \mathcal{Y}$. Again, these constraints are linear in the d_i by inspection.

Consider the family of linear programs $\mathcal{L} = \{L(v)\}_{v \in \Delta^k}$ where the linear program $L(v)$ has the same objective function $\sum_{i=1}^{m} d_i \cdot u(x_i) \cdot p_i$ and budget constraint $\sum_{i=1}^{m} d_i \cdot p_i \le b$ as before, together with additional constraints for each $a \in \mathcal{A}$ as in Eq. (10), where $C_{y_i} = v_i$ for $i = 1, \ldots, k$.

By assumption, there exists a feasible solution to the optimization problem in Eq. (7), so the solution to at least one program in $\mathcal{L}$ is a utility-maximizing policy that satisfies counterfactual predictive parity.

□

C. A stylized example of college admissions

In the example we consider in Section 2.1, the exogenous variables $\mathcal{U} = \{u_A, u_D, u_E, u_M, u_T, u_Y\}$ in the DAG are independently distributed as follows:

$$\begin{aligned}
U_A, U_D, U_Y &\sim \textsc{Unif}(0, 1), \\
U_E, U_M, U_T &\sim \mathrm{N}(0, 1).
\end{aligned}$$

For fixed constants μ_A, $\beta_{E,0}$, $\beta_{E,A}$, $\beta_{M,0}$, $\beta_{M,E}$, $\beta_{T,0}$, $\beta_{T,E}$, $\beta_{T,M}$, $\beta_{T,B}$, $\beta_{T,u}$, $\beta_{Y,0}$, $\beta_{Y,D}$, we define the endogenous variables $\mathcal{V} = \{A, E, M, T, D, Y\}$ in the DAG by the following structural equations:

$$\begin{aligned}
f_A(u_A) &= \begin{cases} a_1 & \text{if } u_A \le \mu_A \\ a_0 & \text{otherwise} \end{cases}, \\
f_E(a, u_E) &= \beta_{E,0} + \beta_{E,A} \cdot \mathbb{1}(a = a_1) + u_E, \\
f_M(e, u_M) &= \beta_{M,0} + \beta_{M,E} \cdot e + u_M, \\
f_T(e, m, u_T) &= \beta_{T,0} + \beta_{T,E} \cdot e \\
&\qquad + \beta_{T,M} \cdot m + \beta_{T,B} \cdot e \cdot m + \beta_{T,u} \cdot u_T, \\
f_D(x, u_D) &= \mathbb{1}(u_D \le d(x)), \\
f_Y(m, u_Y, \delta) &= \mathbb{1}(u_Y \le \text{logit}^{-1}(\beta_{Y,0} + m + \beta_{Y,D} \cdot \delta)),
\end{aligned}$$

where $\text{logit}^{-1}(x) = (1 + \exp(-x))^{-1}$ and $d(x)$ is the decision policy. In our example, we use constants $\mu_A = \frac{1}{3}$, $\beta_{E,0} = 1$, $\beta_{E,A} = -1$ $\beta_{M,0} = 0$, $\beta_{M,E} = 1$, $\beta_{T,0} = 50$, $\beta_{T,E} = 4$, $\beta_{T,M} = 4$, $\beta_{T,u} = 7$, $\beta_{T,B} = 1$, $\beta_{Y,0} = -\frac{1}{2}$, $\beta_{Y,D} = \frac{1}{2}$. We also assume a budget $b = \frac{1}{2}$.

D. Proof of Proposition 1

We begin by more formally defining (multiple) threshold policies. We assume, without loss of generality, that $\Pr(A = a) > 0$ for all $a \in \mathcal{A}$ throughout.

Definition D.1. Let $u(x)$ be a utility function. We say that a policy $d(x)$ is a *threshold policy* with respect to u if there exists some t such that

$$d(x) = \begin{cases} 1 & u(x) > t, \\ 0 & u(x) < t, \end{cases}$$

and $d(x) \in [0, 1]$ is arbitrary if $u(x) = t$.

We say that $d(x)$ is a *multiple threshold policy* with respect to u if there exist group-specific constants t_a for $a \in \mathcal{A}$ such that

$$d(x) = \begin{cases} 1 & u(x) > t_{\alpha(x)}, \\ 0 & u(x) < t_{\alpha(x)}, \end{cases}$$

and $d(x) \in [0, 1]$ is arbitrary if $u(x) = t_{\alpha(x)}$.

Remark 1. In general, it is possible for different thresholds to produce threshold policies that are almost surely equal. For instance, if $u(X) \sim \textsc{Bern}(\frac{1}{2})$, then the policies $\mathbb{1}_{u(X)>p}$ are almost surely equal for all $p \in [0, 1)$. Nevertheless, we speak in general of *the* threshold associated with the threshold policy $d(X)$ unless there is ambiguity.

We first observe that if $\mathcal{U}$ is consistent modulo α, then whether a decision policy $d(x)$ is a multiple threshold policy does not depend on our choice of $u \in \mathcal{U}$.

Lemma D.1. *Let $\mathcal{U}$ be a collection of utilities consistent modulo α, and suppose $d : \mathcal{X} \to [0, 1]$ is a decision rule. If $d(x)$ is a multiple threshold rule with respect to a utility $u^* \in \mathcal{U}$, then $d(x)$ is a multiple threshold rule with respect to every $u \in \mathcal{U}$. In particular, if $d(x)$ can be represented by non-negative thresholds over u^*, it can be represented by non-negative thresholds over any $u \in \mathcal{U}$.*

Proof. Suppose $d(x)$ is represented by thresholds $\{t^*_a\}_{a\in\mathcal{A}}$ with respect to u^*. We construct the thresholds $\{t_a\}_{a\in\mathcal{A}}$ explicitly.

Fix $a \in \mathcal{A}$ and suppose that there exists $x^* \in \alpha^{-1}(a)$ such that $u^*(x^*) = t^*_a$. Then set $t_a = u(x^*)$. Now, if $u(x) > t_a = u(x^*)$ then, by consistency modulo α, $u^*(x) > u^*(x^*) = t^*_a$. Similarly if $u(x) < t_a$ then $u^*(x) < t^*_a$. We also note that by consistency modulo α, $\operatorname{sign}(t_a) = \operatorname{sign}(u(x^*)) = \operatorname{sign}(u^*(x^*)) = \operatorname{sign}(t^*_a)$.

If there is no $x^* \in \alpha^{-1}(a)$ such that $u^*(x^*) = t^*_a$, then let

$$t_a = \inf_{x\in S_a} u(x)$$

where $S_a = \{x \in \alpha^{-1}(a) \mid u^*(x) > t^*_a\}$. Note that since $\operatorname{sign}(u(x)) = \operatorname{sign}(u^*(x))$ for all x by consistency modulo α, if $t^*_a \geq 0$, it follows that $t_a \geq 0$ as well.

We need to show in this case also that if $u(x) > t_a$ then $u^*(x) > t^*_a$, and if $u(x) < t_a$ then $u^*(x) < t^*_a$. To do so, let $x \in \alpha^{-1}(a)$ be arbitrary, and suppose $u(x) > t_a$. Then, by definition, there exists $x' \in \alpha^{-1}(a)$ such that $u(x) > u(x') > t_a$ and $u^*(x') > t^*_a$, whence $u^*(x) > u^*(x') > t^*_a$ by consistency modulo α. On the other hand, if $u(x) < t_a$, it follows by the definition of t_a that $u^*(x) \leq t^*_a$; since $u^*(x) \neq t^*_a$ by hypothesis, it follows that $u^*(x) < t^*_a$.

Therefore, it follows in both cases that for $x \in \alpha^{-1}(a)$, if $u(x) > t_a$ then $u^*(x) > t^*_a$, and if $u(x) < t_a$ then $u^*(x) < t^*_a$. Therefore

$$d(x) = \begin{cases} 1 & \text{if } u(x) > t_{\alpha(x)}, \\ 0 & \text{if } u(x) < t_{\alpha(x)}, \end{cases}$$

i.e., $d(x)$ is a multiple threshold policy with respect to u. Moreover, as noted above, if $t^*_a \geq 0$ for all $a \in \mathcal{A}$, then $t_a \geq 0$ for all $a \in \mathcal{A}$. □

We now prove the following strengthening of Prop. 1.

Lemma D.2. *Let $\mathcal{U}$ be a collection of utilities consistent modulo α. Let $d(x)$ be a decision policy that is not a.s. a multiple threshold policy with non-negative thresholds with respect to $\mathcal{U}$, then $d(x)$ is strongly Pareto dominated.*

Proof. We prove the claim in two parts. First, we show that any policy that is not a multiple threshold policy is strongly Pareto dominated. Then, we show that any multiple threshold policy that cannot be represented with non-negative thresholds is strongly Pareto dominated.

If $d(x)$ is not a multiple threshold policy, then there exists a $u \in \mathcal{U}$ and $a^* \in \mathcal{A}$ such that $d(x)$ is not a threshold policy when restricted to $\alpha^{-1}(a^*)$ with respect to u.

We will construct an alternative policy $d'(x)$ that attains strictly greater utility on $\alpha^{-1}(a^*)$ and is identical elsewhere. Thus, without loss of generality, we assume there is a single group, i.e., $\alpha(x) = a^*$. The proof proceeds heuristically by moving some of the mass below a threshold to above a threshold to create a feasible policy with improved utility.

Let $b = \mathbb{E}[d(X)]$. Define

$$m^{\text{Lo}}(t) = \mathbb{E}[d(X) \cdot \mathbb{1}_{u(X)<t}],$$
$$m^{\text{Up}}(t) = \mathbb{E}[(1 - d(X)) \cdot \mathbb{1}_{u(X)>t}].$$

We show that there exists t^* such that $m^{\text{Up}}(t^*) > 0$ and $m^{\text{Lo}}(t^*) > 0$. For, if not, consider

$$\tilde{t} = \inf\{t \in \mathbb{R} : m^{\text{Up}}(t) = 0\}.$$

Note that $d(X) \cdot \mathbb{1}_{u(X)>\tilde{t}} = \mathbb{1}_{u(X)>\tilde{t}}$ a.s. If $\tilde{t} = -\infty$, then by definition $d(X) = 1$ a.s., which is a threshold policy, violating our assumption on $d(X)$. If $\tilde{t} > -\infty$, then for any

$t' < \tilde{t}$, we have, by definition that $m^{\mathrm{Up}}(t') > 0$, and so by hypothesis $m^{\mathrm{Lo}}(t') = 0$. Therefore $d(X) \cdot \mathbb{1}_{u(X)<\tilde{t}} = 0$ a.s., and so, again, $d(X)$ is a threshold policy, contrary to hypothesis.

Now, with t^* as above, for notational simplicity, let $m^{\mathrm{Up}} = m^{\mathrm{Up}}(t^*)$ and $m^{\mathrm{Lo}} = m^{\mathrm{Lo}}(t^*)$ and consider the alternative policy

$$d'(x) = \begin{cases} (1 - m^{\mathrm{Up}}) \cdot d(x) & u(x) < t^*, \\ d(x) & u(x) = t^*, \\ 1 - (1 - m^{\mathrm{Lo}}) \cdot (1 - d(x)) & u(x) > t^*. \end{cases}$$

Then it follows by construction that

$$\begin{aligned} \mathbb{E}[d'(X)] &= (1 - m^{\mathrm{Up}}) \cdot m^{\mathrm{Lo}} + \mathbb{E}[d(x) \cdot *\mathbb{1}_{u(X)=t*}] \\ &\quad + \Pr(u(X) > t*) - (1 - m^{\mathrm{Lo}}) \cdot m^{\mathrm{Up}} \\ &= m^{\mathrm{Lo}} + \mathbb{E}[d(x) \cdot \mathbb{1}_{u(X)=t*}] \\ &\quad + \Pr(u(X) > t*) - m^{\mathrm{Up}} \\ &= \mathbb{E}[d(X) \cdot \mathbb{1}_{u(X)<t*}] + \mathbb{E}[d(x) \cdot \mathbb{1}_{u(X)=t*}] \\ &\quad + \mathbb{E}[\mathbb{1}_{u(X)>t*}] - \mathbb{E}[(1 - d(X)) \cdot \mathbb{1}_{u(X)>t*}] \\ &= \mathbb{E}[d(X)] \\ &= b, \end{aligned}$$

so $d'(x)$ is feasible. However,

$$\begin{aligned} d'(x) - d(x) = m^{\mathrm{Lo}} \cdot (1 - d(x)) \cdot \mathbb{1}_{u(x)>t^*} \\ - m^{\mathrm{Up}} \cdot d(x) \cdot \mathbb{1}_{u(x)<t^*}, \end{aligned}$$

and so

$$\begin{aligned} &\mathbb{E}[(d'(X) - d(X)) \cdot u(X)] \\ &\quad = m^{\mathrm{Lo}} \cdot \mathbb{E}[(1 - d(X)) \cdot \mathbb{1}_{u(X)>t^*} \cdot u(X)] \\ &\qquad - m^{\mathrm{Up}} \cdot \mathbb{E}[d(X) \cdot \mathbb{1}_{u(X)<t^*} \cdot u(X)] \\ &\quad > m^{\mathrm{Lo}} \cdot t^* \cdot \mathbb{E}[(1 - d(X)) \cdot \mathbb{1}_{u(X)>t^*}] \\ &\qquad - m^{\mathrm{Up}} \cdot t^* \cdot \mathbb{E}[d(X) \cdot \mathbb{1}_{u(X)<t^*}] \\ &\quad = t^* \cdot m^{\mathrm{Lo}} \cdot m^{\mathrm{Up}} - t^* \cdot m^{\mathrm{Up}} \cdot m^{\mathrm{Lo}} \\ &\quad = 0. \end{aligned}$$

Therefore

$$\mathbb{E}[d(X) \cdot u(X)] < \mathbb{E}[d'(X) \cdot u(X)].$$

It remains to hold that $u'(d') > u'(d)$ for arbitrary $u' \in \mathcal{U}$. Let

$$t' = \inf\{u'(x) : d'(x) > d(x)\}.$$

Note that by construction for any $x, x' \in \mathcal{X}$, if $d'(x) > d(x)$ and $d'(x') < d(x')$, then $u(x) > t > u(x')$. It follows by consistency modulo α that $u'(x) \geq t' \geq u'(x')$ for any $u' \in \mathcal{U}$, and, moreover, that at least one of the inequalities is strict.

Without loss of generality, assume $u'(x) > t' \geq u'(x')$. It follows that $\mathbb{E}[(d'(X) - d(X)) \cdot \mathbb{1}_{u'(X)>t'}] = m^{\mathrm{Up}} > 0$. Since $\mathbb{E}[d'(x) - d(x)] = 0$, we see that

$$\begin{aligned} &\mathbb{E}[(d'(X) - d(X)) \cdot u'(X)] \\ &\quad = \mathbb{E}[(d'(X) - d(X)) \cdot \mathbb{1}_{u'(X)>t'} \cdot u'(X)] \\ &\qquad + \mathbb{E}[(d'(X) - d(X)) \cdot \mathbb{1}_{u'(X)\leq t'} \cdot u'(X)] \\ &\quad > t' \cdot \mathbb{E}[(d'(X) - d(X)) \cdot \mathbb{1}_{u'(X)>t'}] \\ &\qquad + t' \cdot \mathbb{E}[(d'(X) - d(X)) \cdot \mathbb{1}_{u'(X)\leq t'}] \\ &\quad = t' \cdot \mathbb{E}[d'(X) - d(X)] \\ &\quad = 0, \end{aligned}$$

where in the inequality we have used the fact that if $d'(x) > d(x)$, $u'(x) > t$, and if $d'(x) < d(x)$, $u'(x) \leq t'$. Therefore

$$\mathbb{E}[d(X) \cdot u'(X)] < \mathbb{E}[d'(X) \cdot u'(X)],$$

i.e., $d'(x)$ strongly Pareto dominates $d(x)$.

Now, we prove the second claim, namely, that a multiple threshold policy $\tau(x)$ that cannot be represented with non-negative thresholds is strongly Pareto dominated. For, if $\tau(x)$ is such a policy, then, by Lemma D.1, for any $u \in \mathcal{U}$, $\mathbb{E}[\tau(X) \cdot \mathbb{1}_{u(X)<0}] > 0$. It follows immediately that $\tau'(x) = \tau(x) \cdot \mathbb{1}_{u(x)>0}$ satisfies $u(\tau') > u(\tau)$. By consistency modulo α, the definition of $\tau'(x)$ does not depend on our choice of u, and so $u(\tau') > u(\tau)$ for every $u \in \mathcal{U}$, i.e., $\tau'(x)$ strongly Pareto dominates $\tau(x)$. □

The following corollary of Lemma D.2 and lemma are useful in the proof of Theorem 1.

Definition D.2. We say that a decision policy $d(x)$ is *budget-exhausting* if

$$\begin{aligned} \min(b, \Pr(u(X) > 0)) &\leq \mathbb{E}[d(X)] \\ &\leq \min(b, \Pr(u(X) \geq 0)). \end{aligned}$$

Remark 2. We note that if $\mathcal{U}$ is consistent modulo α, then whether or not a decision policy $d(x)$ is budget-exhausting does not depend on the choice of $u \in \mathcal{U}$. Further, if $\Pr(u(X) = 0) = 0$—e.g., if the distribution of X is $\mathcal{U}$-fine—then the decision policy is budget-exhausting if and only if $\mathbb{E}[d(X)] = \min(b, \Pr(u(X) > 0))$.

Corollary D.1. *Let $\mathcal{U}$ be a collection of utilities consistent modulo α. If $d(x)$ is not a budget-exhausting multiple threshold policy with non-negative thresholds, then $d(x)$ is strongly Pareto dominated.*

Proof. Suppose $\tau(x)$ is not strongly Pareto dominated. By Prop. 1, it is a multiple threshold policy with non-negative thresholds.

Now, suppose toward a contradiction that $\tau(x)$ is not budget-exhausting. Then, either $\mathbb{E}[\tau(X)] > \min(b, \Pr(u(X) \geq 0))$ or $\mathbb{E}[\tau(X)] < \min(b, \Pr(u(X) > 0))$.

In the first case, since $\tau(x)$ is feasible, it follows that $\mathbb{E}[\tau(X)] > \Pr(u(X) \geq 0)$. It follows that $\tau(x) \cdot \mathbb{1}_{u(x)<0}$ is not almost surely zero. Therefore

$$\mathbb{E}[\tau(X)] < \mathbb{E}[\tau(X) \cdot \mathbb{1}_{u(X)>0}],$$

and, by consistency modulo α, this holds for any $u \in \mathcal{U}$. Therefore $\tau(x)$ is strongly Pareto dominated, contrary to hypothesis.

In the second case, consider

$$d(x) = [\theta + (1-\theta) \cdot \tau(x)] \cdot \mathbb{1}_{u(x)>0}.$$

Since $\mathbb{E}[\tau(X)] < \min(b, \Pr(u(X) > 0))$ and

$$\mathbb{E}[d(X)] = \theta \cdot \Pr(u(X) > 0) + (1-\theta) \cdot \mathbb{E}[\tau(X)],$$

there exists some $\theta > 0$ such that $d(x)$ is feasible. For that θ, a similar calculation shows immediately that $u(d) > u(\tau)$, and, by consistency modulo α, $u'(d) > u'(\tau)$ for all $u' \in \mathcal{U}$. Therefore, again, $d(x)$ strongly Pareto dominates $\tau(x)$, contrary to hypothesis. □

Lemma D.3. *Given a utility u, there exists a mapping T from $[0,1]^{\mathcal{A}}$ to $[-\infty, \infty]^{\mathcal{A}}$ taking sets of quantiles $\{q_a\}_{a \in \mathcal{A}}$ to thresholds $\{t_a\}_{a \in \mathcal{A}}$ such that:*

1. *T is monotonically non-increasing in each coordinate;*
2. *For each set of quantiles, there is a multiple threshold policy $\tau : \mathcal{X} \to [0,1]$ with thresholds $T(\{q_a\})$ with respect to u such that $\mathbb{E}[\tau(X) \mid A = a] = q_a$.*

Proof. Simply choose

$$t_a = \inf\{s \in \mathbb{R} : \Pr(u(X) > s) < q_a\}. \tag{11}$$

Then define

$$p_a = \begin{cases} \frac{q_a - \Pr(u(X) > t_a \mid A=a)}{\Pr(u(X) = t_a \mid A=a)} & \Pr(u(X) = t_a, A = a) > 0 \\ 0 & \Pr(u(X) = t_a, A = a) = 0. \end{cases}$$

Note that $\Pr(u(X) \geq t_a \mid A = a) \geq q_a$, since, by definition, $\Pr(u(X) > t_a - \epsilon \mid A = a) \geq q_a$ for all $\epsilon > 0$. Therefore,

$$\Pr(u(X) > t_a \mid A = a) + \Pr(u(X) = t_a \mid A = a) \geq q_a,$$

and so $p_a \leq 1$. Further, since $\Pr(u(X) > t_a \mid A = a) \leq q_a$, we have that $p_a \geq 0$.

Finally, let

$$d(x) = \begin{cases} 1 & u(x) > t_{\alpha(x)}, \\ p_a & u(x) = t_{\alpha(x)}, \\ 0 & u(x) < t_{\alpha(x)}, \end{cases}$$

and it follows immediately that $\mathbb{E}[d(X) \mid A = a] = q_a$. That t_a is a monotonically non-increasing function of q_a follows immediately from Eq. (11). □

We can further refine Cor. D.1 and Lemma D.3 as follows:

Lemma D.4. *Let u be a utility. Then a policy is utility maximizing if and only if it is a budget-exhausting threshold policy. Moreover, there exists at least one utility maximizing policy.*

Proof. Let $\bar{\alpha}$ be a constant map, i.e., $\bar{\alpha} : \mathcal{X} \to \bar{\mathcal{A}}$, where $|\bar{\mathcal{A}}| = 1$. Then $\mathcal{U} = \{u\}$ is consistent modulo $\bar{\alpha}$, and so by Cor. D.1, any Pareto efficient policy is a budget exhausting multiple threshold policy relative to $\mathcal{U}$. Since $\mathcal{U}$ contains a single element, a policy is Pareto efficient if and only if it is utility maximizing. Since $\bar{\alpha}$ is constant, a policy is a multiple threshold policy relative to $\bar{\alpha}$ if and only if it is a threshold policy. Therefore, a policy is utility maximizing if and only if it is a budget exhausting threshold policy. By Lemma D.3, such a policy exists, and so the maximum is attained. □

E. Prevalence and the Proof of Theorem 1

The notion of a probabilistically "small" set—such as the event in which an idealized dart hits the exact center of a target—is, in finite-dimensional real vector spaces, typically encoded by the idea of a Lebesgue null set.

Here we prove that the set of distributions such that there exists a policy satisfying either counterfactual equalized odds, conditional principal fairness, or counterfactual fairness that is not strongly Pareto domianted is "small" in an analogous sense. The proof turns on the following intuition. Each of the fairness definitions imposes a number of constraints. By Lemma D.2, any policy that is not strongly Pareto dominated is a multiple threshold policy. By adjusting the group-specific thresholds of such a policy, one can potentially satisfy one constraint per group. If there are more constraints than groups constraints, then one has no additional degrees of freedom that can be used to ensure that the remaining constraints are satisfied. If, by chance, those constraints *are* satisfied with the same threshold policy, they are not satisfied robustly—even a minor distribution shift, such as increasing the amount of mass above the threshold by any amount on the relevant subpopulation, will break them. Therefore, over a "typical" distribution, at most $|\mathcal{A}|$ of the constraints can simultaneously be satisfied by a Pareto efficient policy, meaning that typically no Pareto efficient policy fully satisfies all of the conditions of the fairness definitions.

Formalizing this intuition, however, requires considerable care. In Section E.1, we give a brief introduction to a popular generalization of null sets to infinite-dimensional vector spaces, drawing heavily on a review article by Ott & Yorke (2005). In Section E.2 we provide a roadmap of the proof itself. In Section E.3, we establish the main hypotheses necessary to apply the notion of prevalence to a convex set—in

our case, the set of $\mathcal{U}$-fine distributions. In Section E.4, we establish a number of technical lemmata used in the proof of Theorem 1, and provide a proof of the theorem itself in Section E.5. In Section E.6, we show why the hypothesis of $\mathcal{U}$-fineness is important and how conspiracies between atoms in the distribution of $u(X)$ can lead to "robust" counterexamples in the general setting.

E.1. Shyness and Prevalence

Lebesgue measure λ_n on $\mathbb{R}^n$ has a number of desirable properties:

- **Local finiteness:** For any point $v \in \mathbb{R}^n$, there exists an open set U containing x such that $\lambda_n[U] < \infty$;
- **Strict positivity:** For any open set U, if $\lambda_n[U] = 0$, then $U = \emptyset$;
- **Translation invariance:** For any $v \in \mathbb{R}^n$ and measurable set E, $\lambda_n[E + v] = \lambda_n[E]$.

No measure on an infinite-dimensional, separable Banach space, such as $L^1(\mathbb{R})$, can satisfy these three properties (Ott & Yorke, 2005). However, while there is no generalization of Lebesgue measure to infinite dimensions, there is a generalization of Lebesgue null sets—called *shy* sets—to the infinite-dimensional context that preserves many of their desirable properties.

Definition E.3 (Hunt et al. (1992))**.** Let V be a completely metrizable topological vector space. We say that a Borel set $E \subseteq V$ is *shy* if there exists a Borel measure μ on V such that:

1. There exists $C \subseteq V$ such that $0 < \mu[C] < \infty$,
2. For all $v \in V$, $\mu[E + v] = 0$.

An arbitrary set $F \subseteq V$ is shy if there exists a shy Borel set $E \subseteq V$ containing F.

We say that a set is *prevalent* if its complement is shy.

Prevalence generalizes the concept of Lebesgue "full measure" or "co-null" sets (i.e., sets whose complements have null Lebesgue measure) in the following sense:

Proposition E.3 (Hunt et al. (1992))**.** *Let V be a completely metrizable topological vector space. Then:*

- *A prevalent set is dense in V;*
- *If $G \subseteq L$ and G is prevalent, then L is prevalent;*
- *A countable intersection of prevalent sets is prevalent;*
- *Every translate of a prevalent set is prevalent;*
- *If $V = \mathbb{R}^n$, then $G \subseteq \mathbb{R}^n$ is prevalent if and only if $\lambda_n[\mathbb{R}^n \setminus G] = 0$, where λ_n denotes n-dimensional Lebesgue measure.*

As is conventional for sets of full measure in finite-dimensional spaces, if some property holds for every $v \in E$, where E is prevalent, then we say that the property holds for *almost every* $v \in V$ or that it holds *generically in* V.

Prevalence can also be generalized from vector spaces to convex subsets of vector spaces, although additional care must be taken to ensure that a relative version of Prop. E.3 holds.

Definition E.4 (Anderson & Zame (2001))**.** Let V be a topological vector space and let $C \subseteq V$ be a convex subset completely metrizable in the subspace topology induced by V. We say that a universally measurable set $E \subseteq C$ is *shy in C at $c \in C$* if for each $1 \geq \delta > 0$, and each neighborhood U of 0 in V, there is a regular Borel measure μ with compact support such that

$$\text{SUPP}(\mu) \subseteq (\delta(C - c) + c) \cap (U + c),$$

and $\mu[E + v] = 0$ for every $v \in V$.

We say that E is *shy in C* if E is *shy in C at c* for every $c \in C$. An arbitrary set $F \subseteq V$ is shy in C if there exists a universally measurable shy set $E \subseteq C$ containing F.

A set G is *prevalent* in C if $C \setminus G$ is shy in C.

Proposition E.4 (Anderson & Zame (2001))**.** *If E is shy in C for any $c \in C$, E is shy in C.*

Sets that are shy in C enjoy similar properties to sets that are shy in V.

Proposition E.5 (Anderson & Zame (2001))**.** *Let V be a topological vector space and let $C \subseteq V$ be a convex subset completely metrizable in the subspace topology induced by V. Then:*

- *A prevalent set in C is dense in C;*
- *If $G \subseteq L$ and G is prevalent in C, then L is prevalent in C;*
- *A countable intersection of sets prevalent in C is prevalent in C*
- *If G is prevalent in C then $G + v$ is prevalent in $C + v$ for all $v \in V$.*
- *If $V = \mathbb{R}^n$ and $C \subseteq V$ is a convex subset with nonempty interior, then $G \subseteq C$ is prevalent in C if and only if $\lambda_n[C \setminus G] = 0$.*

Sets that are shy in C can often be identified by inspecting their intersections with a finite-dimensional subspace W of V, a strategy we use to prove Theorem 1.

Definition E.5 (Anderson & Zame (2001))**.** A universally measurable set $E \subseteq C$ is said to be *k-shy in C* if there exists a k-dimensional subspace $W \subseteq V$ such that

1. A translate of the set C has positive measure in W, i.e., $\lambda_W[C + v_0] > 0$ for some $v_0 \in V$;
2. Every translate of the set E is a null set in W, i.e., $\lambda_W[E + v] = 0$ for all $v \in V$.

Here λ_W denotes k-dimensional Lebesgue measure supported on W.[12] We refer to such a W as a *k-dimensional probe* witnessing the k-shyness of E, and to an element $w \in W$ as a *perturbation*.

The following intuition motivates the use of probes to detect shy sets. By analogy with Fubini's theorem, one can imagine trying to determine whether a subset of a finite-dimensional vector space is large or small by looking at its cross sections parallel to some subspace $W \subseteq V$. If a set $E \subseteq V$ is small in each cross section—i.e., if $\lambda_W[v + E] = 0$ for all $v \in V$—then E itself is small in V, i.e., E has λ_V-measure zero.

Proposition E.6 (Anderson & Zame (2001))**.** *Every k-shy set in C is shy in C.*

E.2. Outline

To aid the reader in following the application of the theory in Section E.1 to the proof of Theorem 1, we provide the following outline of the argument.

In **Section E.3** we establish the context to which we apply the notion of relative shyness. In particular, we introduce the vector space $\mathbb{K}$ consisting of the *totally bounded Borel measures* on the state space $\mathcal{K}$—where $\mathcal{K}$ is $\mathcal{X} \times \mathcal{Y}$, $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$, or $\mathcal{A} \times \mathcal{X}^{\mathcal{A}}$, depending on which notion of fairness is under consideration. We further isolate the subspace $\mathbf{K} \subseteq \mathbb{K}$ of $\mathcal{U}$-fine totally bounded Borel measures. Within this space, we are interested in the convex set $\mathbf{Q} \subseteq \mathbb{K}$, the set of *$\mathcal{U}$-fine probability distributions* of X and $Y(1)$. Withing $\mathbf{Q}$, we identify $\mathbf{E} \subseteq \mathbf{Q}$, the set of $\mathcal{U}$-fine distributions on $\mathcal{K}$ *over which there exists a policy satisfying the relevant fairness definition that is not strongly Pareto dominated*. The claim of Theorem 1 is that $\mathbf{E}$ is shy relative to $\mathbf{Q}$.

To ensure that relative shyness generalizes Lebesgue null measure in the expected way—i.e., that Prop. E.5 holds—Definition E.4 has three technical requirements: (1) that the ambient vector space V be a topological vector space; (2) that the convex set C be completely metrizable; and (3) that the shy set E be universally measurable. In **Lemma E.7**, we observe that $\mathbb{K}$ is a complete topological vector space under the total variation norm, i.e., $\mathbb{K}$ is a Banach space. We extend this in **Cor. E.2**, and show that $\mathbf{K}$ is also a Banach space. We use this fact in **Lemma E.11** to show that $\mathbf{Q}$ is a completely metrizable subset of $\mathbf{K}$, as well as convex. Lastly, in **Lemma E.13**, we show that the set $\mathbf{E}$ is closed, and therefore universally measurable, where $\mathbf{E}$ is the set of$\mathcal{U}$-fine distributions on $\mathcal{K}$ such that there exists a policy satisfying the relevant fairness definition that is not strongly Pareto dominated.

In **Section E.4**, we develop the machinery needed to construct a probe $\mathbf{W}$ for the proof of Theorem 1 and prove several lemmata simplifying the eventual proof of the theorem. To build the probe, it is necessary to construct measures whose support on the utility scale is maximal. This ensures that the probe can distinguish between any threshold policies between which *any* $\mu \in \mathbf{K}$ can distinguish. The construction of these measures, the $\mu_{\max,a}$, is carried out in **Lemma E.14** and **Cor. E.3**. Next, we introduce the basic style of argument used to show that a subset of $\mathbf{Q}$ is shy in **Lemma E.15** and **Lemma E.16**, in particular, by showing that if a fixed $\mu \in \mathbf{Q}$ "sees" two disjoint sets E_0 and E_1, then so does a typical element of $\mathbf{Q}$. We use then use a technical lemma, **Lemma E.17**, to show, in effect, that a generic element of $\mathbf{Q}$ has support on the utility scale wherever a given fixed distribution $\mu \in \mathbf{Q}$ does. **In Defn. E.12**, we introduce the concept of overlapping and splitting utilities, and show in **Lemma E.19** that this property is generic in $\mathbf{Q}$ unless there exists a ω-stratum that contains no positive-utility observables x. Lastly, in **Lemma E.20**, we provide a mild simplification of the characterization of finitely shy sets that makes the the proof of Theorem 1 more straightforward.

Finally, in **Section E.5**, we give the proof of Theorem 1. We divide the proof into three parts. In the first part, we restrict our attention to the case of counterfactual equalized odds, and show in detail how to combine the lemmata of the previous section to construct the (at most) $2 \cdot |\mathcal{A}|$-dimensional probe $\mathbf{W}$. In the second part we consider two distinct cases. The argument in both cases is conceptually parallel. First, we argue that the balance conditions of counterfactual equalized odds encoded by Eq. (2) must be broken by a typical perturbation in $\mathbf{W}$. In particular, we argue that for a given base distribution μ, there can be at most one budget-exhausting multiple threshold policy that can—although need not necessarily—satisfy counterfactual equalized odds. We show that the form of this policy cannot be altered by an appropriate perturbation in $\mathbf{W}$, but that the conditional probability of a positive decision will, in general, be altered in such a way that Eq. (2) can only hold for a $\lambda_{\mathbf{W}}$-null set of perturbations. In the final section, we lay out modiciations that can be made to the proof given for counterfactual equalized odds in the first two parts that

[12] Note that Lebesgue measure on W is only defined up to a choice of basis; however, since $\lambda[T(A)] = |\det(T)| \cdot \lambda[A]$ for any linear automorphism T and Lebesgue measure λ, whether a set has null measure does not depend on the choice of basis.

adapt the argument to the cases of conditional principal fairness and path-specific fairness. In particular, we show how to construct the probe $\mathbf{W}$ in such a way that the additional conditioning on the reduced covariates $W = \omega(X)$ in Eqs. (3) and (5) does not affect the argument.

E.3. Convexity, complete metrizability, and universal measurability

In this section, we establish the background requirements of Prop. E.6 for the setting of Theorem 1. In particular, we exhibit the $\mathcal{U}$-fine distributions as a convex subset of a topological vector space, the set of totally bounded $\mathcal{U}$-fine Borel measures. We show that the regular $\mathcal{U}$-fine distributions form a completely metrizable subset in the topology it inherits from the space of totally bounded measures. Lastly, we show that the set of regular distributions under which there exists a Pareto efficient policy satisfying one of the three fairness criteria is universally measurable.

E.3.1. Background and notation

We begin by establishing some notational conventions. We let $\mathcal{K}$ denote the underlying state space over which the distributions in Theorem 1 range. Specifically, $\mathcal{K} = \mathcal{X} \times \mathcal{Y}$ in the case of counterfactual equalized odds; $\mathcal{K} = \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ in the case of conditional principal fairness; and $\mathcal{K} = \mathcal{A} \times \mathcal{X}^{\mathcal{A}}$ in the case of path-specific fairness. We note that since $\mathcal{X} \subseteq \mathbb{R}^k$ for some k and $Y \subseteq \mathbb{R}$, $\mathcal{K}$ may equivalently be considered a subset of $\mathbb{R}^n$ for some $n \in \mathbb{N}$, with the subspace topology (and Borel sets) inherited from $\mathbb{R}^n$.

We recall the definition of totally bounded measures.

Definition E.6. Let $\mathcal{M}$ be a σ-algebra on V, and let μ be a countably additive $(V, \mathcal{M})$-measure. Then, we define

$$|\mu|[E] = \sup \sum_{i=1}^{\infty} |\mu[E_i]| \tag{12}$$

where the supremum is taken over all countable partitions $\{E_i\}_{i \in \mathbb{N}}$, i.e., collections such that $\bigcup_{i=1}^{\infty} E_i = E$ and $E_i \cap E_j = \emptyset$ for $j \neq i$. We call $|\mu|$ the *total variation of* μ, and the *total variation norm of* μ is $|\mu|[X]$.

We say that μ *totally bounded* if its total variation norm is finite, i.e., $|\mu|[X] < \infty$.

Lemma E.5. *If μ is totally bounded, then $|\mu|$ is a finite positive measure on $(V, \mathcal{M})$, and $|\mu[E]| \leq |\mu|[E]$ for all $E \in \mathcal{M}$.*

See Theorem 6.2 in Rudin (1987) for proof.

We let $\mathbb{K}$ denote the set of totally bounded Borel measures on $\mathcal{K}$. We note that, in the case of path specific fairness, which involves the joint distributions of counterfactuals, X is not defined directly. Rather, the joint distribution of the counterfactuals $X_{\Pi,A,a'}$ and A defines the distribution of X through consistency, i.e., what would have happened to someone if their group membership were changed to $a' \in \mathcal{A}$ is what actually happens to them if their group membership *is* a'. More formally, $\Pr(X \in E \mid A = a') = \Pr(X_{\Pi,A,a'} \in E \mid A = a')$ for all Borel sets $E \subseteq \mathcal{X}$. (See §3.6.3 in Pearl (2009b).)

For any $\mu \in \mathbb{K}$, we adopt the following notational conventions: if we say that a property holds μ-a.s., then the subset of $\mathcal{K}$ on which the property fails has $|\mu|$-measure zero. If $E \subseteq \mathcal{K}$ is a measurable set, then we denote by $\mu \restriction_E$ the restriction of μ to E, i.e., the measure defined by the mapping $E' \mapsto \mu[E \cap E']$. We let $\mathbb{E}_\mu[f] = \int_{\mathcal{K}} f \, \mathrm{d}\mu$, and for measurable sets E, $\Pr_\mu(E) = \mu[E]$.[13] The fairness criteria we consider involve conditional independence relations. To make sense of conditional independence relations more generally, for Borel measurable f we define $\mathbb{E}_\mu[f \mid \mathcal{F}]$ to be the Radon-Nikodym derivative of the measure $E \mapsto \mathbb{E}_\mu[f \cdot \mathbb{1}_E]$ with respect to the measure μ restricted to the sub–σ-algebra of Borel sets $\mathcal{F}$. (See §34 in Billingsley (1995).) Similarly, we define $\mathbb{E}_\mu[f \mid g]$ to be $\mathbb{E}_\mu[f \mid \sigma(g)]$, where $\sigma(g)$ denotes the sub–σ-algebra of the Borel sets generated by g.

In cases where the condition can occur with non-zero probability, we can instead make use of the elementary definition of discrete conditional probability.

Lemma E.6. *Let g be a Borel function on $\mathcal{K}$, and suppose $\Pr_\mu(g = c) \neq 0$. Then, we have that μ-a.s., for any Borel function f,*

$$\mathbb{E}_\mu[f \mid g] \cdot \mathbb{1}_{g=c} = \frac{\mathbb{E}_\mu[f \cdot \mathbb{1}_{g=c}]}{\Pr_\mu(g = c)} \cdot \mathbb{1}_{g(x)=c}.$$

See Rao (2005) for proof.

With these notational conventions in place, we turn to establishing the background conditions of Prop. E.6.

Lemma E.7. *$\mathbb{K}$ is a Banach space with the metric $d(\mu, \mu') = |\mu - \mu'|[\mathcal{K}]$.*

See, e.g., Steele (2019) for proof.

Remark 3. Since $\mathbb{K}$ is a Banach space, it possesses a topology, and consequently a collection of Borel subsets. These Borel sets are to be distinguished from the Borel subsets of the underlying state space $\mathcal{K}$, which the elements of $\mathbb{K}$ measure. The requirement that the subset E of the convex set C be universally measurable in Proposition E.6 is in reference to the *Borel subsets of* $\mathbb{K}$; the requirement that $\mu \in \mathbb{K}$ be a Borel measure is in reference to the *Borel subsets of* $\mathcal{K}$.

Recall the definition of absolute continuity.

[13] To state and prove our results in a notationally uniform way, we occasionally write $\Pr_\mu(E)$ even when μ ranges over measures that may not be probability measures.

Definition E.7. Let μ and ν be measures on a measure space $(V, \mathcal{M})$. We say that a measure ν is *absolutely continuous with respect to* μ—also written $\nu \ll \mu$—if, whenever $\mu[E] = 0$, $\nu[E] = 0$.

Absolute continuity is a closed property in the topology induced by the total variation norm.

Lemma E.8. *Consider the space of totally bounded measures on a measure space $(V, \mathcal{M})$. The set of ν such that $\nu \ll \mu$ is closed.*

Proof. Let $\{\nu_i\}_{i \in \mathbb{N}}$ be a convergent sequence of measures absolutely continuous with respect to μ. Then, there exists some totally bounded ν such that $\nu_i \to \nu$. We seek to show that $\nu \ll \mu$. Let $E \in \mathcal{M}$ be an arbitrary set such that $\mu[E] = 0$. Then, we have that

$$\begin{aligned}\nu[E] &= \lim_{n \to \infty} \nu_i[E] \\ &= \lim_{n \to \infty} 0 \\ &= 0,\end{aligned}$$

since $\nu_i \ll \mu$ for all i. Since E was arbitrary, the result follows. □

Recall the definition of a pushforward measure.

Definition E.8. Let $f : (V, \mathcal{M}) \to (V', \mathcal{M}')$ be a measurable function. Let μ be a measure on V. We define the *pushforward measure* $\mu \circ f^{-1}$ to be $\mu \circ f^{-1}[E] = \mu[f^{-1}(E)]$ for all $E \in \mathcal{M}'$.

Within $\mathbb{K}$, we define the subspace $\mathbf{K}$ to be the set of totally bounded measures μ on $\mathcal{K}$ such that the pushforward measure $\mu \circ u^{-1}$—or, in the case of path-specific fairness, $\mu \circ (u^{\mathcal{A}})^{-1}$—is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}$ (resp., $\mathbb{R}^{\mathcal{A}}$). By the Radon-Nikodym theorem, these pushforward measures arise from densities, i.e., for any $\mu \in \mathbf{K}$, there exists a corresponding $f_\mu \in L^1(\mathbb{R})$ (resp., $f_\mu \in L^1(\mathbb{R}^{\mathcal{A}})$) such that for any measurable subset E of $\mathbb{R}$ (resp., $\mathbb{R}^{\mathcal{A}}$), we have

$$\mu \circ u^{-1}[E] = \int_E f_\mu \, \mathrm{d}\lambda.$$

We therefore see that $\mathbf{K}$ extends in a natural way with the notion of a $\mathcal{U}$- or $\mathcal{U}^{\mathcal{A}}$-fine distribution, and so, by a slight abuse of notation, refer to $\mathbf{K}$ as the set of *$\mathcal{U}$-fine measures on $\mathcal{K}$*.

Indeed, since $\Pr(u(X) \in E, A = a) \leq \Pr(u(X) \in E)$, it also follows that the conditional distributions of $u(X) \mid A = a$ are also absolutely continuous with respect to Lebesgue measure, and so also have densities. For notational convenience, we set $f_{\mu,a}$ to be the function satisfying

$$\Pr_\mu(u(X) \in E, A = a) = \int_E f_{\mu,a} \, \mathrm{d}\lambda,$$

so that $f_\mu = \sum_{a \in \mathcal{A}} f_{\mu,a}$.

Since absolute continuity is a closed condition, it follows that $\mathbf{K}$ is a closed subspace of $\mathbb{K}$. This leads to the following useful corollary of Lemma E.8.

Corollary E.2. *The collection of $\mathcal{U}$-fine measures on $\mathcal{K}$ is a Banach space.*

Proof. It is straightforward to see that $\mathbf{K}$ is a subspace of $\mathbb{K}$. Since $\mathbf{K}$ is a closed subset of $\mathbb{K}$ by Lemma E.8, it is complete, and therefore a Banach space. □

We note the following useful fact about elements of $\mathbf{K}$.

Lemma E.9. *Consider the mapping $\mu \mapsto f_\mu$ from $\mathbf{K}$ to $L^1(\mathbb{R})$ given by associating a measure μ with the Radon-Nikodym derivative of the pushforward measure $\mu \circ u^{-1}$. This mapping is continuous. Likewise, the mapping $\mu \mapsto f_{\mu,a}$ is continuous for all $a \in \mathcal{A}$, and, in the case of path-specific fairness, the mapping of μ to the Radon-Nikodym derivative of $\mu \circ (u^{\mathcal{A}})^{-1}$ is continuous.*

Proof. We show only the first case. The others follow by virtually identical arguments.

Let $\epsilon > 0$ be arbitrary. Choose $\mu \in \mathbf{K}$, and suppose that $|\mu - \mu'| < \epsilon$. Then, let

$$\begin{aligned}E^{\mathrm{Up}} &= \{x \in \mathbb{R} : f_\mu(x) > f_{\mu'}(x)\} \\ E^{\mathrm{Lo}} &= \{x \in \mathbb{R} : f_\mu(x) < f_{\mu'}(x)\}.\end{aligned}$$

Then E^{Up} and E^{Lo} are disjoint, so we have that

$$\begin{aligned}\|f_\mu - f_{\mu'}\|_{L^1(\mathbb{R})} &= \left| \int_{E^{\mathrm{Up}}} f_\mu - f_{\mu'} \, \mathrm{d}\lambda \right| \\ &\quad + \left| \int_{E^{\mathrm{Lo}}} f_\mu - f_{\mu'} \, \mathrm{d}\lambda \right| \\ &= |(\mu - \mu')[u^{-1}(E^{\mathrm{Up}}]| \\ &\quad + |(\mu - \mu')[u^{-1}(E^{\mathrm{Lo}})]| \\ &< \epsilon,\end{aligned}$$

where the second equality follows by the definition of pushforward measures and the inequality follows from Lemma E.5. Since ϵ was arbitrary, the claim follows. □

Finally, we define $\mathbf{Q}$. We let $\mathbf{Q}$ be the subset of $\mathbf{K}$ consisting of all $\mathcal{U}$-fine probability measures, i.e., measures $\mu \in \mathbb{K}$ such that:

1. The measure μ is $\mathcal{U}$-fine;
2. For all Borel sets $E \subseteq \mathcal{K}$, $\mu[E] \geq 0$;
3. The measure of the whole space is unity, i.e., $\mu[\mathcal{K}] = 1$.

We conclude the background and notation by observing that threshold policies are somewhat simpler for distributions in $\mathbf{K}$ and $\mathbf{Q}$ than for general distributions.

Lemma E.10. *For any $\mu \in \mathbf{K}$, if $\tau_0(x)$ and $\tau_1(x)$ are two multiple threshold policies with the same thresholds, then $\tau_0(X) = \tau_1(X)$ μ-a.s. Likewise, suppose that for $\mu \in \mathbf{Q}$*

$$\mathbb{E}_\mu[\tau_0(X) \mid A = a] = \mathbb{E}_\mu[\tau_1(X) \mid A = a]$$

for all $a \in \mathcal{A}$ such that $\Pr_\mu(A = a) > 0$. Then $\tau_0(X) = \tau_1(X)$ μ-a.s.

In the case of path-specific fairness, under the same conditions, we have in addition that $\tau_0(X_{\Pi,A,a}) = \tau_1(X_{\Pi,A,a})$ for any $a \in \mathcal{A}$.

Proof. First, we show that threshold policies with the same thresholds are equal, then we show that threshold policies that distribute positive decisions across groups in the same way are equal.

Let $\{t_a\}_{a\in\mathcal{A}}$ denote the shared set of thresholds. It follows that if $\tau_0(x) \neq \tau_1(x)$, then $u(x) = t_{\alpha(x)}$. Now,

$$\Pr(u(X) = t_a, A = a) = \int_{t_a}^{t_a} f_{\mu,a}\,\mathrm{d}\mu = 0,$$

so $\Pr_\mu(\tau_0(X) \neq \tau_1(X)) = 0$.

Next, suppose

$$\mathbb{E}_\mu[\tau_0(X) \mid A = a] = \mathbb{E}_\mu[\tau_1(X) \mid A = a].$$

If the thresholds of the two policies agree for all $a \in \mathcal{A}$ such that $\Pr_\mu(A = a) > 0$, then we are done by the previous paragraph. Therefore, suppose $t_a^0 \neq t_a^1$ for some suitable $a \in \mathcal{A}$, where t_a^i represents the threshold for group $a \in \mathcal{A}$ under the policy $\tau_i(x)$. Without loss of generality, suppose $t_a^0 < t_a^1$. Then, it follows that

$$\begin{aligned}\int_{t_a^0}^{t_a^1} f_{\mu,a}\,\mathrm{d}\lambda &= \mathbb{E}_\mu[\tau_1(X) \mid A = a] - \mathbb{E}_\mu[\tau_0(X) \mid A = a] \\ &= 0.\end{aligned}$$

Since $\mu \in \mathbf{Q}$, $\mu = |\mu|$, whence

$$\Pr_{|\mu|}(t_0^a \leq u(X) \leq t_a^1 \mid A = a) = 0.$$

Since this is true for all $a \in \mathcal{A}$ such that $\Pr_\mu(A = a) > 0$, $\tau_0(X) = \tau_1(X)$ μ-a.s.

The proof in the case of path-specific fairness is almost identical. □

E.3.2. Convexity, complete metrizability, and universal measurability

The set of regular $\mathcal{U}$-fine probability measures $\mathbf{Q}$ is the set to which we wish to apply Prop. E.6. To do so, we must show that $\mathbf{Q}$ is a convex and completely metrizable subset of $\mathbf{K}$.

Lemma E.11. *The set of regular probability measures $\mathbf{Q}$ is convex and completely metrizable.*

Proof. The proof proceeds in two pieces. First, we show that the $\mathcal{U}$-fine probability distributions are convex, as can be verified by direct calculation. Then, we show that $\mathbf{Q}$ is closed and therefore complete in the original metric of $\mathbf{K}$.

We begin by verifying convexity. Let $\mu, \mu' \in \mathbf{Q}$ and let $E \subseteq \mathcal{K}$ be an arbitrary Borel subset of $\mathcal{K}$. Then, choose $\theta \in [0, 1]$, and note that

$$\begin{aligned}(\theta \cdot \mu + [1-\theta] \cdot \mu')[E] &= \theta \cdot \mu[E] + [1-\theta] \cdot \mu'[E] \\ &\geq \theta \cdot 0 + [1-\theta] \cdot 0 \\ &= 0,\end{aligned}$$

and, likewise, that

$$\begin{aligned}(\theta \cdot \mu + [1-\theta] \cdot \mu')[\mathcal{K}] &= \theta \cdot \mu[\mathcal{K}] + [1-\theta] \cdot \mu'[\mathcal{K}] \\ &= \theta \cdot 1 + [1-\theta] \cdot 1 \\ &= 1.\end{aligned}$$

It remains only to show that $\mathbf{Q}$ is completely metrizable. To prove this, it suffices to show that it is closed, since closed subsets of complete spaces are complete, and $\mathbf{K}$ is a Banach space by Cor. E.2, and therefore complete.

Suppose $\{\mu_i\}_{i\in\mathbb{N}}$ is a convergent sequence of probability measures in $\mathbf{K}$ with limit μ. Then

$$\mu[E] = \lim_{i\to\infty} \mu_i[E] \geq \lim_{i\to\infty} 0 = 0$$

and

$$\mu[\mathcal{K}] = \lim_{i\to\infty} \mu_i[\mathcal{K}] = \lim_{i\to\infty} 1 = 1.$$

Therefore $\mathbf{Q}$ is closed, and therefore complete, and hence is a convex, completely metrizable subset of $\mathbf{K}$. □

Next we prove that the set $\mathbf{E}$ of regular $\mathcal{U}$-fine densities over which there exists a policy satisfying the relevant counterfactual fairness definition that is not strongly Pareto dominated is universally measurable.

Recall the definition of universal measurability.

Definition E.9. Let V be a complete topological space. Then $E \subseteq V$ is *universally measurable* if V is measurable by the completion of every finite Borel measure on V, i.e., if for every finite Borel measure μ, there exist Borel sets E' and S such that $E \triangle E' \subseteq S$ and $\mu[S] = 0$.

We note that if a set is Borel, it is by definition universally measurable. Moreover, if a set is open or closed, it is by definition Borel.

To show that $\mathbf{E}$ is closed, we show that any convergent sequence in $\mathbf{E}$ has a limit in $\mathbf{E}$. The technical complication of the argument stems from the following fact that satisfying the fairness conditions, e.g., Eq. (4), involves conditional expectations, about which very little can be said in the absence of a density, and which are difficult to compare when taken across distinct measures.

To handle these difficulties, we begin with a technical lemma, Lemma E.12, which gives a coarse bound on how different the conditional expectations of the same variable can be with respect to a sub–σ-algebra $\mathcal{F}$ over two different distributions, μ and μ', before applying the results to the proof of Lemma E.13.

Definition E.10. Let μ be a measure on a measure space $(V, \mathcal{M})$, and let f be μ-measurable. Consider the equivalence class of $\mathcal{M}$-measurable functions $C = \{g : g = f \ \mu\text{-a.e.}\}$.[14] We say that any $g \in C$ is a *version* of f, and that $g \in C$ is a *standard version* if $g(v) \leq C$ for some constant C and all $v \in V$.

Remark 4. It is straightforward to see that for $f \in L^\infty(\mu)$, a standard version always exists with $C = \|f\|_\infty$.

Remark 5. Note that in general, the conditional expectation $\mathbb{E}_{\mu'}[f \mid \mathcal{F}]$ is defined only μ'-a.e. If μ is not assumed to be absolutely continuous with respect to μ', it follows that

$$\|\mathbb{E}_{\mu}[f \mid \mathcal{F}] - \mathbb{E}_{\mu'}[f \mid \mathcal{F}]\|_{L^1(\mu)} \tag{13}$$

is not entirely well-defined, in that its value depends on what version of $\mathbb{E}_{\mu'}[f \mid \mathcal{F}]$ one chooses. For appropriate f, however, one can nevertheless bound Eq. (13) for any standard version of $\mathbb{E}_{\mu'}[f \mid \mathcal{F}]$.

Lemma E.12. *Let μ, μ' be totally bounded measures on a measure space $(V, \mathcal{M})$. Let $f \in L^\infty(\mu) \cap L^\infty(\mu')$. Let $\mathcal{F}$ be a sub–σ-algebra of $\mathcal{M}$. Let*

$$C = \max(\|f\|_{L^\infty(\mu)}, \|f\|_{L^\infty(\mu')}).$$

Then, if g is a standard version of $\mathbb{E}_{\mu'}[f \mid \mathcal{F}]$, we have that

$$\int_V |\mathbb{E}_{\mu}[f \mid \mathcal{F}] - g| \, \mathrm{d}\mu \leq 4C \cdot |\mu - \mu'|[V]. \tag{14}$$

Proof. First, we note that both $\mathbb{E}_{\mu}[f \mid \mathcal{F}]$ and g are $\mathcal{F}$-measurable. Therefore, the sets

$$E^{\mathrm{Up}} = \{v \in V : \mathbb{E}_{\mu}[f \mid \mathcal{F}](v) > g(v)\}$$

and

$$E^{\mathrm{Lo}} = \{v \in V : \mathbb{E}_{\mu}[f \mid \mathcal{F}](v) < g(v)\}$$

are in $\mathcal{F}$. Now, note that

$$\int_V |\mathbb{E}_{\mu}[f \mid \mathcal{F}] - g| \, \mathrm{d}\mu = \int_{E^{\mathrm{Up}}} \mathbb{E}_{\mu}[f \mid \mathcal{F}] - g \, \mathrm{d}\mu + \int_{E^{\mathrm{Lo}}} g - \mathbb{E}_{\mu}[f \mid \mathcal{F}] \, \mathrm{d}\mu.$$

First consider E^{Up}. Then, we have that

$$\begin{aligned}
&\int_{E^{\mathrm{Up}}} \mathbb{E}_{\mu}[f \mid \mathcal{F}] - g \, \mathrm{d}\mu \\
&\qquad = \int_{E^{\mathrm{Up}}} \mathbb{E}_{\mu}[f \mid \mathcal{F}] - g \, \mathrm{d}\mu \\
&\qquad\quad + \int_{E^{\mathrm{Up}}} g - g \, \mathrm{d}\mu' \\
&\qquad \leq \left| \int_{E^{\mathrm{Up}}} \mathbb{E}_{\mu}[f \mid \mathcal{F}] \, \mathrm{d}\mu - \int_{E^{\mathrm{Up}}} g \, \mathrm{d}\mu' \right| \\
&\qquad\quad + \int_{E^{up}} g \, \mathrm{d}|\mu - \mu'| \\
&\qquad \leq \left| \int_{E^{\mathrm{Up}}} f \, \mathrm{d}\mu - \int_{E^{\mathrm{Up}}} f \, \mathrm{d}\mu' \right| \\
&\qquad\quad + \int_{E^{up}} C \, \mathrm{d}|\mu - \mu'|,
\end{aligned}$$

where in the final inequality, we have used the fact that, since g is a standard version of $\mathbb{E}_{\mu'}[f \mid \mathcal{F}]$,

$$g(v) \leq \|\mathbb{E}_{\mu'}[f \mid \mathcal{F}]\|_{L^\infty(\mu')} \leq C$$

for all $v \in V$, and the fact that, by the definition of conditional expectation,

$$\int_E \mathbb{E}_{\nu}[h \mid \mathcal{F}] \, \mathrm{d}\nu = \int_E h \, \mathrm{d}\nu$$

for any $E \in \mathcal{F}$.

Since C is everywhere bounded by C, applying Lemma E.5 yields that this final expression is less than or equal to $2C \cdot |\mu - \mu'|[V]$. An identical argument shows that

$$\int_{E^{\mathrm{Lo}}} g - \mathbb{E}_{\mu}[f \mid \mathcal{F}] \, \mathrm{d}\mu \leq 2C \cdot |\mu - \mu'|[V],$$

whence the result follows. □

Lemma E.13. *Let $\mathbf{E} \subseteq \mathbf{Q}$ denote the set of joint densities on $\mathcal{K}$ such that there exists a policy satisfying the relevant fairness definition that is not strongly Pareto dominated. Then, $\mathbf{E}$ is closed, and therefore universally measurable.*

Proof. For simplicity, we consider the case of counterfactual equalized odds. The proofs in the other two cases are virtually identical.

Suppose $\mu_i \to \mu$ in $\mathbf{K}$, where $\{\mu_i\}_{i \in \mathbb{N}} \subseteq \mathbf{E}$. Then, by Lemma E.9, $f_{\mu_i, a} \to f_{\mu, a}$ in $L^1(\mathbb{R})$. Moreover, by

[14] Some authors define $L^p(\mu)$ spaces to consist of such equivalence classes, rather than the definition we use here.

Lemma D.2, there exists a sequence of threshold policies $\{\tau_i(x)\}_{i\in\mathbb{N}}$ such that both

$$\mathbb{E}_{\mu_i}[\tau(X)] = \min(b, \Pr_{\mu_i}(u(X) > 0))$$

and

$$\mathbb{E}_{\mu}[\tau_i(X) \mid A, Y(1)] = \mathbb{E}_{\mu}[\tau_i(X) \mid Y(1)].$$

Let $\{q_{a,i}\}_{a\in\mathcal{A}}$ be defined by

$$q_{a,i} = \mathbb{E}_{\mu_i}[\tau_i(X) \mid A = a]$$

if $\Pr_{\mu_i}(A = a) > 0$, and $q_{a,i} = 0$ otherwise.

Since $[0,1]^{\mathcal{A}}$ is compact, there exists a convergent subsequence $\{\{q_{a,n_i}\}_{a\in\mathcal{A}}\}_{i\in\mathbb{N}}$. Let it converge to the collection of quantiles $\{q_a\}_{a\in\mathcal{A}}$ defining, by Lemma D.3, a multiple threshold policy $\tau(x)$ over μ.

Because $\mu_i \to \mu$ and $\{q_{a,n_i}\}_{a\in\mathcal{A}} \to \{q_a\}_{a\in\mathcal{A}}$, we have that

$$\mathbb{E}_{\mu}[\tau_{a,n_i}(X) \mid A = a] \to \mathbb{E}_{\mu}[\tau(X) \mid A = a]$$

for all $a \in \mathcal{A}$ such that $\Pr_{\mu}(A = a) > 0$. Therefore, by Lemma E.9, $\tau_{n_i}(X) \to \tau(X)$ in $L^1(\mu)$.

Choose $\epsilon > 0$ arbitrarily. Then, choose N so large that for i greater than N,

$$|\mu - \mu_{n_i}|[\mathcal{K}] < \tfrac{\epsilon}{10}, \quad \|\tau(X) - \tau_{n_i}(X)\|_{L^1(\mu)} \leq \tfrac{\epsilon}{10}.$$

Then, observe that $\tau(x), \tau_i(x) \leq 1$, and recall that

$$\mathbb{E}_{\mu_{n_i}}[\tau_{n_i}(X) \mid A, Y(1)] = \mathbb{E}_{\mu_{n_i}}[\tau_{n_i}(X) \mid Y(1)]. \quad (15)$$

Therefore, let $g_i(x)$ be a standard version of $\mathbb{E}_{\mu_{n_i}}[\tau_{n_i}(X) \mid Y(1)]$ over μ_{n_i}. By Eq. (15), $g_i(x)$ is also a standard version of $\mathbb{E}_{\mu_{n_i}}[\tau_{n_i}(X) \mid A, Y(1)]$ over μ_{n_i}. Then, by Lemma E.12, we have that

$$\begin{aligned}
&\|\mathbb{E}_{\mu}[\tau(X) \mid A, Y(1)] - \mathbb{E}_{\mu_{n_i}}[\tau_{n_i}(X) \mid Y(1)]\|_{L^1(\mu)} \\
&\quad \leq \|\mathbb{E}_{\mu}[\tau(X) \mid A, Y(1)] \\
&\qquad\quad - \mathbb{E}_{\mu}[\tau_{n_i}(X) \mid A, Y(1)]\|_{L^1(\mu)} \\
&\qquad + \|\mathbb{E}_{\mu}[\tau_{n_i}(X) \mid A, Y(1)] - g_i(X)\|_{L^1(\mu)} \\
&\qquad + \|g_i(X) - \mathbb{E}_{\mu}[\tau_{n_i}(X) \mid Y(1)]\|_{L^1(\mu)}\|_{L^1(\mu)} \\
&\qquad + \|\mathbb{E}_{\mu}[\tau_{n_i}(X) \mid Y(1) - \mathbb{E}_{\mu}[\tau(X) \mid Y(1)]\|_{L^1(\mu)} \\
&\quad < \frac{\epsilon}{10} + \frac{4\epsilon}{10} + \frac{4\epsilon}{10} + \frac{\epsilon}{10}.
\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, it follows that, μ-a.e.,

$$\mathbb{E}_{\mu}[\tau(X) \mid A, Y(1)] = \mathbb{E}_{\mu}[\tau(X) \mid Y(1)].$$

Recall the standard fact that for independent random variables X and U,

$$\mathbb{E}[f(X,U) \mid X] = \int f(X,u)\,\mathrm{d}F_U(u),$$

where F_U is the distribution of U.[15] Further recall that $D = \mathbb{1}_{U_D \leq \tau(X)}$, where $U_D \perp\!\!\!\perp X, Y(1)$. It follows that

$$\Pr_{\mu}(D = 1 \mid X, Y(1)) = \int_0^1 \mathbb{1}_{u_d < \tau(X)}\,\mathrm{d}\lambda(u_d) = \tau(X).$$

Hence, by the law of iterated expectations,

$$\begin{aligned}
&\Pr_{\mu}(D = 1 \mid A, Y(1)) \\
&\qquad = \mathbb{E}_{\mu}[\Pr_{\mu}(D = 1 \mid X, Y(1)) \mid A, Y(1)] \\
&\qquad = \mathbb{E}_{\mu}[\tau(X) \mid A, Y(1)] \\
&\qquad = \mathbb{E}_{\mu}[\tau(X) \mid Y(1)] \\
&\qquad = \mathbb{E}_{\mu}[\Pr_{\mu}(D = 1 \mid X, Y(1)) \mid Y(1)] \\
&\qquad = \Pr_{\mu}(D = 1 \mid Y(1)).
\end{aligned}$$

Therefore $D \perp\!\!\!\perp A \mid Y(1)$ over μ, i.e., counterfactual equalized odds holds for the decision policy $\tau(x)$ over the distribution μ. Consequently $\mu \in \mathbf{E}$, and so $\mathbf{E}$ is closed and therefore universally measurable. □

E.4. Shy sets and probes

We require a number of additional technical lemmata for the proof of Theorem 1. The probe must be constructed carefully, so that, on the utility scale, an arbitrary element of μ is absolutely continuous with respect to a typical perturbation. In addition, it is useful to show that a number of properties are generic to simplify certain aspects of the proof of Theorem 1. For instance, Lemma E.16 is used in Theorem 1 that a certain conditional expectation is generically well-defined, avoiding the need to separately treat certain corner cases.

Cor. E.3 concerns the construction of the probe used in the proof of Theorem 1. Lemmata E.17 to E.20 use Cor. E.3 to provide additional simplifications to the proof of Theorem 1.

E.4.1. Maximal support

First, to construct the probe used in the proof of Theorem 1, we require an element $\mu \in \mathbf{K}$ such that f_μ has "maximal" support. To produce such an element, we use the following measure-theoretic construction.

Definition E.11. Let $\{E_\alpha\}_{\alpha\in\mathcal{I}}$ be an arbitrary collection of μ-measurable sets for some positive measure μ on a measure space $(M, \mathcal{M})$. We say that E is the *measure-theoretic union* of $\{E_\alpha\}_{\alpha\in\mathcal{I}}$ if $\mu[E - E_\alpha] = 0$ and $E = \bigcup_{i=1}^{\infty} E_{\alpha_i}$ for some countable subcollection $\{\alpha_i\} \subseteq \mathbb{N}$.

While measure-theoretic unions themselves are known in the literature (cf. Silva (2008), Rudin (1991)), their existence is a folk theorem, and so we include the following proof for completeness.

Lemma E.14. *Let μ be a finite measure. Then an arbitrary*

[15] For a proof of this fact see, e.g., Brozius (2019).

collection of μ-measurable sets has a measure-theoretic union.

Proof. For each countable subcollection $\mathcal{I}' \subseteq \mathcal{I}$, consider the "error term"

$$r(\mathcal{I}') = \sup_{\alpha \in \mathcal{I}} \mu \left[E_\alpha \setminus \bigcup_{\alpha' \in \mathcal{I}'} E_{\alpha'} \right]$$

We claim that the infimum of $r(\mathcal{I}')$ over all countable subcollections $\mathcal{I}' \subseteq \mathcal{I}$ must be zero.

For, toward a contradiction, suppose it were greater than or equal to $\epsilon > 0$. Choose any set E_{α_1} such that $\mu[E_{\alpha_1}] \geq \epsilon$. Such a set must exist, since otherwise $r(\emptyset) < \epsilon$. Choose E_{α_2} such that $\mu[E_{\alpha_2} \setminus E_{\alpha_1}] > \epsilon$. Again, some such set must exist, since otherwise $r(\{E_{\alpha_1}\}) < \epsilon$. Continuing in this way, we construct a countable collection $\{E_{\alpha_i}\}_{i \in \mathbb{N}}$. Therefore, we see that

$$\mu[M] \geq \mu \left[\bigcup_{i=1}^{n} E_{\alpha_i} \right] = \sum_{i=1}^{n} \mu \left[E_{\alpha_i} \setminus \bigcup_{j=1}^{i} E_{\alpha_j} \right].$$

By construction, every term in the final sum is greater than or equal to ϵ, contradicting the fact that $\mu[M] < \infty$. □

The construction of the "maximal" elements used to construct the probe in the proof of Theorem 1 follows as a corollary of Lemma E.14

Corollary E.3. *There are measures $\mu_{\max,a} \in \mathbf{Q}$ such that for every $a \in \mathcal{A}$ and any $\mu \in \mathbf{K}$,*

$$\lambda[\textsc{Supp}(f_{\mu',a}) \setminus \textsc{Supp}(f_{\mu_{\max},a})] = 0.$$

Proof. Consider the collection of measurable sets $\{\textsc{Supp}(f_\mu, a)\}_{\mu \in \mathbf{K}}$. By Lemma E.14, there exists a countable collection of measures $\{\mu_i\}_{i \in \mathbb{N}}$ such that for any $\mu \in \mathbf{K}$,

$$\lambda \left[\textsc{Supp}(f_{\mu,a}) \setminus \bigcup_{i=1}^{\infty} \textsc{Supp}(f_{\mu_i,a}) \right] = 0$$

By the hypothesis of Theorem 1, there exists $\mu \in \mathbf{Q}$ such that $\Pr_\mu(A = a) > 0$. Therefore, we can define the probability measure μ_a where

$$\mu_a[E] = \sum_{i=1}^{n} 2^{-i} \cdot \frac{|\mu_i \restriction_{A=a}|}{|\mu_i \restriction_{A=a}|[\mathcal{K}]}.$$

It follows immediately by construction that

$$\textsc{Supp}(f_{\mu_{\max},a}) = \bigcup_{i=1}^{\infty} \textsc{Supp}(f_{\mu_i,a}),$$

and that $\mu_{\max,a} \in \mathbf{Q}$. □

For notational simplicity, we refer to $\textsc{Supp}(f_{\mu_{\max,a}})$ as S_a and $\lambda \restriction_{S_a}$ throughout.

In the case of conditional principal fairness and path-specific fairness, we need a mild refinement of the previous result that accounts for ω.

Corollary E.4. *There are measures $\mu_{\max,a,w} \in \mathbf{Q}$ defined for every $w \in \mathcal{W} = \textsc{Img}(\omega)$ and any $a \in \mathcal{A}$ such that for some $\nu \in \mathbf{K}$, $\Pr_\nu(W = w, A = a) > 0$. These measures have the property that for any $\mu \in \mathbf{K}$,*

$$\lambda[\textsc{Supp}(f_{\mu',a,w}) \setminus \textsc{Supp}(f_{\mu_{\max},a,w})] = 0,$$

where $f_{\mu',a,w}$ is the density of the pushforward measure $(\mu' \restriction_{W=w,A=a}) \circ u^{-1}$.

Recalling that $|\textsc{Img}(\omega)| < \infty$, the proof is the same, and we analogously refer to $\textsc{Supp}(f_{\mu_{\max,a,w}})$ as $S_{a,w}$. Here, we have assumed without loss of generality—as we continue to assume in the sequel—that there is some $\mu \in \mathbf{K}$ such that $\Pr_\mu(W = w) > 0$ for all $w \in \mathcal{W}$.

Remark 6. Because their support is maximal, the hypotheses of Theorem 1 in addition to implying that $\mu_{\max,a}$ is well-defined for all $a \in \mathcal{A}$ also that that $\Pr_{\mu_{\max,a}}(u(X) > 0) > 0$. In the case of conditional principal fairness, they imply that $\Pr_{\mu_{\max,a}}(W = w) > 0$ for all $w \in \mathcal{W}$ and $a \in \mathcal{A}$. In the case of path-specific fairness, they imply that $\Pr_{\mu_{\max,a}}(W = w_i) > 0$ for $i = 0, 1$ and some $a \in \mathcal{A}$.

E.4.2. Shy sets and probes

In the following lemmata, we demonstrate that a number of useful properties are generic in $\mathbf{Q}$. We also demonstrate a short technical lemma, Lemma E.20, which allows us to use these generic properties to simplify the proof of Theorem 1.

We begin with the following lemma, which is useful in verifying that subspaces of $\mathbf{K}$ form probes.

Lemma E.15. *Let $\mathbf{W}$ be a non-trivial finite dimensional subspace of $\mathbf{K}$ such that $\nu[\mathcal{K}] = 0$ for all $\nu \in \mathbf{W}$. Then, there exists $\mu \in \mathbf{W}$ such that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$.*

Proof. Set

$$\mu = \sum_{i=1}^{n} \frac{|\mu_i|}{|\mu_i|[\mathcal{K}]},$$

where $\mu_1, \ldots, \mu_n$ form a basis of $\mathbf{W}$. Then, if $|\beta_i| < \frac{1}{|\mu_i|[\mathcal{K}]}$, it follows that

$$\mu + \sum_{i=1}^{n} \beta_i \cdot \mu_i \in \mathbf{Q}.$$

Since

$$\lambda_n \left[\prod_{i=1}^{n} \left(-\frac{1}{|\mu_i|[\mathcal{K}]}, \frac{1}{|\mu_i|[\mathcal{K}]} \right) \right] = \frac{2^n}{\prod_{i=1}^{n} |\mu_i|[\mathcal{K}]} > 0,$$

it follows that $\lambda_W[\mathbf{Q} - \mu] > 0$. □

Next we show that, given a $\nu \in \mathbf{Q}$, a generic element of $\mathbf{Q}$ "sees" events to which ν assigns non-zero probability. While Lemma E.18 alone in principle suffices for the proof of Theorem 1, we include Lemma E.16 both for conceptual clarity and to introduce at a high level the style of argument used in the subsequent Lemmata and in the proof of Theorem 1 to show that a set is shy relative to $\mathbf{Q}$.

Lemma E.16. *Suppose there exist disjoint Borel sets $E_0, E_1 \subseteq \mathcal{K}$ and $\nu \in \mathbf{Q}$ such that $\nu[E_i] > 0$ for $i = 0, 1$. Then the set of $\mu \in \mathbf{Q}$ such that $\mu[E_0] = 0$ or $\mu[E_1] = 0$ is shy.*

Proof. First, we note that the set of $\mu \in \mathbf{Q}$ such that $\mu[E_i] = 0$ for $i = 0$ or $i = 1$ is closed and therefore universally measurable. For, if $\{\mu_i\}_{i\in\mathbb{N}} \subseteq \mathbf{Q}$ is a convergent sequence with limit μ, then

$$\begin{aligned}\mu[E_i] &= \lim_{n\to\infty} \mu_i[E] \\ &= \lim_{n\to\infty} \\ &= 0.\end{aligned}$$

Next, consider the measure

$$\tilde{\nu}[E] = \nu[E_1] \cdot \nu[E \cap E_0] - \nu[E_0] \cdot \nu[E \cap E_1].$$

Then, let $\mathbf{W} = \textsc{Span}(\tilde{\nu})$. Since $\tilde{\nu} \neq 0$ and

$$\tilde{\nu}[\mathcal{K}] = \nu[E_1] \cdot \nu[E_0] - \nu[E_0] \cdot \nu[E_1] = 0,$$

it follows by Lemma E.15 that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$ for some μ.

Now, for arbitrary $\mu \in \mathbf{Q}$, note that if $(\mu + \beta \cdot \tilde{\nu})[E_0] = 0$, then

$$\mu[E_0] + \beta \cdot \nu[E_1] \cdot \nu[E_0] = 0,$$

i.e.,

$$\beta = -\frac{\mu[E_0]}{\nu[E_1] \cdot \nu[E_0]}.$$

A similar argument for E_1 shows that there at most two $\beta \in \mathbb{R}$ such that $(\mu + \beta \cdot \tilde{\nu})[E_i] = 0$ for some $i \in \{0, 1\}$, which is a $\lambda_{\mathbf{W}}$-null set. Therefore, by Prop. E.6, the set of $\mu \in \mathbf{Q}$ such that $\mu[E_0] = 0$ or $\mu[E_1] = 0$ is shy, as desired. □

While Lemma E.16 shows that a typical element of $\mathbf{Q}$ "sees" individual events, in the proof of Theorem 1, we require a stronger condition, namely, that a typical element of $\mathbf{Q}$ "sees" certain uncountable collections of events. To demonstrate this more complex property, we require the following technical result, which is closely related to the real analysis folk theorem that any convergent uncountable "sum" can contain only countably many non-zero terms. (See, e.g., Benji (2020).)

Lemma E.17. *Suppose μ is a totally bounded measure on $(V, \mathcal{M})$, f and g are μ-measurable real-valued functions, and $g \neq 0$ μ-a.e. Then the set*

$$Z_\beta = \{v \in V : f(v) + \beta \cdot g(v) = 0\}$$

has non-zero μ measure for at most countably many $\beta \in \mathbb{R}$.

Proof. First, we show that for any countable collection $\{\beta_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R}$, the sum $\sum_{i=1}^{\infty} \mu[Z_{\beta_i}]$ converges. Then, we show how this implies that $\mu[Z_\beta] = 0$ for all but countably many $\beta \in \mathbb{R}$.

First, we note that for distinct $\beta, \beta' \in \mathbb{R}$,

$$Z_\beta \cap Z_{\beta'} = \{v \in V : (\beta - \beta') \cdot g(v) = 0\}.$$

Now, by hypothesis,

$$\mu[\{v \in V : g(v) = 0\}] = 0,$$

and since $\beta - \beta' \neq 0$, it follows that

$$\mu[\{v \in V : (\beta - \beta') \cdot g(v) = 0\}] = 0$$

as well. Consequently, it follows that

$$\begin{aligned}\sum_{i=1}^{\infty} \mu[Z_{\beta_i}] = \mu\left[\bigcup_{i=1}^{\infty} Z_{\beta_i}\right] \\ \leq \mu[V] \\ < \infty.\end{aligned}$$

Therefore for any countable collection $\{\beta_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R}$, the sum $\sum_{i=1}^{\infty} \mu[Z_{\beta_i}]$ converges.

To see that this implies that $\mu[Z_\beta] > 0$ for only countably many $\beta \in \mathbb{R}$, let $G_\epsilon \subseteq \mathbb{R}$ consist of those β such that $\mu[Z_\beta] \geq \epsilon$. Then G_ϵ must be finite for all $\epsilon > 0$, since otherwise we could form a collection $\{\beta_i\}_{i\in\mathbb{N}} \subseteq G_\epsilon$, in which case

$$\sum_{i=1}^{\infty} \mu[Z_{\beta_i}] \geq \sum_{i=1}^{\infty} \epsilon = \infty,$$

contrary to what was just shown. Therefore G_ϵ is finite for all $\epsilon > 0$, and so

$$\{\beta \in \mathbb{R} : \mu[Z_\beta] > 0\} = \bigcup_{i=1}^{\infty} G_{1/i}$$

is countable. □

We now apply Lemma E.17 to the following lemma, which states, informally, that, under a generic element of $\mathbf{Q}$, $u(X)$ is supported everywhere it is supported under some particular fixed element of $\mathbf{Q}$. For instance, Lemma E.17 can be used to show that for a generic element of $\mathbf{Q}$, the density of $u(X) \mid A = a$ is positive $\lambda \restriction_{S_a}$-a.e.

Lemma E.18. *Let $\nu \in \mathbf{Q}$ and suppose ν is supported on E, i.e., $\nu[\mathcal{K} \setminus E] = 0$. Then the set of $\mu \in \mathbf{Q}$ such that $\nu \circ u^{-1} \ll (\mu \restriction_E) \circ u^{-1}$ is prevalent relative to $\mathbf{Q}$.*

Lemma E.18 states, informally, that for generic $\mu \in \mathbf{Q}$, $f_{\mu \restriction_E}$ is supported everywhere f_ν is supported.

Proof. We begin by showing that the set of $\mu \in \mathbf{Q}$ such that $\nu \circ u^{-1} \ll (\mu \restriction_E) \circ u^{-1}$ is Borel, and therefore universally measurable. Then, we construct a probe $\mathbf{W}$ and use it to show that this collection is finitely shy.

To begin, let U_q denote the set of $\mu \in \mathbf{Q}$ such that

$$\nu \circ u^{-1}[\{|f_{\mu \restriction_E}| = 0\}] < q.$$

We note that U_q is open. For, if $\mu \in U_q$, then there exists some $r > 0$ such that

$$\nu \circ u^{-1}[\{|f_{\mu \restriction_E}| < r\}] < q.$$

Let

$$\epsilon = q - \nu \circ u^{-1}[\{|f_{\mu \restriction_E}| < r\}].$$

Now, since $\nu \circ u^{-1} \ll \lambda$, there exists a δ such that if $\lambda[E'] < \delta$, then $\nu \circ u^{-1}[E'] < \epsilon$. Choose μ' arbitrarily so that $|\mu - \mu'|[\mathcal{K}] < \delta \cdot r$.

Then, by Markov's inequality, we have that

$$\lambda[\{|f_{\mu \restriction_E} - f_{\mu' \restriction_E}| > r\}] < \delta,$$

i.e.,

$$\nu \circ u^{-1}[\{f_{\mu \restriction_E} - f_{\mu' \restriction_E}| > r\}] < \epsilon.$$

Now, note that by the triangle inequality, wherever $|f_{\mu' \restriction_E}| = 0$, either $|f_{\mu \restriction_E}| < r$ or $|f_{\mu \restriction_E} - f_{\mu' \restriction_E}| > r$. Therefore

$$\begin{aligned}\lambda[\{|f_{\mu' \restriction_E}| = 0\}] &\leq \nu \circ u^{-1}[\{|f_{\mu \restriction_E} - f_{\mu' \restriction_E}| > r\}] \\ &\quad + \mu \circ u^{-1}[\{|f_{\mu \restriction_E}| < r] \\ &< \epsilon + \mu \circ u^{-1}[\{|f_{\mu \restriction_E}| < r] \\ &< q.\end{aligned}$$

We conclude that $\mu' \in U_q$, and so U_q is open.

Note that $\nu \circ u^{-1} \ll (\mu \restriction_E) \circ u^{-1}$ if and only if

$$\lambda[\textsc{Supp}(f_\nu) \setminus \textsc{Supp}(f_{\mu \restriction_E})] = 0$$

By the definition of the support of a function, $\lambda \restriction_{\textsc{Supp}(f_\mu)} \ll \mu \circ u^{-1}$. Therefore, it follows that

$$\lambda[\textsc{Supp}(f_\mu) \setminus \textsc{Supp}(f_{\nu \restriction_E})] = 0$$

if and only if

$$\mu \circ u^{-1}[\textsc{Supp}(f_\mu) \setminus \textsc{Supp}(f_{\nu \restriction_E})] = 0.$$

Then, it follows immediately that the set of $\nu \in \mathbf{Q}$ such that $\mu \circ u^{-1} \ll (\nu \restriction_E) \circ u^{-1}$ is $\bigcap_{i=1}^n U_{1/i}$, which is, by construction, Borel, and therefore universally measurable.

Now, since

$$\Pr_\nu(u(X) < t) = \int_{-\infty}^{t} f_\nu \, \mathrm{d}\lambda$$

is a continuous function of t, by the intermediate value theorem, there exists t such that $\Pr_\nu(u(X) \in S_0) = \Pr_\nu(u(X) \in S_1)$, where $S_0 = \textsc{Supp}(f_\nu) \cap (-\infty, t)$ and $S_1 = \textsc{Supp}(f_\nu) \cap [t, \infty)$. Then, we let

$$\tilde{\nu}[E'] = \int_{E'} \mathbb{1}_{u^{-1}(S_{a,0})} - \mathbb{1}_{u^{-1}(S_{a,1})} \, \mathrm{d}\nu.$$

Take $\mathbf{W} = \textsc{Span}(\tilde{\nu})$. Since $\tilde{\nu} \neq 0$ and $\tilde{\nu}[\mathcal{K}] = 0$, we have by Lemma E.15 that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$ for some μ.

By the definition of a density, $f_{\tilde{\nu}}$ is positive $(\tilde{\nu} \circ u^{-1})$-a.e. Consequently, by the definition of $\tilde{\nu}$, $f_{\tilde{\nu}}$ is positive $(\mu \circ u^{-1})$-a.e. Therefore, by Lemma E.17, there exist only countably many $\beta \in \mathbb{R}$ such that the density of $(\mu + \beta \cdot \tilde{\nu}) \circ u^{-1}$ equals zero on a set of positive $\mu \circ u^{-1}$-measure. Since countable sets have λ-measure zero and ν is arbitrary, the set of $\mu \in \mathbf{Q}$ such that $\nu \circ u^{-1} \ll (\mu \restriction_E) \circ u^{-1}$ is prevalent relative to $\mathbf{Q}$ by Prop. E.6. □

The following definition and technical lemma are needed to extend Theorem 1 to the cases of conditional principal fairness and path-specific fairness, which involve additional conditioning on $W = \omega(X)$. In particular, one corner case we wish to avoid in the proof of Theorem 1 is when the decision policy is non-trivial (i.e., some individuals receive a positive decision and others do not) but from the perspective of each ω-stratum, the policy is trivial (i.e., everyone in the stratum is accepted or rejected). Definition E.12 formalizes this pathology, and Lemma E.19 shows that this issue—under a mild hypothesis—does not arise for a generic element of $\mathbf{Q}$.

Definition E.12. We say that $\mu \in \mathbf{Q}$ *overlaps utilities* when, for any budget-exhausting multiple threshold policy $\tau(x)$, if

$$0 < \mathbb{E}_\mu[\tau(X)] < 1,$$

then there exists $w \in \mathcal{W}$ such that

$$0 < \mathbb{E}_\mu[\tau(X) \mid W = w] < 1.$$

If there exists a budget-exhausting multiple threshold policy $\tau(x)$ such that

$$0 < \mathbb{E}_\mu[\tau(X)] < 1,$$

but for all $w \in \mathcal{W}$,

$$\mathbb{E}_\mu[\tau(X) \mid W = w] \in \{0, 1\},$$

then we say that $\tau(x)$ *splits utilities* over μ.

Informally, having overlapped utilities prevents a budget-exhausting threshold policy from having thresholds that fall on the utility scale exactly between the strata induced by ω—i.e., a threshold policy that splits utilities. This is almost a generic condition in $\mathbf{Q}$, as we shown in Lemma E.19.

Lemma E.19. *Let $0 < b < 1$. Suppose that for all $w \in \mathcal{W}$ there exists $\mu \in \mathbf{Q}$ such that $\Pr_\mu(u(X) > 0, W = w) > 0$. Then almost every $\mu \in \mathbf{Q}$ overlaps utilities.*

Proof. Our goal is to show that the set $\mathbf{E}'$ of measures $\mu \in \mathbf{Q}$ such that there exists a splitting policy $\tau(x)$ is shy. To simplify the proof, we divide an conquer, showing that the set $\mathbf{E}_\mathcal{I}$ of measures $\mu \in \mathbf{Q}$ such that there exists a splitting policy where the thresholds fall below $w \in \mathcal{I} \subseteq \mathcal{W}$ and above $w \notin \mathcal{I}$ is Borel, before constructing a probe that shows that it is shy. Then, we argue that $\mathbf{E}' = \bigcup_{\mathcal{I} \subseteq \mathcal{W}} \mathbf{E}_\mathcal{I}$, which shows that $\mathbf{E}'$ is shy.

We begin by considering the linear map $\Phi : \mathbf{K} \to \mathbb{R} \times \mathbb{R}^\mathcal{W}$ given by

$$\Phi(\mu) = \left(\Pr_\mu(u(X) = 0), (\Pr_\mu(W = w))_{w \in \mathcal{W}}\right).$$

For any $\mathcal{I} \subseteq \mathcal{W}$, the sets

$$F^{\mathrm{Up}}_\mathcal{I} = \{x \in \mathbb{R} \times \mathbb{R}^\mathcal{W} : x_0 \geq b, b = \sum_{z \in \mathcal{I}} x_z\},$$
$$F^{\mathrm{Lo}}_\mathcal{I} = \{x \in \mathbb{R} \times \mathbb{R}^\mathcal{W} : x_0 \leq b, x_0 = \sum_{z \in \mathcal{I}} x_z\},$$

are closed by construction. Therefore, since Φ is continuous,

$$\mathbf{E}_\mathcal{I} = \mathbf{Q} \cap \Phi^{-1}\left(\bigcup_{\mathcal{I} \subseteq \mathcal{W}} F^{\mathrm{Up}}_\mathcal{I} \cup F^{\mathrm{Lo}}_\mathcal{I}\right) \tag{16}$$

is closed, and therefore universally measurable.

Note that by our hypothesis and Cor. E.4, for all $w \in \mathcal{W}$ there exists some $a_w \in \mathcal{A}$ such that

$$\Pr_{\mu_{\max, a_w, w}}(u(X) > 0).$$

We use this to show that $\mathbf{E}_\mathcal{I}$ is shy. Pick $w^* \in \mathcal{W}$ arbitrarily, and consider the measures ν_w for $w \neq w^*$ defined by

$$\nu_w = \frac{\mu_{\max, a_w, w} \restriction_{u(X) > 0}}{\Pr_{\mu_{\max, a_w, w}}(u(X) > 0)} - \frac{\mu_{\max, a_{w^*}, w^*} \restriction_{u(X) > 0}}{\Pr_{\mu_{\max, a_{w^*}, w^*}}(u(X) > 0)}.$$

We note that $\nu_w[\mathcal{K}] = 0$ by construction. Therefore, if $\mathbf{W}_w = \textsc{Span}(\nu_w)$, then $\lambda_{\mathbf{W}_w}[\mathbf{Q} - \mu_w] > 0$ for some μ_w by Lemma E.15.

Moreover, we have that $\Pr_\nu(u(X) > 0) = 0$ for all $\nu \in \mathbf{W}$, i.e.,

$$\Pr_\mu(u(X) > 0) = \Pr_{\mu + \nu}(u(X) > 0).$$

Now, since $b \neq 0, 1$, ω partitions $\mathcal{X}$ and has finite image, and μ is regular, it follows that

$$\mathbf{E}_\mathcal{W} = \mathbf{E}_\emptyset = \emptyset.$$

Since $\lambda_\mathbf{W}[\emptyset] = 0$ for any $\mathbf{W}$, we can assume without loss of generality that $\mathcal{I} \neq \mathcal{W}, \emptyset$.

In that case, there exists $w_\mathcal{I} \in \mathcal{W}$ such that if $w^* \in \mathcal{I}$, then $w_\mathcal{I} \notin \mathcal{I}$, and vice versa. Without loss of generality, assume $w_\mathcal{I} \in \mathcal{I}$ and $w^* \notin \mathcal{I}$. It then follows that for arbitrary $\mu \in \mathbf{Q}$,

$$\Phi(\mu + \beta \cdot \nu_{w_\mathcal{I}}) = \Phi(\mu) + \beta \cdot \mathbf{e}_{w_\mathcal{I}} - \beta \cdot \mathbf{e}_{w^*},$$

where $\mathbf{e}_w$ is the basis vector corresponding to $w \in \mathcal{W}$. From this, it follows immediately by Eq. (E.4.2) that

$$\mu + \beta \cdot \nu_{w_\mathcal{I}} \in \mathbf{E}_\mathcal{I}$$

only if

$$\beta = \min(b, \Pr_\mu(u(X) > 0)) - \sum_{w \in \mathcal{I}} \Pr_\mu(W = w).$$

This is a measure zero subset of $\mathbb{R}$, and so it follows that

$$\lambda_{\mathbf{W}_{w_\mathcal{I}}}[\mathbf{E}_\mathcal{I} - \mu] = 0$$

for all $\mu \in \mathbf{K}$. Therefore, by Prop. E.6, $\mathbf{E}_\mathcal{I}$ is shy in $\mathbf{Q}$. Taking the union over $\mathcal{I} \subseteq \mathcal{W}$, it follows by Prop. E.5 that $\bigcup_{\mathcal{I} \subseteq \mathcal{W}} \mathbf{E}_\mathcal{I}$ is shy.

Now, we must show that $\mathbf{E}' = \bigcup_{\mathcal{I} \subseteq \mathcal{W}} \mathbf{E}_\mathcal{I}$. By construction, $\mathbf{E}_\mathcal{I} \subseteq \mathbf{E}'$, since the policy $\tau(x) = \mathbb{1}_{\omega(x) \in \mathcal{I}}$ is budget-exhausting and separates utilities. To see the reverse inclusion, suppose $\mu \in \mathbf{E}'$, i.e., that there exists a budget-exhausting multiple threshold policy $\tau(x)$ that splits utilities over μ. Then, let

$$\mathcal{I}_\mu = \{w \in \mathcal{W} : \mathbb{E}_\mu[\tau(X) \mid W = w] = 1\}.$$

Since $\tau(x)$ is budget-exhausting, it follows immediately that $\mu \in \mathbf{E}_{\mathcal{I}_\mu}$. Therefore, $\mathbf{E}' = \bigcup_{\mathcal{I} \subseteq \mathcal{W}} \mathbf{E}_\mathcal{I}$, and so $\mathbf{E}'$ is shy, as desired. □

We conclude our discussion of shyness and shy sets with the following general lemma, which simplifies relative prevalence proofs by showing that one can, without loss of generality, restrict one's attention to the elements of the shy set itself in applying Prop. E.6.

Lemma E.20. *Suppose E is a universally measurable subset of a convex, completely-metrizable set C in a topological vector space V. If for some finite-dimensional subspace V' and all $v \in E$,*

$$\lambda_{V'}[\{v' \in V' : v + v' \in E\}] = 0, \tag{17}$$

then it follows that E is shy relative to C.

Proof. Let v be arbitrary. Then, either $(v + V') \cap E$ is empty or not.

First, suppose it is empty. Since $\lambda_{V'}[\emptyset] = 0$ by definition, it follows immediately that in this case $\lambda_{V'}[E - v] = 0$.

Next, suppose the intersection is not empty, and let $v + v^* \in E$ for some fixed $v^* \in V'$. It follows that

$$\begin{aligned}\lambda_{V'}[E - v] &= \lambda_{V'}[\{v' \in V' : v + v' \in E\}] \\ &= \lambda_{V'}[\{v' \in V' : (v + v^*) + v' \in E\}] \\ &= 0,\end{aligned}$$

where the first equality follows by definition; the second equality follows by the translation invariance of $\lambda_{V'}$, and the fact that $v^* + V' = V'$; and the final inequality follows from Eq. (17).

Therefore $\lambda_{V'}[E - v] = 0$ for arbitrary v, and so E is shy. □

E.5. Proof of Theorem 1

Using the lemmata above, we can prove Theorem 1. We briefly summarize what has been established so far:

- **Lemma E.7:** The set $\mathbf{K}$ of $\mathcal{U}$-fine distributions on $\mathcal{K}$ is a Banach space;
- **Lemma E.11:** The subset $\mathbf{Q}$ of $\mathcal{U}$-fine probability measures on $\mathcal{K}$ is a convex, completely metrizable subset of $\mathbf{K}$;
- **Lemma E.13:** The subset $\mathbf{E}$ of $\mathbf{Q}$ is a universally measurable subset of $\mathbf{K}$, where $\mathbf{E}$ is the set consisting of $\mathcal{U}$-fine probability measures over which there exists a policy satisfying counterfactual equalized odds (resp., conditional principal fairness, or path-specific fairness) that is not strongly Pareto dominated.

Therefore, to apply Prop. E.6, it follows that we simply need to construct a probe $\mathbf{W}$ and show that $\lambda_{\mathbf{W}}[\mathbf{Q} + \mu_0] > 0$ for some $\mu_0 \in \mathbf{K}$ but $\lambda_{\mathbf{W}}[\mathbf{E} + \mu] = 0$ for all $\mu \in \mathbf{K}$.

Proof. We divide the proof into three pieces. First, we illustrate how to construct the probe $\mathbf{W}$ from a particular collection of distributions $\{\nu_a^{\mathrm{Up}}, \nu_a^{\mathrm{Lo}}\}_{a \in \mathcal{A}}$. Second, we show that $\lambda_{\mathbf{W}}[\mathbf{E} - \mu] = 0$ for all $\mu \in \mathbf{K}$. For notational and expository simplicity, we focus in these first two sections on the case of counterfactual equalized odds. Therefore, in the third section, we show how to generalize the argument to conditional principal fairness and path-specific fairness.

Construction of the probe We will construct our probe to address two different cases. We recall that, by Cor. D.1, any policy that is not strongly Pareto dominated must be a budget-exhausting multiple threshold policy with non-negative thresholds. In the first case, we consider when the candidate budget-exhausting multiple threshold policy is $\mathbb{1}_{u(x)>0}$. By perturbing the underlying distribution by $\nu \in \mathbf{W}^{\mathrm{Lo}}$, we will be able to break the balance requirements implied by Eq. (2). In the second case, we treat the possibility that the candidate budget-exhausting multiple threshold policy has a non-trivial positive threshold. By perturbing the underlying distribution by $\nu \in \mathbf{W}^{\mathrm{Up}}$ for an alternative set of perturbations $\mathbf{W}^{\mathrm{Up}}$, we will again be able to break the balance requirements.

More specifically, to construct our probe $\mathbf{W} = \mathbf{W}^{\mathrm{Up}} + \mathbf{W}^{\mathrm{Lo}}$, we want $\mathbf{W}^{\mathrm{Up}}$ and $\mathbf{W}^{\mathrm{Lo}}$ to have a number of properties.

In particular, for all $\nu \in \mathbf{W}$, perturbation by ν should not affect whether the underlying distribution is a probability distribution, and should not affect how much of the budget is available to budget-exhausting policies. Specifically, for all $\nu \in \mathbf{W}$,

$$\int_{\mathcal{K}} 1 \, \mathrm{d}\nu = 0, \tag{18}$$

and

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)>0} \, \mathrm{d}\nu = 0. \tag{19}$$

In fact, the amount of budget available to budget-exhausting policies will not change within group, i.e., for all $a \in \mathcal{A}$ and $\nu \in \mathbf{W}$,

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)>0, A=a} \, \mathrm{d}\nu = 0. \tag{20}$$

Additionally, for some distinguished $y_0, y_1 \in \mathcal{Y}$, non-zero perturbations in $\nu^{\mathrm{Lo}} \in \mathbf{W}^{\mathrm{Lo}}$ should move mass between y_0 and y_1. That is, they should have the property that if $\Pr_{|\nu^{\mathrm{Lo}}|}(A = a) > 0$, then

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)<0, Y=y_i, A=a} \, \mathrm{d}\nu^{\mathrm{Lo}} \neq 0. \tag{21}$$

Finally, perturbations in $\mathbf{W}^{\mathrm{Up}}$ should have the property that for any non-trivial $t > 0$, some mass is moved either above or below $t > 0$. More precisely, for any $\mu \in \mathbf{Q}$ and any t such that

$$0 < \Pr_{\mu}(u(X) > t \mid A = a) < 1,$$

if $\nu^{\mathrm{Up}} \in \mathbf{W}^{\mathrm{Up}}$ is such that $\Pr_{|\nu^{\mathrm{Up}}|}(A = a) > 0$, then

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)>t, A=a} \, \mathrm{d}\nu^{\mathrm{Up}} \neq 0. \tag{22}$$

To carry out the construction, choose distinct $y_0, y_1 \in \mathcal{Y}$. Then, since

$$\mu_{\max,a} \circ u^{-1}[S_a \cap [0, r_a)] - \mu_{\max,a} \circ u^{-1}[S_a \cap [r_a, \infty)]$$

is a continuous function of r_a, it follows by the intermediate value theorem that we can partition S_a into three pieces,

$$\begin{aligned} S_a^{\mathrm{Lo}} &= S_a \cap (-\infty, 0), \\ S_{a,0}^{\mathrm{Up}} &= S_a \cap [0, r_a), \\ S_{a,1}^{\mathrm{Up}} &= S_a \cap [r_a, \infty), \end{aligned}$$

so that

$$\Pr_{\mu_{\max,a}}(u(X) \in S_{a,0}^{\mathrm{Up}}) = \Pr_{\mu_{\max,a}}(u(X) \in S_{a,1}^{\mathrm{Up}}).$$

Recall that $\mathcal{K} = \mathcal{X} \times \mathcal{Y}$. Let $\pi_{\mathcal{X}} : \mathcal{K} \to \mathcal{X}$ denote projection onto $\mathcal{X}$, and $\iota_y : \mathcal{X} \to \mathcal{K}$ be the injection $x \mapsto (x, y)$. We define

$$\begin{aligned} \nu_a^{\mathrm{Up}}[E] = \mu_{\max,a} \circ (\iota_{y_1} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_{a,1}^{\mathrm{Up}})], \\ - \mu_{\max,a} \circ (\iota_{y_1} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_{a,0}^{\mathrm{Up}})], \\ \nu_a^{\mathrm{Lo}}[E] = \mu_{\max,a} \circ (\iota_{y_1} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_a^{\mathrm{Lo}})] \\ - \mu_{\max,a} \circ (\iota_{y_0} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_a^{\mathrm{Lo}})]. \end{aligned}$$

By construction, ν_a^{Up} concentrates on

$$\{y_1\} \times u^{-1}(S_a \cap [0, \infty)),$$

while ν_a^{Lo} concentrates on

$$\{y_0, y_1\} \times u^{-1}(S_a \cap (-\infty, 0)).$$

Moreover, if we set

$$\begin{aligned} \mathbf{W}^{\mathrm{Up}} &= \textsc{Span}(\nu_a^{\mathrm{Up}})_{a \in \mathcal{A}}, \\ \mathbf{W}^{\mathrm{Lo}} &= \textsc{Span}(\nu_a^{\mathrm{Lo}})_{a \in \mathcal{A}}, \end{aligned}$$

then it is easy to see that Eqs. (18) to (21) will hold. The only non-trivial case is Eq. (22). However, by Cor. E.3, the support of $f_{\mu_{\max,a}}$ is maximal. That is, for $\mu \in \mathbf{Q}$, if

$$0 < \Pr_\mu(u(X) > t \mid A = a, u(X) > 0) < 1,$$

then it follows that $0 < t < \sup S_a$. Either $t \le r_a$ or $t > r_a$. Assume $t \le r_a$; then, it follows by the construction of ν_a^{Up} that

$$\begin{aligned} \nu_a^{\mathrm{Up}} \circ u^{-1}[(t, \infty)] &= \int_{r_a}^{\infty} f_{\max,a} \, \mathrm{d}\lambda \\ &\quad - \int_t^{r_a} f_{\max,a} \, \mathrm{d}\lambda \\ &> \int_{r_a}^{\infty} f_{\max,a} \, \mathrm{d}\lambda \\ &\quad - \int_0^{r_a} f_{\max,a} \, \mathrm{d}\lambda \\ &= 0. \end{aligned}$$

Similarly, if $t > r_a$,

$$\begin{aligned} \nu_a^{\mathrm{Up}} \circ u^{-1}[(t, \infty)] &= \int_t^{\infty} f_{\max,a} \, \mathrm{d}\lambda \\ &> \int_{\sup S_a}^{\infty} f_{\max,a} \, \mathrm{d}\lambda \\ &= 0. \end{aligned}$$

Therefore Eq. (22) holds.

Since $\mathbf{W}$ is non-trivial[16] and $\nu[\mathcal{K}] = 0$ for all $\nu \in \mathbf{W}$, it follows by Lemma E.15 that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$ for some $\mu \in \mathbf{K}$.

Shyness Recall that, by Prop. E.5, a set E is shy if and only if, for an arbitrary shy set E', $E \setminus E'$ is shy. By Lemma E.16, a generic element of $\mu \in \mathbf{Q}$ satisfies $\Pr_\mu(u(X) > 0, Y(1) = y_i, A = a) > 0$ for $i = 0, 1$, and $a \in \mathcal{A}$. Likewise, by Lemma E.18, a generic $\mu \in \mathbf{Q}$ satisfies $\nu_a^{\mathrm{Up}} \circ u^{-1} \ll (\mu \restriction_{\mathcal{X} \times \{y_i\}}) \circ u^{-1}$ for $i = 0, 1$. Therefore, to simplify our task and recalling Remark 6, we may instead to demonstrate the shyness of the set of $\mu \in \mathbf{Q}$ such that:

- There exists a budget-exhausting multiple policy $\tau(x)$ with non-negative thresholds satisfying counterfactual equalized odds over μ;
- For $i = 0, 1$,

$$\Pr_\mu(u(X) > 0, A = a, Y(1) = y_i) > 0; \tag{23}$$

- For $i, j = 0, 1$ and all $a \in \mathcal{A}$,

$$\nu_{a,i}^{\mathrm{Up}} \circ u^{-1} \ll (\mu \restriction_{\alpha^{-1}(a) \times \{y_j\}}) \circ u^{-1}. \tag{24}$$

By a slight abuse of notation, we continue to refer to this set as $\mathbf{E}$. We note that, by the construction of $\mathbf{W}$, Eq (23) is not affected by perturbation by $\nu \in \mathbf{W}$, and Eq. (24) is not affected by perturbation by $\nu^{\mathrm{Lo}} \in \mathbf{W}$.

In particular, by Lemma E.20, it suffices to show that $\lambda_{\mathbf{W}}[\mathbf{E} - \mu] = 0$ for $\mu \in \mathbf{E}$.

Therefore, let $\mu \in \mathbf{E}$ be arbitrary. Let the budget-exhausting multiple threshold policy satisfying counterfactual equalized odds over it be $\tau(x)$, so that

$$\mathbb{E}_\mu[\tau(X)] = \min(b, \Pr_\mu(u(X) > 0)),$$

with thresholds $\{t_a\}_{a \in \mathcal{A}}$. We split into two cases based on whether $\tau(X) = \mathbb{1}_{u(X)>0}$ μ-a.s. or not.

[16] In general, some or all of the ν^{Lo} may be zero depending on the λ-measure of S_a^{Lo}. However, as noted in Remark 6, the $\nu_{a,i}^{\mathrm{Up}}$ cannot be zero, since $\Pr_{\mu_{\max,a}}(u(X) > 0) > 0$ for all $a \in \mathcal{A}$. Therefore $\mathbf{W} \neq \{0\}$.

In both cases, we make use of the following two useful observations.

First, note that as $\mathbf{E} \subseteq \mathbf{Q}$, if $\mu + \nu$ is not a probability measure, then $\mu + \nu \notin \mathbf{E}$. Therefore, without loss of generality, we assume throughout that $\mu + \nu$ is a probability measure.

Second, suppose $\tau'(x)$ is a policy satisfying counterfactual equalized odds over some $\nu \in \mathbf{Q}$. Then, if $0 < \mathbb{E}_\mu[\tau'(X)] < 1$, it follows that for all $a \in \mathcal{A}$,

$$0 < \mathbb{E}_\mu[\tau'(X) \mid A = a] < 1. \tag{25}$$

For, suppose not. Then, without loss of generality, there must be $a_0, a_1 \in \mathcal{A}$ such that

$$\mathbb{E}_\mu[\tau'(X) \mid A = a_0] = 0$$

and

$$\mathbb{E}_\mu[\tau'(X) \mid A = a_1] > 0.$$

But then, by the law of iterated expectation, there must be some $\mathcal{Y}' \subseteq \mathcal{Y}$ such that $\mu[\mathcal{X} \times \mathcal{Y}'] > 0$ and, on $\mathcal{X} \times \mathcal{Y}'$,

$$\mathbb{E}_\mu[\tau'(X) \mid A = a_1, Y(1)] > 0 = \mathbb{E}_\mu[\tau'(X) \mid A = a_0, Y(1)],$$

contradicting the fact that $\tau'(x)$ satisfies counterfactual equalized odds over μ. Therefore, in what follows, we can assume that Eq. (25) holds.

Our goal is to show that $\lambda_{\mathbf{W}}[\mathbf{E} - \mu] = 0$.

Case 1 ($\tau(X) = \mathbb{1}_{u(X)>0}$). We argue as follows. First, we show that $\mathbb{1}_{u(X)>0}$ is the unique budget-exhausting multiple threshold policy with non-negative thresholds over $\mu + \nu$ for all $\nu \in \mathbf{W}$. Then, we show that the set of $\nu \in \mathbf{W}$ such that $\mathbb{1}_{u(x)>0}$ satisfies counterfactual equalized odds over $\mu + \nu$ is a $\lambda_{\mathbf{W}}$ null set.

We begin by observing that $\mathbf{W}^{\mathrm{Lo}} \neq \{0\}$. For, if that were the case, then Eq. (25) would not hold for $\tau(x)$.

Next, we note that by Eq. (19), for any $\nu \in \mathbf{W}$,

$$\Pr_{\mu+\nu}(u(X) > 0) = \Pr_\mu(u(X) > 0)$$

and so

$$\mathbb{E}_{\mu+\nu}[\mathbb{1}_{u(X)>0}] = \min(b, \Pr_{\mu+\nu}(u(X) > 0)).$$

If $\tau'(x)$ is a feasible multiple threshold policy with non-negative thresholds and $\tau'(X) \neq \mathbb{1}_{u(X)<0}$ $(\mu + \nu)$-a.s., then, as a consequence,

$$\mathbb{E}_{\mu+\nu}[\tau'(X)] < \Pr_{\mu+\nu}(u(X) > 0) \leq b.$$

Therefore, it follows that $\mathbb{1}_{u(X)>0}$ is the unique budget-exhausting multiple threshold policy over $\mu + \nu$ with non-negative thresholds.

Now, note that if counterfactual fairness holds with decision policy $\tau(x) = \mathbb{1}_{u(x)>0}$, then, by Eq. (4) and Lemma E.6, we must have that

$$\Pr_{\mu+\nu}(u(X) > 0 \mid A = a, Y(1) = y_1) = \Pr_{\mu+\nu}(u(X) > 0 \mid A = a', Y(1) = y_1)$$

for $a, a' \in \mathcal{A}$.[17]

Now, we will show that a typical element of $\mathbf{W}$ breaks this balance requirement. Choose a^* such that $\nu^{\mathrm{Lo}}_{a*} \neq 0$. Recall that ν is fixed, and let $\nu' = \nu - \beta^{\mathrm{Lo}}_{a^*} \cdot \nu^{\mathrm{Lo}}_{a^*}$. Let

$$p_a = \Pr_{\mu+\nu'}(u(X) > 0 \mid A = a', Y(1) = y_1).$$

Note that it cannot be the case that $p_a = 0$ for all $a \in \mathcal{A}$, since, by Eq. (23),

$$\Pr_{\mu+\nu'}(u(X) > 0 \mid Y(1) = y_1) > 0.$$

Therefore, by the foregoing discussion, either $p_{a^*} > 0$ or $p_{a^*} = 0$ and we can choose $a' \in \mathcal{A}$ such that $p_{a'} > 0$. Since the ν^{Lo}_a, $\nu^{\mathrm{Up}}_{a,i}$ are all mutually singular, it follows that counterfactual equalized odds can only hold over $\mu + \nu$ if

$$p_{a'} = \Pr_{\mu+\nu}(u(X) > 0 \mid A = a^*, Y(1) = y_1).$$

Now, we observe that by Lemma E.6, that

$$\Pr_{\mu+\nu}(u(X) > 0 \mid A = a^*, Y(1) = y_1) = \frac{\eta}{\pi + \beta^{\mathrm{Lo}}_{a^*} \cdot \rho}$$

where

$$\eta = \Pr_\mu(u(X) > 0, A = a^*, Y(1) = y_1)$$
$$\pi = \Pr_\mu(A = a^*, Y(1) = y_1),$$
$$\rho = \int_{\mathcal{K}} \mathbb{1}_{A=a^*, Y(1)=y_1} \, \mathrm{d}\nu^{\mathrm{Lo}}_a.$$

since

$$0 = \int_{\mathcal{K}} \mathbb{1}_{u(X)>0, A=a^*, Y(1)=y_1} \, \mathrm{d}\nu^{\mathrm{Lo}}_a,$$
$$0 \neq \int_{\mathcal{K}} \mathbb{1}_{A=a^*, Y(1)=y_1} \, \mathrm{d}\nu^{\mathrm{Lo}}_a.$$

Here, the equality follows by the fact that ν^{Lo} is supported on $S^{\mathrm{Lo}}_a \times \{y_0, y_1\}$.

Therefore, if, in the first case, $p_{a'} > 0$, then counterfactual equalized odds only holds if

$$\beta^{\mathrm{Lo}}_{a^*} = \frac{e - p_{a'} \cdot \pi}{p_{a'} \cdot \rho},$$

[17] To ensure that both quantities are well-defined, here and throughout the remainder of the proof we use the fact that by Eqs. (20) and (23), $\Pr_{\mu+\nu}(u(X) > 0, A = a, Y(1) = y_1) > 0$.

since $\rho \neq 0$ by Eq. (21). In the second case, if $p_{a'} = 0$, then counterfactual equalized odds can only hold if

$$e = p_{a^*} \cdot \pi = 0.$$

Since we chose a' so that $p_{a^*} > 0$ if $p_{a'} = 0$ and $\pi > 0$ by Eq. (23), this is impossible.

In either case, we see that the set of $\beta^{\text{Lo}}_{a^*} \in \mathbb{R}$ such that there a budget-exhausting threshold policy with positive thresholds satisfying counterfactual equalized odds over $\mu + \nu' + \beta^{\text{Lo}}_a \cdot \nu^{\text{Lo}}_a$ has λ-measure zero. That is

$$\lambda_{\text{SPAN}(\nu^{\text{Lo}}_{a^*})}[\mathbf{E} - \mu - \nu'] = 0.$$

Since ν' was arbitrary, it follows by Fubini's theorem that $\lambda_{\mathbf{W}}[\mathbf{E} - \mu] = 0$.

Case 2 ($\tau(X) \neq \mathbb{1}_{u(X)>0}$). Our proof strategy is similar to the previous case. First, we show that, for a given fixed $\nu^{\text{Lo}} \in \mathbf{W}^{\text{Up}}$, there is a unique candidate policy $\tilde{\tau}(x)$ for being a budget-exhausting multiple threshold policy with non-negative thresholds and satisfying counterfactual equalized odds over $\mu + \nu^{\text{Lo}} + \nu^{\text{Up}}$ for any $\nu^{\text{Up}} \in \mathbf{W}^{\text{Up}}$. Then, we show that the set of ν^{Up} such that $\tilde{\tau}(X)$ satisfies counterfactual equalized odds has $\lambda_{\mathbf{W}^{\text{Up}}}$ measure zero. Finally, we argue using that this in turn implies that the set of $\nu \in \mathbf{W}$ such that there exists a Pareto efficient policy satisfying counterfactual equalized odds over $\mu + \nu$ has $\lambda_{\mathbf{W}}$-measure zero.

We seek to show that $\lambda_{\mathbf{W}^{\text{Up}}}[\mathbf{E} - (\mu + \nu^{\text{Lo}})] = 0$. To begin, we note that since $\nu^{\text{Up}}_{a,i}$ concentrates on $\{y_1\} \times \mathcal{X}$ for all $a \in \mathcal{A}$, it follows that

$$\begin{aligned}\mathbb{E}_{\mu+\nu^{\text{Lo}}}[d(X) \mid A = a, Y(1) = y_0] \\ = \mathbb{E}_{\mu+\nu^{\text{Lo}}+\nu^{\text{Up}}}[d(X) \mid A = a, Y(1) = y_0]\end{aligned}$$

for any $\nu^{\text{Up}} \in \mathbf{W}^{\text{Up}}$.

Now, suppose there exists some $\nu^{\text{Up}} \in \mathbf{W}^{\text{Up}}$ such that there exists a budget-exhausting multiple threshold policy $\tilde{\tau}(x)$ with non-negative thresholds such that counterfactual equalized odds is satisfied over $\mu + \nu^{\text{Lo}} + \nu^{\text{Up}}$. (If not, then we are done and $\lambda_{\mathbf{W}^{\text{Up}}}[\mathbf{E} - (\mu + \nu^{\text{Lo}})] = 0$, as the measure of the empty set is zero.) Let

$$p = \mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}(X) \mid A = a, Y(1) = y_0].$$

Suppose that $\tilde{\tau}'(x)$ is an alternative budget-exhausting multiple threshold policy with non-negative thresholds such that counterfactual equalized odds is satisfied. We seek to show that $\tau'(X) = \tau(X)$ $(\mu + \nu^{\text{Lo}} + \nu^{\text{Up}})$-a.e. for any $\nu^{\text{Up}} \in \mathbf{W}^{\text{Up}}$. Toward a contradiction, suppose that for some $a_0 \in \mathcal{A}$,

$$\mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}'(X) \mid A = a_0, Y(1) = y_0] < p.$$

Since, by Eq. (23), $\Pr_{\mu+\nu^{\text{Lo}}}(A = a_0, Y(1) = y_0) > 0$, it follows that

$$\mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}'(X) \mid A = a_0] < \mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}(X) \mid A = a_0].$$

Therefore, since $\tilde{\tau}(x)'$ is budget exhausting, there must be some a_1 such that

$$\mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}'(X) \mid A = a_1] > \mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}(X) \mid A = a_1].$$

From this, it follows $\tilde{\tau}'(x)$ can be represented by a threshold greater than or equal to that of $\tilde{\tau}(x)$ on $\alpha^{-1}(a_1)$, and hence

$$\begin{aligned}\mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}'(X) \mid A = a_1, Y(1) = y_0] \\ \geq \mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}(X) \mid A = a_0, Y(1) = y_0] \\ = p \\ > \mathbb{E}_{\mu+\nu^{\text{Lo}}}[\tilde{\tau}'(X) \mid A = a_0, Y(1) = y_0],\end{aligned}$$

contradicting the fact that $\tilde{\tau}'(x)$ satisfies counterfactual equalized odds.

By the preceding discussion, Lemma E.10, and the fact that ν^{Lo} is supported on $u^{-1}((-\infty, 0])$,

$$\tilde{\tau}(X) = \tilde{\tau}'(X) \quad (\mu \restriction_{\mathcal{X} \times \{y_0\}})\text{-a.e.}$$

By Eq. (24), it follows that $\tilde{\tau}(X) = \tilde{\tau}'(X)$ $\nu^{\text{Up}}_{a,i}$-a.e. for $i = 0, 1$. As a consequence,

$$\tilde{\tau}(X) = \tilde{\tau}'(X) \quad (\mu + \nu^{\text{Lo}} + \nu^{up})\text{-a.e.}$$

for all $\nu^{\text{Up}} \in \mathbf{W}^{\text{Up}}$. Therefore $\tilde{\tau}(X)$ is, indeed, unique, as desired.

Now, we note that since $\tau(X) \neq \mathbb{1}_{u(X)>0}$, it follows that $\mathbb{E}[\tau(X)] < \Pr_\mu(u(X) > 0)$. It follows that $\mathbb{E}_\mu[\tau(X)] = b$, since $\tau(x)$ is budget exhausting. Therefore, by Eq. (19), it follows that for any budget-exhausting policy $\tilde{\tau}(X)$, $\mathbb{E}[\tilde{\tau}(X)] = b$, and so $\tilde{\tau}(X) \neq \mathbb{1}_{u(X)>0}$ over $\mu + \nu$.

Therefore, fix ν^{Lo} and $\tilde{\tau}(X)$. By Eq. (25), there is some a^* such that

$$0 < \Pr_{\mu+\nu^{\text{Lo}}}(u(X) > \tilde{t}_{a^*} \mid A = a^*) < 1.$$

Then, it follows by Eq. (22) that

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)>\tilde{t}_{a^*}} \, \mathrm{d}\nu^{\text{Up}}_{a^*} \neq 0.$$

Fix $\nu' = \nu - \beta^{\text{Up}}_{a^*} \cdot \nu^{\text{Up}}_{a^*}$. Then, for some $a \neq a^*$, set

$$p^* = \mathbb{E}_{\mu+\nu'}[\tilde{\tau}(X) \mid A = a, Y(1) = y_1].$$

Since the ν^{Lo}_a, ν^{Up}_a are all mutually singular, it follows that counterfactual equalized odds can only hold over $\mu + \nu$ if

$$p^* = \Pr_{\mu+\nu}(u(X) > \tilde{t}_{a^*} \mid A = a^*, Y(1) = y_1).$$

Now, we observe that by Lemma E.6, that

$$\begin{aligned}\Pr_{\mu+\nu}(u(X) > \tilde{t}_{a^*} \mid A = a^*, Y(1) = y_1) \\ = \frac{\eta + \beta_a^{\mathrm{Up}} \cdot \gamma}{\pi}\end{aligned} \tag{26}$$

where

$$\begin{aligned}\eta &= \Pr_{\mu+\nu^{\mathrm{Lo}}}(u(X) > \tilde{t}_{a^*} \mid A = a^*, Y(1) = y_1),\\ \pi &= \Pr_{\mu+\nu^{\mathrm{Lo}}}(A = a^*, Y(1) = y_1),\\ \gamma &= \int_{\mathcal{K}} \mathbb{1}_{u(X)>\tilde{t}_{a^*}, A=a^*, Y(1)=y_1} \,\mathrm{d}\nu_a^{\mathrm{Up}},\end{aligned}$$

and we note that

$$0 = \int_{\mathcal{K}} \mathbb{1}_{A=a^*, Y(1)=y_1} \,\mathrm{d}\nu_a^{\mathrm{Lo}}.$$

Eq. (26) can be rearranged to

$$(p^* \cdot \pi - \eta) - \beta \cdot \gamma = 0.$$

This can only hold if

$$\beta = \frac{p^* \cdot \pi - \eta}{\gamma},$$

since by Eq. (22), $\gamma \neq 0$. Since any countable subset of $\mathbb{R}$ is a λ-null set,

$$\lambda_{\textsc{Span}(\nu_{a^*}^{\mathrm{Up}})}[\mathbf{E} - \mu - \nu'] = 0.$$

Since ν' was arbitrary, it follows by Fubini's theorem that $\lambda_{\mathbf{W}^{\mathrm{Up}}}[\mathbf{E} - \mu - \nu^{\mathrm{Lo}}] = 0$ in this case as well. Lastly, since ν^{Lo} was also arbitrary, applying Fubini's theorem a final time gives that $\lambda_{\mathbf{W}}[\mathbf{E} - \mu] = 0$.

Conditional principal fairness and path-specific fairness The extension of these results to conditional principal fairness and path-specific fairness is straightforward. All that is required is a minor modification of the probe.

In the case of conditional principal fairness, we set

$$\begin{aligned}\nu_{a,w}^{\mathrm{Up}}[E] &= \mu_{\max,a,w} \circ (\iota_{(y_1,y_1)} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_{a,1}^{\mathrm{Up}})],\\ &\quad - \mu_{\max,a} \circ (\iota_{(y_1,y_1)} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_{a,w}^{\mathrm{Up}})],\\ \nu_{a,w}^{\mathrm{Lo}}[E] &= \mu_{\max,a,w} \circ (\iota_{(y_1,y_1)} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_a^{\mathrm{Lo}})]\\ &\quad - \mu_{\max,a} \circ (\iota_{(y_0,y_0)} \circ \pi_{\mathcal{X}})^{-1}[E \cap u^{-1}(S_{a,w}^{\mathrm{Lo}})],\end{aligned}$$

where $\iota_{(y,y')} : \mathcal{X} \to \mathcal{K}$ is the injection $x \mapsto (x, y, y')$. Our probe is then given by

$$\begin{aligned}\mathbf{W}^{\mathrm{Up}} &= \textsc{Span}(\nu_{a,w}^{\mathrm{Up}}),\\ \mathbf{W}^{\mathrm{Lo}} &= \textsc{Span}(\nu_{a,w}^{\mathrm{Lo}}),\end{aligned}$$

almost as before.

The proof otherwise proceeds virtually identically, except for two points. First, recalling Remark 6, we use $\Pr_\mu(A = a, W = w) > 0$ in place of the condition $\Pr_\mu(A = a) > 0$. Second, we use the fact that ω overlaps utility in place of Eq. (25). In particular, If ω does not overlap utilities for a generic $\mu \in \mathbf{Q}$, then, by Lemma E.19, there exists $w \in \mathcal{W}$ such that $\Pr_\mu(u(X) > 0, W = w) = 0$ for all $\mu \in \mathbf{Q}$. If this occurs, we can show that no budget-exhausting multiple threshold policy with positive thresholds satisfies conditional principal fairness, exactly as we did to show Eq. (25).

In the case of path-specific fairness, we instead define

$$\begin{aligned}S_{a,w}^{\mathrm{Lo}} &= S_{a,w} \cap (-\infty, r_{a,w}),\\ S_{a,w}^{\mathrm{Up}} &= S_{a,w} \cap [r_{a,w}, \infty),\end{aligned}$$

where $r_{a,w}$ is chosen so that

$$\Pr_{\mu_{\max,a,w}}(u(X) \in S_{a,w}^{\mathrm{Lo}}) = \Pr_{\mu_{\max,a,w}}(u(X) \in S_{a,w}^{\mathrm{Up}}).$$

Let π_X denote the projection from $\mathcal{K} = \mathcal{A} \times \mathcal{X}^{\mathcal{A}}$ given by

$$(a, (x_{a'})_{a' \in \mathcal{A}}) = x_a.$$

Let $\pi_{a'}$ denote the projection from the a'-th component. (That is, given $\mu \in \mathbf{K}$, the distribution of $X_{\Pi,A,a'}$ over μ is given by $\mu \circ \pi_{a'}^{-1}$ and the distribution of X is given by $\mu \circ \pi_X^{-1}$.) Then, we let $\tilde{\mu}_{\max,a,w}$ be the measure on $\mathcal{X}$ given by

$$\begin{aligned}\tilde{\mu}_{\max,a,w}[E] = \mu_{\max,a,w}[E \cap (u \circ \pi_a)^{-1}(s_{a,w}^{\mathrm{Up}})]\\ - \mu_{\max,a,w}[E \cap (u \circ \pi_a)^{-1}(S_{a,w}^{\mathrm{Lo}})].\end{aligned}$$

Finally, let $\phi : \mathcal{A} \to \mathcal{A}$ be a permutation of the groups with no fixed points, i.e., so that $a' \neq \phi(a')$ for all $a' \in \mathcal{A}$. Then, we define

$$\nu_{a'} = \delta_{a'} \times \tilde{\mu}_{\max,\phi(a'),w_1} \times \prod_{a \neq \phi(a')} \mu_{\max,a,w_1} \circ \pi_a^{-1},$$

where δ_a is the measure on $\mathcal{A}$ given by $\delta_a[\{a'\}] = \mathbb{1}_{a=a'}$. Then, simply let

$$\mathbf{W} = \textsc{Span}(\nu_a)_{a \in \mathcal{A}}.$$

Since $\tilde{\mu}_{\max,a,w}[\mathcal{X}] = 0$ for all $a \in \mathcal{A}$, it follows that $\nu_{a,w} \circ \pi_X = 0$, i.e.,

$$\Pr_\mu(X \in E) = \Pr_{\mu+\nu}(X \in E)$$

for any $\nu \in \mathbf{W}$ and $\mu \in \mathbf{Q}$. Therefore Eqs. (18) and (19) hold.

Moreover, the ν_a satisfy the following strengthening of Eq. (22). Perturbations in $\mathbf{W}$ have the property that for any non-trivial t—*not* necessarily positive—some of the

mass of $u(X_{\Pi,A,a})$ is moved either above or below t. More precisely, for any $\mu \in \mathbf{Q}$ and any t such that

$$0 < \Pr_\mu(u(X) > t \mid A = a) < 1,$$

if $\nu \in \mathbf{W}$ is such that $\Pr_{|\nu|}(A = \phi^{-1}(a)) > 0$, then

$$\int_{\mathcal{K}} \mathbb{1}_{u(X_{\Pi,A,a})>t} \, \mathrm{d}\nu_a \neq 0. \tag{27}$$

This stronger property means that we need not separately treat the case where $\tau(X) = \mathbb{1}_{u(X)>0}$ μ-a.e.

Other than this difference the proof proceeds in the same way, except for two points. First, we again make use of the fact that ω can be assumed to overlap utilities in place of Eq. (25), as in the case of conditional principal fairness. Second, w_0 and w_1 take the place of y_0 and y_1. In particular, to establish the uniqueness of $\tilde{\tau}(x)$ given μ and ν^{Lo} in the second case, instead of conditioning on y_0, we instead condition on w_0, where, following the discussion in Remark 6 and Lemma E.16, this conditioning is well-defined for a generic element of $\mathbf{Q}$. □

E.6. General measures on $\mathcal{K}$

Theorem 1 is restricted to $\mathcal{U}$-fine and $\mathcal{U}^{\mathcal{A}}$-fine distributions on the state space. The reason for this restriction is that when the distribution of X induces atoms on the utility scale, threshold policies will, in general, possess additional—or even infinite—degrees of freedom when the threshold falls exactly on an atom. In particular circumstances, these degrees of freedom can be used to prevent causal fairness notions, such as counterfactual equalized odds, from failing in a locally robust way. In particular, the generalization of Theorem 1 beyond $\mathcal{U}$-fine measures to all totally bounded measures on the state space is false, as illustrated by the following proposition.

Proposition E.7. *Consider the set $\mathbf{E}'$ of distributions—not necessarily $\mathcal{U}$-fine—on $\mathcal{K} = \mathcal{X} \times \mathcal{Y}$ over which there exists a Pareto efficient policy satisfying counterfactual equalized odds. There exist b, $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{U}$ such that $\mathbf{E}'$ is* not *shy.*

Proof. We adopt the notational conventions of Section E.3. We note that by Prop. E.5, a set can only be shy if it has empty interior. Therefore, we will construct an example in which an open ball of distributions on $\mathcal{K}$ in the total variation norm all allow for a Pareto efficient policy satisfying counterfactual equalized odds, i.e., are contained in $\mathbf{E}'$.

Let $b = \frac{3}{4}$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{A} = \{a_0, a_1\}$, and $\mathcal{X} = \{0, 1\} \times \{a_0, a_1\} \times \mathbb{R}$. Let $\alpha : \mathcal{X} \to \mathcal{A}$ be given by $\alpha : (y, a, v) \mapsto a$ for arbitrary $(y, a, v) \in \mathcal{X}$. Likewise, let $u : \mathcal{X} \to \mathbb{R}$ be given by $u : (y, a, v) \mapsto v$. Then, if $\mathcal{U} = \{u\}$, $\mathcal{U}$ is vacuously consistent modulo α.

Consider the joint distribution μ on $\mathcal{K} = \mathcal{X} \times \mathcal{Y}$ where for all $y, y' \in \mathcal{Y}$, $a \in \mathcal{A}$, and $u \in \mathbb{R}$,

$$\Pr_\mu(X = (a, y, u), Y(1) = y') = \frac{1}{4} \cdot \mathbb{1}_{y=y'} \cdot \Pr_\mu(u(X) = u),$$

where, over μ, $u(X)$ is distributed as a $\frac{1}{2}$-mixture of $\textsc{Unif}(1, 2)$ and $\delta(1)$; that is, $\Pr(u(X) = 1) = \frac{1}{2}$ and $\Pr(a < u(X) < b) = \frac{b-a}{2}$ for $0 \leq a \leq b \leq 1$.

We first observe that there exists a Pareto efficient threshold policy $\tau(x)$ such that counterfactual equalized odds is satisfied with respect to the decision policy $\tau(X)$. Namely, let

$$\tau(a, y, u) = \begin{cases} 1 & u > 1, \\ \frac{1}{2} & u = 1, \\ 0 & u < 1. \end{cases}$$

Then, it immediately follows that $\mathbb{E}[\tau(X)] = \frac{3}{4} = b$. $\tau(x)$ is a threshold policy and exhausts the budget, it is utility maximizing by Lemma D.4.

Moreover, if $D = \mathbb{1}_{U_D \leq \tau(x)}$ for some $U_D \sim \textsc{Unif}(0, 1)$ independent of X and $Y(1)$, then $D \perp\!\!\!\perp A \mid Y(1)$, where $D = \mathbb{1}_{U_D \leq \tau(x)}$. Moreover, since $u(X) \perp\!\!\!\perp A, Y(1)$, it follows that

$$\begin{aligned} \Pr_\mu(D = 1 \mid A = a, Y(1) = y) &= \Pr(U_D \leq \tau(X) \mid A = a, Y(1) = y) \\ &= \Pr(U_D \leq \tau(X)) \\ &= \mathbb{E}_\mu[\tau(X)], \end{aligned}$$

Therefore Eq. (2) is satisfied, i.e., counterfactual equalized odds holds.

Now, using μ, we construct an open ball of distributions over which we can construct similar threshold properties. In particular, suppose μ' is any distribution such that $|\mu - \mu'|[\mathcal{K}] < \frac{1}{64}$.

Then, we claim that there exists a budget-exhausting threshold policy satisfying counterfactual equalized odds over μ'. For, we note that

$$\begin{aligned} \Pr_{\mu'}(U > 1) &< \Pr_\mu(U > 1) + \frac{1}{64} = \frac{33}{64}, \\ \Pr_{\mu'}(U \geq 1) &> \Pr_\mu(U \geq 1) - \frac{1}{64} = \frac{63}{64}, \end{aligned}$$

and so any threshold policy $\tau'(x)$ satisfying $\mathbb{E}[\tau'(X)] = b = \frac{3}{4}$ must have $t = 1$ as its threshold.

We will now construct a threshold policy $\tau'(x)$ satisfying counterfactual equalized odds over μ'. Consider a threshold

policy of the form

$$\tau'(a,y,u) = \begin{cases} 1 & u > 1, \\ p_{a,y} & u = 1, \\ 0 & u < 1. \end{cases}$$

For notational simplicity, let

$$\begin{aligned} q_{a,y} &= \Pr_{\mu'}(A = a, Y = y, U > 1), \\ r_{a,y} &= \Pr_{\mu'}(A = a, Y = y, U = 1), \\ \pi_{a,y} &= \Pr_{\mu'}(A = a, Y = y). \end{aligned}$$

Then, we have that

$$\mathbb{E}_{\mu'}[\tau'(X)] = \sum_{a,y} q_{a,y} + p_{a,y} \cdot r_{a,y},$$

$$\mathbb{E}_{\mu'}[\tau'(X) \mid A = a, Y = y] = \frac{q_{a,y} + p_{a,y} \cdot r_{a,y}}{\pi_{a,y}}.$$

Therefore, the policy will be budget exhausting if

$$\sum_{a,y} q_{a,y} + p_{a,y} \cdot r_{a,y} = \tfrac{3}{4},$$

and it will satisfy counterfactual equalized odds if

$$\begin{aligned} &\pi_{a_1,0} \cdot (q_{a_0,0} + p_{a_0,0} \cdot r_{a_0,0}) \\ &\qquad = \pi_{a_0,0} \cdot (q_{a_1,0} + p_{a_1,0} \cdot r_{a_1,0}), \\ &\pi_{a_1,1} \cdot (q_{a_0,1} + p_{a_0,1} \cdot r_{a_0,1}) \\ &\qquad = \pi_{a_0,1} \cdot (q_{a_1,1} + p_{a_1,1} \cdot r_{a_1,1}), \end{aligned} \tag{28}$$

since, as above,

$$\begin{aligned} &\Pr(D = 1 \mid A = a, Y(1) = y) \\ &\qquad = \mathbb{E}[\tau'(X) \mid A = a, Y(1) = y). \end{aligned}$$

Again, for notational simplicity, let

$$S = \frac{\frac{3}{4} - \Pr_{\mu'}(U > 1)}{\Pr_{\mu'}(U = 1)}.$$

Then, a straightforward algebraic manipulation shows that Eq. (28) is solved by setting $p_{a_0,y}$ to be

$$\frac{S \cdot \pi_{a_0,y} \cdot (r_{a_0,y} + r_{a_1,y}) + \pi_{a_0,y} \cdot q_{a_1,y} - \pi_{a_1,y} \cdot q_{a_0,y}}{r_{a_0,y} \cdot (\pi_{a_0,y} + \pi_{a_1,y})},$$

and $p_{a_1,y}$ to be

$$\frac{S \cdot \pi_{a_1,y} \cdot (r_{a_0,y} + r_{a_1,y}) + \pi_{a_1,y} \cdot q_{a_0,y} - \pi_{a_0,y} \cdot q_{a_1,y}}{r_{a_1,y} \cdot (\pi_{a_0,y} + \pi_{a_1,y})}.$$

In order for $\tau'(x)$ to be a well-defined policy, we need to show that $p_{a,y} \in [0,1]$ for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$. To that end, note that

$$\begin{aligned} q_{a,y} &= \Pr_{\mu'}(A = a, Y = y, U > 1), \\ r_{a,y} &= \Pr_{\mu'}(A = a, Y = y, U = 1), \\ \pi_{a,y} &= \Pr_{\mu'}(A = a, Y = y), \\ r_{a_0,y} + r_{a_1,y} &= \Pr_{\mu'}(Y = y, U = 1), \\ \pi_{a_0,y} + \pi_{a_1,y} &= \Pr_{\mu'}(Y = y), \\ S &= \frac{\frac{3}{4} - \Pr_{\mu'}(U > 1)}{\Pr_{\mu'}(U = 1)}. \end{aligned}$$

Now, we recall that $|\Pr_{\mu'}(E) - \Pr_{\mu}(E)| < \frac{1}{64}$ for any event E by hypothesis. Therefore,

$$\begin{aligned} \frac{7}{64} \leq &\quad q_{a,y} \quad \leq \frac{9}{64}, \\ \frac{7}{64} \leq &\quad r_{a,y} \quad \leq \frac{9}{64}, \\ \frac{7}{64} \leq &\quad \pi_{a,y} \quad \leq \frac{17}{64}, \\ \frac{15}{64} \leq &\quad r_{a_0,y} + r_{a_1,y} \quad \leq \frac{17}{64}, \\ \frac{31}{64} \leq &\quad \pi_{a_0,y} + \pi_{a_1,y} \quad \leq \frac{33}{64}, \\ \frac{15}{31} \leq &\quad S \quad \leq \frac{17}{33}. \end{aligned}$$

Using these bounds and the expressions for $p_{a,y}$ derived above, we see that

$$\frac{629}{3069} < p_{a,y} < \frac{6497}{7161},$$

and hence $p_{a,y} \in [0,1]$ for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$.

Therefore, the policy $\tau'(x)$ is well-defined, and, by construction, is budget-exhausting and therefore utility-maximizing by Lemma D.4. It also satisfies counterfactual equalized odds by construction.

Since μ' was arbitrary, it follows that the set of distributions on $\mathcal{K}$ such that there exists a Pareto efficient policy satisfying counterfactual equalized odds contains an open ball, and hence is not shy. □

F. Theorem 2 and Related Results

We first prove a variant of Theorem 2 for general, continuous covariates $\mathcal{X}$. Then, we extend and generalize Theorem 2 using the theory of finite Markov chains, offering a proof of the theorem different from the sketch included in the main text.

F.1. Extension to Continuous Covariates

Here we follow the proof sketch in the main text for Theorem 2, which assumes a finite covariate-space $\mathcal{X}$. In that

case, we start with a point x^* with maximum decision probability $d(x^*)$, and then assume, toward a contradiction, that there exists a point with strictly lower decision probability. The general case is more involved since it is not immediately clear that the maximum value of $d(x)$ is achieved with positive probability in $\mathcal{X}$. We start with the lemma below before proving the main result.

Lemma F.21. *A decision policy $d(x)$ satisfies path-specific fairness with $W = X$ if and only if any $a' \in \mathcal{A}$,*

$$\mathbb{E}[d(X_{\Pi,A,a'}) \mid X] = d(X).$$

Proof. First, suppose that $d(x)$ satisfies path-specific fairness. To show the result, we use the standard fact that for independent random variables X and U,

$$\mathbb{E}[f(X,U) \mid X] = \int f(X,u)\,\mathrm{d}F_U(u), \tag{29}$$

where F_U is the distribution of U. (For a proof of this fact see, for example, Brozius, 2019)

Now, we have that

$$\begin{aligned}\mathbb{E}[D_{\Pi,A,a'} \mid X_{\Pi,A,a'}] &= \mathbb{E}[\mathbb{1}_{U_D \le d(X_{\Pi,A,a'})} \mid X_{\Pi,A,a'}] \\ &= \int_0^1 \mathbb{1}_{u \le d(X_{\Pi,A,a'})}\,\mathrm{d}u \\ &= d(X_{\Pi,A,a'}),\end{aligned}$$

where the first equality follows from the definition of $D_{\Pi,A,a'}$, and the second from Eq. (29), since the exogenous variable $U_D \sim \textsc{Unif}(0,1)$ is independent of the counterfactual covariates $X_{\Pi,A,a'}$. An analogous argument shows that $\mathbb{E}[D \mid X] = d(X)$.

Finally, conditioning on X, we have

$$\begin{aligned}\mathbb{E}[d(X_{\Pi,A,a'}) \mid X] &= \mathbb{E}[\mathbb{E}[D_{\Pi,A,a'} \mid X_{\Pi,A,a'}] \mid X] \\ &= \mathbb{E}[\mathbb{E}[D_{\Pi,A,a'} \mid X_{\Pi,A,a'}, X] \mid X] \\ &= \mathbb{E}[D_{\Pi,A,a'} \mid X] \\ &= \mathbb{E}[D \mid X] \\ &= d(X),\end{aligned}$$

where the second equality follows from the fact that $D_{\Pi,A,a'} \perp\!\!\!\perp X \mid X_{\Pi,A,a'}$, the third from the law of iterated expectations, and the fourth from the definition of path-specific fairness.

Next, suppose that

$$\mathbb{E}[d(X_{\Pi,A,a'} \mid X] = d(X)$$

for all $a' \in \mathcal{A}$. Then, since $W = X$ and $X \perp\!\!\!\perp U_D$, using Eq. (29), we have that for all $a' \in \mathcal{A}$,

$$\begin{aligned}\mathbb{E}[D_{\Pi,A,a'} \mid X] &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{U_D \le d(X_{\Pi,A,a'})} \mid X_{\Pi,A,a'}, X] \mid X] \\ &= \mathbb{E}[\mathbb{E}[d(X_{\Pi,A,a'}) \mid X_{\Pi,A,a'}, X] \mid X] \\ &= \mathbb{E}[d(X_{\Pi,A,a'}) \mid X] \\ &= d(X) \\ &= \mathbb{E}[d(X) \mid X] \\ &= \mathbb{E}[D \mid X].\end{aligned}$$

This is exactly Eq. (5), and so the result follows. □

We are now ready to prove a continuous variant of Theorem 2. The technical hypotheses of the theorem 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—for example, the conditions ensure that the conditional distribution of $X_{\Pi,A,a} \mid X$ does not have atoms that X itself does not have, and *vice versa*. For notational and conceptual simplicity, we only consider the case of trivial ζ, i.e., where $\zeta(x) = \zeta(x')$ for all $x, x' \in \mathcal{X}$.

Proposition F.8. *Suppose that*

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

Then, for $W = X$, any Π-fair policy $d(x)$ is constant a.s. (i.e., $d(X) = p$ a.s. for some $0 \le p \le 1$).

Proof. Let $d_{\max} = \|d(x)\|_\infty$, the essential supremum of d.

To establish the theorem statement, we show that $\Pr(d(X) = d_{\max} \mid A = a) = 1$ for all $a \in \mathcal{A}$. To do that, we begin by showing that there exists some $a \in \mathcal{A}$ such that $\Pr(d(X) = d_{\max} \mid A = a) > 0$.

Assume, toward a contradiction, that for all $a \in \mathcal{A}$,

$$\Pr(d(X) = d_{\max} \mid A = a) = 0. \tag{30}$$

Because $\mathcal{A}$ is finite, there must be some $a_0 \in \mathcal{A}$ such that

$$\Pr(d_{\max} - d(X) < \epsilon \mid A = a_0) > 0 \tag{31}$$

for all $\epsilon > 0$.

Choose $a_1 \neq a_0$. We show that for values of x such that $d(x)$ is close to $d_{\max}$, the distribution of $d(X_{\Pi,A,a_1}) \mid X =$

x must be concentrated near $d_{\max}$ with high probability to satisfy the definition of path-specific fairness, in Eq. (5). But, under the assumption in Eq. (30), we also show that the concentration occurs with low probability, by the continuity hypothesis in the statement of the theorem, establishing the contradiction.

Specifically, by Markov's inequality, for any $\rho > 0$, a.s.,

$$\begin{aligned}\Pr(d_{\max} - d(X_{\Pi,A,a_1}) \geq \rho \mid X) \\ \leq \frac{\mathbb{E}[d_{\max} - d(X_{\Pi,A,a_1}) \mid X]}{\rho} \\ = \frac{d_{\max} - d(X)}{\rho},\end{aligned}$$

where the final equality follows from Lemma F.21. Rearranging, it follows that for any $\rho > 0$, a.s.,

$$\begin{aligned}\Pr(d_{\max} - d(X_{\Pi,A,a_1}) < \rho \mid X) \\ \geq 1 - \frac{d_{\max} - d(X)}{\rho}.\end{aligned} \tag{32}$$

Now let $S = \{x \in \mathcal{X} : d_{\max} - d(x) < \rho\}$. By the second hypothesis of the theorem, we can choose δ sufficiently small that if

$$\Pr(X \in S \mid A = a_1) < \delta$$

then, a.s.,

$$\Pr(X_{\Pi,A,a_1} \in S, A \neq a_1 \mid X) < \tfrac{1}{2}.$$

In other words, we can chose δ such that if

$$\Pr(d_{\max} - d(X) < \rho \mid A = a_1) < \delta$$

then, a.s.,

$$\Pr(d_{\max} - d(X_{\Pi,A,a_1}) < \rho, A \neq a_1 \mid X) < \tfrac{1}{2}$$

By Eq. (30), we can choose $\epsilon > 0$ so small that

$$\Pr(d_{\max} - d(X) < \epsilon \mid A = a_1) < \delta.$$

Then, we have that

$$\Pr(d_{\max} - d(X_{\Pi,A,a_1}) < \epsilon, A \neq a_1 \mid X) < \tfrac{1}{2}$$

a.s. Further, by the definition of the essential supremum and a_0, and the fact that $a_0 \neq a_1$, we have that

$$\Pr(d_{\max} - d(X) < \tfrac{\epsilon}{2}, A \neq a_1) > 0.$$

Therefore, with positive probability, we have that

$$\begin{aligned}1 - \frac{d_{\max} - d(X)}{\epsilon} \\ > 1 - \frac{\frac{\epsilon}{2}}{\epsilon} \\ = \frac{1}{2} \\ > \Pr(d_{\max} - d(X_{\Pi,A,a_1}) < \epsilon, A \neq a_1 \mid X).\end{aligned}$$

This contradicts Eq. (32), and so it cannot be the case that $\Pr(d(X) = d_{\max} \mid A = a_0) = 0$, meaning $\Pr(d(X) = d_{\max} \mid A = a_0) > 0$.

Now, we show that $\Pr(d(X) = d_{\max} \mid A = a_1) = 1$. Suppose, toward a contradiction, that

$$\Pr(d(X) < d_{\max} \mid A = a_1) > 0.$$

Then, by the first hypothesis, a.s.,

$$\Pr(d(X_{\Pi,A,a_1}) < d_{\max} \vee A = a_1 \mid X) > 0$$

As a consequence,

$$\begin{aligned}d_{\max} &= \mathbb{E}[d(X) \mid d(X) = d_{\max}, A = a_0] \\ &= \mathbb{E}[\mathbb{E}[d(X_{\Pi,A,a_1}) \mid X] \mid d(X) = d_{\max}, A = a_0] \\ &< \mathbb{E}[\mathbb{E}[d_{\max} \mid X] \mid d(X) = d_{\max}, A = a_0] \\ &= \mathbb{E}[d_{\max} \mid d(X) = d_{\max}, A = a_0] \\ &= d_{\max},\end{aligned}$$

where we can condition on the set $\{d(X) = d_{\max}, A = a_0\}$ since $\Pr(d(X) = d_{\max} \mid A = a_0) > 0$; and the second equality above follows from Lemma F.21. This establishes the contradiction, and so $\Pr(d(X) = d_{\max} \mid A = a_1) = 1$.

Finally, we extend this equality to all $a \in \mathcal{A}$. Since, $\Pr(d(X) \neq d_{\max} \mid A = a_1) = 0$, we have, by the second hypothesis of the theorem, that, a.s.,

$$\Pr(d(X_{\Pi,A,a_1}) \neq d_{\max}, A \neq a_1 \mid X) = 0.$$

Since, by definition, $\Pr(X_{\Pi,A,a_1} = X \mid A = a_1) = 1$, and $\Pr(d(X) = d_{\max} \mid A = a_1) = 1$, we can strengthen this to

$$\Pr(d(X_{\Pi,A,a_1}) \neq d_{\max} \mid X) = 0.$$

Consequently, a.s.,

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

where the first equality follows from Lemma F.21, establishing the result. □

F.2. A Markov Chain Perspective

The theory of Markov chains illuminates—and allows us to extend—the proof of Theorem 2. Suppose $\mathcal{X} = \{x_1, \ldots, x_n\}$.[18] For any $a' \in \mathcal{A}$, let $P_{a'} = [p^{a'}_{i,j}]$ where

[18] Because of the technical difficulties associated with characterizing the long-run behavior of arbitrary infinite Markov chains, we restrict our attention in this section to Markov chains with finite state spaces.

$p^{a'}_{i,j} = \Pr(X_{\Pi,A,a'} = x_j \mid X = x_i)$. Then $P_{a'}$ is a stochastic matrix.

To motivate the subsequent discussion, we first note that this perspective conceptually simplifies some of our earlier results. Lemma F.21 can be recast as stating that when $W = X$, a policy d is Π-fair if and only if $P_{a'}d = d$—i.e., if and only if d is a 1-eigenvector of $P_{a'}$—for all $a' \in \mathcal{A}$.

The 1-eigenvectors of Markov chains have a particularly simple structure, which we derive here for completeness.

Lemma F.22. *Let $S_1, \ldots, S_m$ denote the recurrent classes of a finite Markov chain with transition matrix P. If d is a 1-eigenvector of P, then d takes a constant value p_k on each S_k, $k = 1, \ldots, m$, and*

$$d_i = \sum_{k=1}^{m} \left[\lim_{n\to\infty} \sum_{j \in S_k} P^n_{ij} \right] \cdot p_k. \tag{33}$$

Remark 7. We note that $\lim_{n\to\infty} \sum_{j\in S_k} P^n_{i,j}$ always exists and is the probability that the Markov chain, beginning at state i, is eventually absorbed by the recurrent class S_k.

Proof. Note that, possibly by reordering the states, we can arrange that the stochastic matrix P is in canonical form, i.e., that

$$P = \begin{bmatrix} B & \\ R' & Q \end{bmatrix},$$

where Q is a sub-stochastic matrix, R is non-negative, and

$$B = \begin{bmatrix} P_1 & & & \\ & P_2 & & \\ & & \ddots & \\ & & & P_m \end{bmatrix}$$

is a block-diagonal matrix with the stochastic matrix P_i corresponding to the transition probabilities on the recurrent set S_i in the i-th position along the diagonal.

Now, consider a 1-eigenvector $v = [v_1\ v_2]^\top$ of P. We must have that $Pv = v$, i.e., $Bv_1 = v_1$ and $R'v_1 + Qv_2 = v_2$. Therefore v_1 is a 1-eigenvector of B. Since B is block diagonal, and each diagonal element is a positive stochastic matrix, it follows by the Perron-Frobenius theorem that the 1-eigenvectors of B are given by $\textsc{Span}(\mathbf{1}_{S_i})_{i=1,\ldots,m}$, where $\mathbf{1}_{S_i}$ is the vector which is 1 at index j if $j \in S_i$ and is 0 otherwise.

Now, for $v_1 \in \textsc{Span}(\mathbf{1}_{S_i})_{i=1,\ldots,m}$, we must find v_2 such that $R'v_1 + Qv_2 = v_2$.

Note that every finite Markov chain M can be canonically associated with an absorbing Markov chain M^{ABS} where the set of states of M^{ABS} is exactly the union of the transitive states of M and the recurrent sets of M. (In essence, one tracks which state of M the Markov chain is in until it is absorbed by one of the recurrent sets, at which point the entire recurrent set is treated as a single absorbent state.) The transition matrix P^{ABS} associated with M^{ABS} is given by

$$P^{\text{ABS}} = \begin{bmatrix} I & \\ R & Q \end{bmatrix}$$

where $R = R'[\mathbf{1}_{S_1}\ \ldots\ \mathbf{1}_{S_m}]$. In particular, it follows that $v = [v_1\ v_2]^\top$ is a 1-eigenvector of P if and only if $[Tv_1\ v_2]^\top$ is a 1-eigenvector of P^{ABS}, where $T : \mathbf{1}_{S_i} \mapsto \mathbf{e}_i$.

Now, if v is a 1-eigenvector of P^{ABS}, then it is a 1-eigenvector of $(P^{\text{ABS}})^k$ for all k. Since Q is sub-stochastic, the series $\sum_{k=0}^{\infty} Q^k$ converges to $(I-Q)^{-1}$. Since

$$(P^{\text{ABS}})^k = \begin{bmatrix} I & \\ (I + Q + \cdots + Q^{k-1})R & Q^k \end{bmatrix},$$

it follows that

$$\lim_{k\to\infty} (P^{\text{ABS}})^k = \begin{bmatrix} I & \\ (I-Q)^{-1}R & 0 \end{bmatrix}.$$

Therefore, if $v = [v_1\ v_2]^\top$ is a 1-eigenvector of P^{ABS}, we must have that $(I-Q)^{-1}Rv_1 = v_2$. By Theorem 3.3.7 in Kemeny & Snell (1976), the (i,k) entry of $(I-Q)^{-1}R$ is exactly the probability that, conditional on $X_0 = x_i$, the Markov chain is eventually absorbed by the recurrent set S_k. This is, in turn, by the Chapman-Kolmogorov equations and the definition of S_k, equal to $\lim_{n\to\infty} \sum_{j\in S_k} P^n_{i,j}$, and therefore the result follows. □

We arrive at the following simple necessary condition on Π-fair policies.

Corollary F.5. *Suppose $\mathcal{X}$ is finite, and let $P = \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} P_{a'}$. If $d(x)$ is a Π-fair policy then it is constant on the recurrent classes of P.*

Proof. By Lemma F.21, d is Π-fair if and only if $P_{a'}d = d$ for all $a' \in \mathcal{A}$. Therefore,

$$\frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} P_{a'}d = \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} d = d, \tag{34}$$

and so d is a 1-eigenvector of P. Therefore it is constant on the recurrent classes of P by Lemma F.22. □

We note that Theorem 2 follows immediately from this.

Proof of Theorem 2. Note that $\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} P_a$ decomposes into a block diagonal stochastic matrix, where each block corresponds to a single stratum of ζ and is irreducible. Consequently, each stratum forms a recurrent class, and the result follows. □

G. Proof of Proposition 2

To prove the proposition, we must characterize the conditional tail risks of the beta distribution. Note that in the main text, we parameterize beta distributions in terms of their mean μ and sample size v; here, for mathematical simplicity, we parameterize them in terms of successes, α, and failures, β, where $\mu = \frac{\alpha}{\alpha+\beta}$ and $v = \alpha + \beta$.

Lemma G.23. *Suppose $Z_i \sim \text{BETA}(\alpha_i, \beta_i)$ for $i = 0, 1$, and that $\alpha_0 > \alpha_1 > 1$, $1 < \beta_0 < \beta_1$. Then, for all $t \in (0, 1]$, $\mathbb{E}[Z_1 \mid Z_1 < t] < \mathbb{E}[Z_0 \mid Z_0 < t]$.*

Proof. Let $Z(\alpha, \beta) \sim \text{BETA}(\alpha, \beta)$. Then,

$$\begin{aligned}\mathbb{E}[Z(\alpha,\beta) \mid Z(\alpha,\beta) < t] &= \frac{\int_0^t x \cdot \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}x}{\int_0^t \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}x} \\ &= \frac{\int_0^t x^{\alpha}(1-x)^{\beta-1}\,\mathrm{d}x}{\int_0^t x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x}.\end{aligned}$$

Since $\alpha > 1$, we may take the partial derivative with respect to α by differentiating under the integral sign, which yields that $\frac{\partial}{\partial\alpha}\mathbb{E}[Z(\alpha,\beta) \mid Z(\alpha,\beta) < t]$ equals

$$\frac{\alpha \cdot I(t,\alpha,\beta)^2 - [\alpha-1] \cdot I(t,\alpha+1,\beta) \cdot I(t,\alpha-1,\beta)}{I(t,\alpha,\beta)^2},$$

where $I(x,\alpha,\beta) = \int_0^t x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x$. Rearranging gives that this is greater than zero when

$$\begin{aligned}0 < \alpha \cdot I(t,\alpha+1,\beta) \cdot \int_0^t (x^{\alpha-2} - x^{\alpha-1})(1-x)^{\beta}\,\mathrm{d}x \\ + \alpha \cdot I(t,\alpha,\beta) \cdot \int_0^t (x^{\alpha-1} - x^{\alpha})(1-x)^{\beta}\,\mathrm{d}x \\ + I(t,\alpha+1,\beta) \cdot I(t,\alpha-1,\beta).\end{aligned}$$

Since all of the integrands are positive, $\frac{\partial}{\partial\alpha}\mathbb{E}[Z(\alpha,\beta) \mid Z(\alpha,\beta) < t] > 0$.

A virtually identical argument shows that $\frac{\partial}{\partial\beta}\mathbb{E}[Z(\alpha,\beta) \mid Z(\alpha,\beta) < t] < 0$. Therefore, the result follows. □

We use this lemma to prove a modest generalization of Prop. 2.

Lemma G.24. *Suppose $\mathcal{A} = \{a_0, a_1\}$, and consider the family $\mathcal{U}$ of utility functions of the form*

$$u(x) = r(x) + \lambda \cdot \mathbb{1}_{\alpha(x)=a_1},$$

indexed by $\lambda \geq 0$, where $r(x) = \mathbb{E}[Y(1) \mid X = x]$. Suppose the conditional distributions of $r(X)$ given A are beta distributed, i.e.,

$$\mathcal{D}(r(X) \mid A = a) = \text{BETA}(\alpha_a, \beta_a),$$

with $1 < \alpha_{a_1} < \alpha_{a_0}$ and $1 < \beta_{a_0} < \beta_{a_1}$. Then any policy satisfying counterfactual predictive parity is strongly Pareto dominated.

Proof. Suppose there were a Pareto efficient policy satisfying counterfactual predictive parity. Let $\lambda = 0$. Then, by Prop. 1, we may without loss of generality assume that there exist thresholds t_{a_0}, t_{a_1} and a constant p such that a threshold policy $\tau(x)$ witnessing Pareto efficiency is given by

$$\tau(x) = \begin{cases} 1 & r(x) > t_{\alpha(x)}, \\ 0 & r(x) < t_{\alpha(x)}. \end{cases}$$

(Note that by our distributional assumption, $\Pr(u(x) = t) = 0$ for all $t \in [0, 1]$.) Since $\lambda \geq 0$, we must have that $t_{a_0} \geq t_{a_1}$. Since $b < 1$, $0 < t_{a_0}$. Therefore,

$$\begin{aligned}&\mathbb{E}[Y(1) \mid A = a_0, D = 0] \\ &\qquad = \mathbb{E}[r(X) \mid A = a_0, u(X) < t_{a_0}] \\ &\qquad \geq \mathbb{E}[r(X) \mid A = a_0, u(X) < t_{a_1}] \\ &\qquad > \mathbb{E}[r(X) \mid A = a_1, u(X) < t_{a_1}] \\ &\qquad = \mathbb{E}[Y(1) \mid A = a_1, D = 0],\end{aligned}$$

where the first equality follows by the law of iterated expectation, the second from the fact that $t_{a_1} \leq t_{a_0}$, the third from our distributional assumption and Lemma G.23, and the final again from the law of iterated expectation. However, since counterfactual predictive parity is satisfied, $\mathbb{E}[Y(1) \mid A = a_0, D = 0] = \mathbb{E}[Y(1) \mid A = a_1, D = 0]$, which is a contradiction. Therefore, no such threshold policy exists. □

After accounting for the difference in parameterization, Prop. 2 follows as a corollary.

Proof of Prop. 2. Note that $\alpha_a = \mu_a \cdot v_a > v \cdot \frac{1}{v} = 1$ and $\beta_a = v - \alpha_a = v \cdot (1 - \mu_a) > v \cdot (1 - \frac{1}{v}) > 2 - 1 = 1$. Moreover, since $\mu_{a_0} > \mu_{a_1}$, $\alpha_{a_0} = v \cdot \mu_{a_0} > v \cdot \mu_{a_1} = \alpha_{a_1}$ and $\beta_{a_0} = v \cdot (1 - \mu_{a_0}) < v \cdot (1 - \mu_{a_1}) = \beta_{a_1}$. Therefore $1 < \alpha_{a_0} < \alpha_{a_1}$ and $1 < \alpha_{a_1} < \alpha_{a_0}$, and so, by Lemma G.24, the proposition follows. □