Journal of Machine Learning Research 24 (2023) 1-117 Submitted 12/22; Revised 8/23; Published 8/23

The Measure and Mismeasure of Fairness

Sam Corbett-Davies* SAMCORBETTDAVIES@GMAIL.COM

Johann D. Gaebler* JGAEBLER@FAS.HARVARD.EDU
Department of Statistics
Harvard University
Cambridge, MA 02138, USA

Hamed Nilforoshan* HAMEDN@CS.STANFORD.EDU
Department of Computer Science
Stanford University
Stanford, CA 94305, USA

Ravi Shroff RAVI.SHROFF@NYU.EDU
Department of Applied Statistics, Social Science, and Humanities
New York University
New York, NY 10003, USA

Sharad Goel SGOEL@HKS.HARVARD.EDU
Harvard Kennedy School
Harvard University
Cambridge, MA 02138, USA

* *Authors contributed equally.*

Editor: Kilian Weinberger

Abstract

The field of fair machine learning aims to ensure that decisions guided by algorithms are equitable. Over the last decade, several formal, mathematical definitions of fairness have gained prominence. Here we first assemble and categorize these definitions into two broad families: (1) those that constrain the effects of decisions on disparities; and (2) those that constrain the effects of legally protected characteristics, like race and gender, on decisions. We then show, analytically and empirically, that both families of definitions typically result in strongly Pareto dominated decision policies. For example, in the case of college admissions, adhering to popular formal conceptions of fairness would simultaneously result in lower student-body diversity and a less academically prepared class, relative to what one could achieve by explicitly tailoring admissions policies to achieve desired outcomes. In this sense, requiring that these fairness definitions hold can, perversely, harm the very groups they were designed to protect. In contrast to axiomatic notions of fairness, we argue that the equitable design of algorithms requires grappling with their context-specific consequences, akin to the equitable design of policy. We conclude by listing several open challenges in fair machine learning and offering strategies to ensure algorithms are better aligned with policy goals.

Keywords: fair machine learning, consequentialism, discrimination

Contents

1 Introduction **4**

2 Mathematical Definitions of Fairness **6**
2.1 Formal Setting 6
2.2 Limiting the Effect of Decisions on Disparities 8
2.3 Limiting the Effect of Attributes on Decisions 9

3 Equitable Decisions in the Absence of Externalities **13**
3.1 Utility, Risk, and Threshold Rules 14
3.2 The Problem of Inframarginality 15
3.3 The Problem with Fairness through Unawareness 18

4 Equitable Decisions in the Presence of Externalities **23**
4.1 The Geometry of Fair Decision Making 23
4.2 A Formal Theory of Fairness in the Presence of Externalities 26

5 A Path Forward **30**
5.1 Competing Notions of Ethical Decision Making: Process vs. Outcomes . . . 31
5.2 Designing Equitable Algorithms 33
5.3 Case Study: An Algorithm to Allocate Limited Medical Resources 38

6 Conclusion **43**

Appendices **44**

A Path-specific Counterfactuals **44**

B Constructing Causally Fair Policies **46**

C A Stylized Example of College Admissions **49**

D Proof of Theorem 10 **49**

E Proof of Proposition 15 **51**

F Prevalence and the Proof of Theorem 17 **56**
F.1 Shyness and Prevalence 56
F.2 Outline 58
F.3 Convexity, Complete Metrizability, and Universal Measurability 60
F.4 Shy Sets and Probes 69
F.5 Proof of Theorem 17 78
F.6 Proof of Theorem 9 88
F.7 Proof of Corollary 18 90
F.8 General Measures on $\mathcal{K}$ 91

G Theorem 11 and Related Results **94**
G.1 Extension to Continuous Covariates 95
G.2 A Markov Chain Perspective 98

H Proofs of Propositions 19 and 20 **100**
H.1 Beta Distributions and Stochastic Dominance 100
H.2 Proof of Proposition 19 103
H.3 Proof of Proposition 20 106

1. Introduction

In banking, criminal justice, medicine, and beyond, consequential decisions are often informed by machine learning algorithms (Barocas and Selbst, 2016; Berk, 2012; Chouldechova et al., 2018; Shroff, 2017). As the influence and scope of algorithms increase, academics, policymakers, and journalists have raised concerns that these tools might inadvertently encode and entrench human biases. Such concerns have sparked tremendous interest in developing *fair* machine-learning algorithms, and, accordingly, a plethora of formal fairness criteria have been proposed in the computer science community (Berk et al., 2021; Carey and Wu, 2022; Chiappa, 2019; Chouldechova, 2017; Chouldechova and Roth, 2020; Cleary, 1968; Corbett-Davies et al., 2017; Coston et al., 2020; Darlington, 1971; Dwork et al., 2012; Galhotra et al., 2022; Hardt et al., 2016; Imai and Jiang, 2020; Imai et al., 2020; Kilbertus et al., 2017; Kleinberg et al., 2017b; Kusner et al., 2017; Loftus et al., 2018; Mhasawade and Chunara, 2021; Nabi and Shpitser, 2018; Wang et al., 2019; Woodworth et al., 2017; Wu et al., 2019; Zafar et al., 2017a,b; Zhang and Bareinboim, 2018; Zhang et al., 2017). Here we synthesize and critically examine the statistical properties of popular formal fairness approaches as well as the consequences of enforcing them. Using both theory and empirical evidence, we argue that these approaches, when used as algorithmic design principles, can often cause more harm than good. In contrast to popular axiomatic approaches to algorithmic fairness, we advocate for a consequentialist perspective that directly grapples with the difficult policy trade-offs inherent to many algorithmically guided decisions.

We begin, in Section 2, by proposing a two-part taxonomy of formal fairness definitions. Our first category of definitions encompasses those that consider the effects of decisions on disparities. Imagine, for example, designing an algorithm to guide decisions for college admissions. 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. Our second category of definitions encompasses those that seek to limit both the direct and indirect effects of one's group membership on decisions. Following the principle that decisions should be agnostic to legally protected attributes like race and gender (cf. Dwork et al., 2012), one might mandate that these features not be provided to the algorithm. Further, 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.

These formalizations of fairness have considerable intuitive appeal. It can feel natural to exclude protected characteristics in a drive for equity; and one might understandably interpret disparities in error rates as indicating problems with an algorithm's design or with the data on which it was trained. However, in Sections 3 and 4, we show that both classes of algorithmic fairness definitions suffer from deep statistical limitations. For example, for natural families of utility functions—like those that prefer both higher academic preparedness and more student-body diversity—we prove that common fairness criteria almost always, in a measure theoretic sense, lead to strongly Pareto dominated decision policies.[1]

1. A policy is strongly Pareto dominated if there is an alternative feasible policy that is preferred under every utility function in the family (cf. Section 4.2).

In particular, in our running college admissions example, adhering to several of the popular conceptions of fairness we consider would simultaneously result in lower student-body diversity and a less academically prepared class, relative to what one could attain by explicitly tailoring admissions policies to achieve desired outcomes. In fact, under one prominent definition of 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 formal fairness criteria are thus often at odds with policy goals, and, perversely, can harm the very same groups one ostensibly sought to protect by developing and adopting axiomatic notions of fairness.

How, then, can we ensure algorithms are fair? There are no easy solutions, but we conclude in Section 5 by offering several observations and suggestions for designing more equitable algorithms. Most importantly, we believe it is critical to acknowledge and tackle head-on the substantive trade-offs at the heart of many decision problems. For example, when creating a college admissions policy, one must necessarily make difficult choices that balance competing priorities. Formal fairness axioms are poor tools for engaging with these challenging issues. Our overarching exhortation is thus to recognize algorithms as encoding policy choices, and to accordingly tailor their design.

Contributions. To summarize, we offer three main contributions. First, we survey the fairness literature, describing existing fairness definitions and organizing them into a two-part taxonomy. Our categorization of formal fairness definitions proposed in the computer science literature highlights their connections to influential legal and economic notions of discrimination. Second, we lay out a consequentialist framework for designing equitable algorithms. Our framework is motivated by viewing algorithmic fairness as a policy objective rather than as a technical problem. This approach exposes the statistical and normative limitations of many popular formal fairness definitions. Finally, we apply our consequentialist framework to develop a positive vision for addressing problems of fairness and equity in algorithm design.

Much of the content we present synthesizes and builds on research that we and our collaborators have conducted over the last several years (Cai et al., 2020; Chohlas-Wood et al., 2023a,b; Corbett-Davies et al., 2017; Koenecke et al., 2023). In particular, we draw heavily on two papers by Corbett-Davies and Goel (2018) and Nilforoshan et al. (2022). In addition to synthesis, we broaden the formal theoretical results presented in this line of work and offer new, concrete illustrations of our theoretical arguments. Some of the results and arguments we present date back five years, and the field of algorithmic fairness has since moved forward in many ways. For example, in the intervening time, there has been increasing recognition of the shortcomings of popular formal fairness definitions (Barocas et al., 2019). Nevertheless, we believe our message is as relevant as ever. For instance, within the research community, new algorithmic fairness definitions are regularly introduced that, while different in some respects, frequently suffer from the same statistical and conceptual limitations as the notions we survey here. In the broader world, policymakers, algorithm designers, journalists, and advocates often still evaluate algorithms—and accordingly influence decisions—by turning to these formal fairness definitions without necessarily appreciating their shortcomings. For example, proposed legislation in Idaho sought to require that pretrial risk assessment algorithms have equal error rates across groups (Idaho H.B. 118, 2019). Although the proposed bill was never passed, it the illustrates the ways

in which these formal measures have garnered significant attention beyond the academic community.

The call to build equitable algorithms will only grow over time as automated decisions become even more widespread. As such, it is imperative to address limitations in past formulations of fairness, to identify best practices moving forward, and to outline important open research questions. By synthesizing and critically examining recent developments in fair machine learning, we hope to help both researchers and practitioners advance this increasingly influential field.

2. Mathematical Definitions of Fairness

We start by assembling and categorizing definitions of algorithmic fairness into a two-part taxonomy: those that seek to limit the effect of decisions on disparities, and those that seek to limit the effect of protected attributes like race or gender on the decisions themselves. We first introduce formal notation and concrete examples of decision problems in which one might seek to apply these fairness definitions, before reviewing prominent examples of both approaches in turn.

2.1 Formal Setting

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 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, $D = 1$.[2,3] Given a budget b with $0 < b \leq 1$, we require the decision rule to satisfy $\mathbb{E}[D] \leq b$. Finally, we suppose that each individual has some associated binary outcome Y. In some cases, we will be concerned with the causal effect of the decision D on Y, in which case we imagine that there exist two potential outcomes, $Y(0)$ and $Y(1)$, corresponding to what happens to the individual depending on whether they receive a negative or positive decision.[4]

To make our discussion concrete, we imagine two running examples corresponding to this formal setting: diabetes screening and college admissions. As we discuss in detail below, these two examples differ in the extent to which there is agreement about the ultimate value of different decision policies, which in turn impacts our mathematical analysis. Diabetes is a common and serious health condition that afflicts many American adults. If caught early, it is often possible to avoid some of the most significant consequences of the disease through, for example, changes to one's diet and physical routine. A blood test can be used to determine whether an individual has diabetes, but as with many screening tools, there are risks and inconveniences associated with screening (e.g., a patient may need to

2. That is, $D = \mathbb{1}_{U_D \leq d(X)}$, where U_D is an independent uniform random variable on $[0, 1]$.
3. By "positive," we simply mean the decision D is greater than zero, without ascribing any normative position to the decision. Individuals may or may not have a preferences for "positive" decisions in this sense.
4. As is implicit in our notation, we assume that there are no spillover effects between units (Imbens and Rubin, 2015).

take time off from work). In particular, if an individual were *certain* that they did not have diabetes, then they would prefer not to undergo screening. Our goal is to design an equitable screening policy $d(x)$ to determine which patients have $(Y = 1)$ or do *not* have $(Y = 0)$ diabetes, based on a set of covariates X. For example, following Aggarwal et al. (2022), the screening decision may be based on a patient's age, body mass index (BMI) and race. (Those authors argue that consideration of race, while controversial, leads to more precise and equitable estimates of diabetes risk, a point we return to in Section 3.3.) We further imagine the budget b equals 1, corresponding to the fact that everyone could be screened in principle.

Our second example concerns college admissions. Here, the population of interest is applicants to a particular college, and the decision D is the admissions committee's binary admissions decision. To simplify our exposition, we assume all admitted students attend the school. In this setting, the covariates X may, for example, consist of an applicant's test score and race $A \in \{a_0, a_1\}$, and Y is a binary variable that indicates college graduation (i.e., degree attainment). In contrast to our diabetes example, here we imagine that the decision itself may affect the outcomes. Specifically, $Y(1)$ and $Y(0)$ describe whether an applicant *would* attain a college degree if admitted to or if rejected from the school we consider, respectively. Note that $Y(0)$ is not necessarily zero, as a rejected applicant may attend—and graduate from—a different university. Further, in this case we set the budget b to be less than one to reflect the fact that the admissions committee has limited resources and is unable to admit every candidate.

As mentioned above, a key distinction between these two examples is the extent to which stakeholders may agree on the value of different potential decision policies. For example, in college admissions, there may be significant disagreement on how to balance competing priorities, such as academic preparedness and class diversity.[5] Admissions committees may seek to increase both dimensions, but there is often an inherent trade-off, particularly since there is a limit on the number of students that can be admitted by the college (i.e., $b < 1$). Our diabetes example, in contrast, reflects a setting where there is ostensibly broader agreement on the value of different decision policies. Indeed, since there is effectively no limit on the number of diabetes tests that can be administered (i.e., $b = 1$), we can model the value of a decision policy as the sum of each individual's value for being screened.[6] In Sections 3 and 4, we in turn examine the structure of equitable decision making in the absence and presence of such trade-offs. First, though, we introduce several formal fairness criteria.

5. In some jurisdictions, explicit considerations of racial diversity may be prohibited. For instance, a recent U.S. Supreme Court case bars colleges from explicitly considering race in admissions; however, colleges may consider "an applicant's discussion of how race affected the applicant's life" (SFFA v. Harvard, 2023). U.S. colleges may also consider other forms of diversity, such as economic or geographic diversity.
6. In the case of infectious diseases—which involve greater externalities—there is again often disagreement about the value of different screening and vaccination policies. Paulus and Kent (2020) similarly draw a distinction between *polar* settings (in which parties have competing interests, like our admissions example) and *non-polar* settings (where there is broad alignment, as in our diabetes example).

2.2 Limiting the Effect of Decisions on Disparities

A popular class of fairness definitions requires that error rates (e.g., false positive and false negative rates) are equal across protected groups (Hardt et al., 2016).[7] We refer to these definitions as examples of "classification parity," meaning that some given measure of classification error is equal across groups defined by attributes such as race and gender. In particular, we include in this definition any measure that can be computed from the two-by-two confusion matrix tabulating the joint distribution of decisions D and outcomes Y for a group. Berk et al. (2021) enumerate seven such statistics, including false positive rate, false negative rate, precision, recall, and the proportion of decisions that are positive. The proportion of positive decisions is not, strictly speaking, a measure of "error", but we nonetheless include it under classification parity since it can be computed from a confusion matrix. We also include the area under the ROC curve (AUC), a popular measure among practitioners examining the fairness of algorithms (Skeem and Lowenkamp, 2016).

Two of the above measures—the proportion of decisions that are positive, and the false positive rate—have received considerable attention in the machine learning community (Agarwal et al., 2018; Blum and Stangl, 2019; Calders and Verwer, 2010; Chouldechova, 2017; Edwards and Storkey, 2016; Feldman et al., 2015; Hardt et al., 2016; Jung et al., 2020a; Kamiran et al., 2013; Pedreshi et al., 2008; Zafar et al., 2017a,c; Zemel et al., 2013).

Definition 1 *We say that* demographic parity *holds when*[8]

$$D \perp\!\!\!\perp A. \tag{1}$$

Definition 2 *We say that* equalized false positive rates *holds when*

$$D \perp\!\!\!\perp A \mid Y = 0. \tag{2}$$

In our running diabetes example, demographic parity means that the proportion of patients who are screened for the disease is equal across race groups. Similarly, in our college admissions example, demographic parity means an equal proportion of students is admitted across race groups. Equalized false positive rates, in our diabetes example, means that among individuals who in reality do not have diabetes—and thus for whom screening, *ex post*, would not have been beneficial—screening rates are equal across race groups.[9]

Causal analogues of these definitions have also recently been proposed (Coston et al., 2020; Imai and 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.[10] Below we list three representative examples of this class of fairness

7. Some work relaxes strict equality of error rates or other metrics to requiring only that the difference be at most some fixed ϵ (e.g., Nabi and Shpitser, 2018). For ease of exposition, we consider strict equality throughout, though we emphasize that the spirit of the critique we develop applies also in cases where fairness constraints are approximate, rather than exact.
8. We use the notation $X \perp\!\!\!\perp Y$ throughout to mean that the random variables X and Y are independent.
9. In our college admissions example, the decision D impacts the outcome Y. One could, in theory, apply the definition of error rate parity above to that case by recognizing that $Y = Y(D)$. However, that interpretation does not seem aligned with the original intent of the definition. We instead discuss the admissions example in the context of the explicitly causal definitions of fairness below.
10. 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;

definitions: counterfactual predictive parity (Coston et al., 2020), counterfactual equalized odds (Coston et al., 2020; Mishler et al., 2021), and conditional principal fairness (Imai and Jiang, 2020).[11]

Definition 3 *We say that* counterfactual predictive parity *holds when*

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

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. (For our diabetes example, because the screening decision does not affect whether a patient actually has diabetes, $Y(0) = Y(1) = Y$, and so counterfactual predictive parity, as well as the causal definitions below, reduce to their non-causal analogues).

Definition 4 *We say that* counterfactual equalized odds *holds when*

$$D \perp\!\!\!\perp A \mid Y(1). \tag{4}$$

In our running college admissions 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 5 *We say that* conditional principal fairness *holds when*

$$D \perp\!\!\!\perp A \mid Y(0), Y(1), W, \tag{5}$$

where, for some 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 the college admissions 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 framework for understanding fairness considers the effects of protected attributes on decisions. This approach can be understood as codifying the legal notion of disparate treatment (Goel et al., 2017; Zafar et al., 2017a)—which we discuss further in Section 5.1. Perhaps the simplest way to limit the effects of protected attributes on decisions is to require that the decisions do not explicitly depend on them, what some call "fairness through unawareness" (cf. Dwork et al., 2012).

Wang et al., 2019), here we focus exclusively on decisions, with predictions implicitly impacting decisions but not explicitly appearing in our definitions.

11. 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 6 *Suppose that the covariates can be partitioned into the protected attributes and all other covariates, i.e., that $\mathcal{X} = \mathcal{X}_u \times \mathcal{A}$, where $\mathcal{X}_u$ consists of "unprotected" attributes. Then, we say that* blinding *holds when, for all $a, a' \in \mathcal{A}$ and $x_u \in \mathcal{X}_u$,*

$$d(x_u, a) = d(x_u, a'). \tag{6}$$

In our running diabetes example, blinding holds when the screening decision depends solely on factors like age and BMI, and, in particular, does not depend on the patient's race. We similarly say college admissions decisions satisfy blinding when the decisions depend on factors like test scores and extracurricular activities, but not race.

Blinding is closely tied to the notion of *calibration*, the requirement that, conditional on the estimated probability of some outcome (such as graduation from college or having diabetes), the outcome is independent of group membership. For example, among people with an estimated diabetes risk of 1%, calibration would require that the proportion of individuals who actually have diabetes be the same across groups. Many authors treat calibration as a kind of fairness constraint—in particular, to ensure that the meaning of estimated risks do not differ across groups—and it has received considerable attention in the fairness literature (e.g., Hébert-Johnson et al., 2018; Rothblum and Yona, 2022). We note, though, that miscalibration is equivalent to blindness in practice. In particular, when estimation error is small, risk estimates that are allowed to depend on group membership are calibrated; conversely, risk estimates that are blind to group membership typically are miscalibrated—an empirical phenomenon shown and discussed in Figure 3 below. Because of this close relationship, we do not treat calibration as a separate fairness constraint, but we do discuss calibration and its relationship to blinding in detail in Sections 3.3.1 and 5.2.

In contrast to blinding—in which race and other protected attributes are barred from being an explicit input to a decision rule—the causal versions of this idea consider both the direct and indirect effects of protected attributes on decisions (Kilbertus et al., 2017; Kusner et al., 2017; Mhasawade and Chunara, 2021; Nabi and Shpitser, 2018; Wang et al., 2019; Wu et al., 2019; Zhang and Bareinboim, 2018; Zhang et al., 2017). 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. In this vein, a decision rule is deemed 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).[12] 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).

12. 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 and Kohler-Hausmann, 2020). Such difficulties can sometimes be addressed by considering a change in the *perception* of race by a decision maker (Greiner and Rubin, 2011)—for instance, by changing the name listed on an employment application (Bertrand and Mullainathan, 2004), or by masking an individual's appearance (Chohlas-Wood et al., 2021; Goldin and Rouse, 2000; Grogger and Ridgeway, 2006; Pierson et al., 2020).

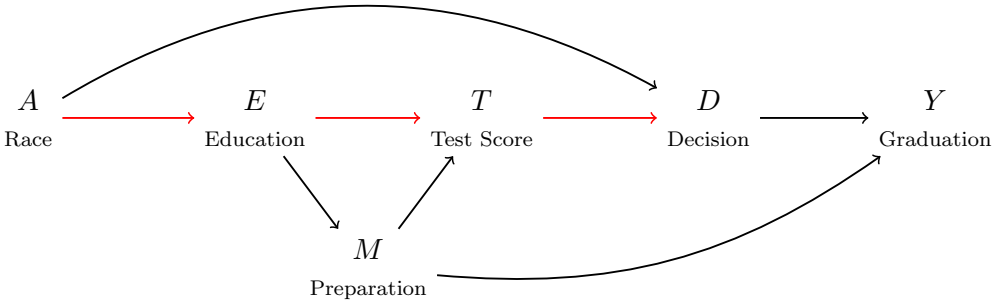


Figure 1: A causal DAG illustrating a hypothetical process for college admissions. Under path-specific fairness, one may require, for example, that race does not affect decisions along the path highlighted in red.

Definition 7 Counterfactual fairness *holds when*

$$\mathbb{E}[D(a') \mid X] = \mathbb{E}[D \mid X], \tag{7}$$

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

In our running college admissions 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 and 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 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.[13]

To formalize 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 score, highlighted in red in Figure 1. We describe the distribution of $D_{\Pi,a,a'}$ generatively, formally showing how to produce a draw from this distribution. To start, we draw values U_E^*, U_M^*, U_T^*, U_D^* of the exogenous variables. Now, the first column in Table 1 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 $D_{\Pi,a,a'}$.

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 8 *Let Π be a collection of paths, and, for some function ω 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 $a' \in \mathcal{A}$,*

$$\mathbb{E}[D_{\Pi,A,a'} \mid W] = \mathbb{E}[D \mid W]. \tag{8}$$

13. 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.

$A^* = a$	$\overline{A}^* = a'$
$E^* = f_E(A^*, U_E^*)$	$\overline{E}^* = f_E(\overline{A}^*, U_E^*)$,
$M^* = f_M(E^*, U_M^*)$	$\overline{M}^* = f_M(E^*, U_M^*)$,
$T^* = f_T(E^*, M^*, U_T^*)$	$\overline{T}^* = f_T(\overline{E}^*, M^*, U_T^*)$
$D^* = f_D(A^*, T^*, U_D^*)$	$\overline{D}^* = f_D(A^*, \overline{T}^*, U_D^*)$

Table 1: Computing path-specific counterfactuals for the DAG in Figure 1. The first column corresponds to draws V^* for each node V, 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 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. Equitable Decisions in the Absence of Externalities

In many decision-making settings, the decision maker is free to make the optimal decision for each individual, without consideration of spillover effects or other externalities. For instance, in our diabetes screening example, one could, in principle, screen all patients if that course of action were medically advisable.

To investigate notions of fairness in these settings, we first introduce a framework for utilitarian decision analysis. Specifically, we consider in this section situations in which there is broad agreement on the utility of different potential courses of action. (In the subsequent section, we consider cases where stakeholders disagree on the precise form of the utility.) In this setting, "threshold rules" maximize utility. We then describe the statistical phenomenon of inframarginality, a property that is endemic to fairness definitions that seek to enforce some form of classification parity. In particular, we discuss, both informally and mathematically, why inframarginality almost surely—in a measure theoretic sense—renders optimal decision making incompatible with classification parity. Finally, we discuss blinding. In parallel to our discussion of classification parity, we see that in many important settings, the information loss associated with, e.g., removing protected information from a predictive model, results in less efficient decision making without compensatory benefits. Moreover, in general, we see that the more stringent the standard of masking—e.g., removing not only

direct but also indirect effects of protected attributes—the greater the potential harm that results from enforcing it.

3.1 Utility, Risk, and Threshold Rules

A natural way to analyze a decision, such as deciding whether an individual should be screened for diabetes, is to consider the costs and benefits of various possible outcomes under different courses of action. For instance, a patient screened for diabetes who does not have the disease still has to bear the risks, discomfort, and inconvenience associated with the blood test itself, while a patient who is *not* screened but does in fact have the disease loses out on the opportunity to start treatment.

In general, the benefit of making decision $D = 1$ *over* $D = 0$ when the outcome Y equals y can be represented by $v(y)$. For instance, in our diabetes example, $v(1)$ represents the net benefit of screening over not screening when the patient has diabetes; and $-v(0)$ is the net cost of screening when the patient does not have diabetes, including both monetary and non-monetary costs, such as discomfort and loss of time.[14] Let $r(x) = \Pr(Y = 1 \mid X = x)$ be the *risk* of Y equalling 1 when $X = x$. Then the expected benefit of making decision $D = 1$ over $D = 0$ for an individual with covariates $X = x$ is

$$\begin{aligned} u(x) &= \mathbb{E}[v(Y) \mid X = x] \\ &= r(x) \cdot v(1) + [1 - r(x)] \cdot v(0). \end{aligned}$$

Here, for ease of interpretation, we restrict our utility to be of the form $u(x) = \mathbb{E}[v(Y) \mid X = x]$ for some function v, and we also assume there is no budget constraint (i.e., $b = 1$). In Section 4, we allow the utility $u(x)$ to be an arbitrary function on $\mathcal{X}$ and consider $b < 1$, which induces the trade-offs in decisions that are central to our later discussion.

The *aggregate* expected utility of a decision policy $d(x)$—relative to the baseline policy of taking action $D = 0$ for all individuals—is then given by $u(d) = \mathbb{E}[d(X) \cdot u(X)]$. We say a decision policy $d^*(x)$ is utility-maximizing if

$$u(d^*) = \max_d u(d).$$

It is better, in expectation, for an individual with covariates $X = x$ to take action $D = 1$ instead of $D = 0$ when $u(x) > 0$; that is, when[15]

$$r(x) > \frac{v(0)}{v(0) - v(1)}. \tag{9}$$

Thus, the decision with the maximum utility can be determined by comparing an individual's risk against a particular risk threshold t, defined by the right-hand side of Eq. (9).

14. For ease of exposition, we assume that costs and benefits are identical across individuals; in reality, these could vary, e.g., depending on age. When utilities vary by person, the optimal decision rule is to screen only those with positive individual utility, in line with our subsequent discussion.

15. We assume, without loss of generality, that $v(1) > v(0)$. If $v(1) < v(0)$, we can take $Y' = 1 - Y$ as our outcome of interest; relative to Y', the inequality will be reversed. If $v(1) = v(0)$, then the outcome is irrelevant. In this degenerate case, the higher utility decision depends on the sign of $v(1)$ alone, and not the risk.

We refer to this kind of policy as a *threshold policy*. In particular, we see that a utility-maximizing decision for each individual—i.e., $d(x) = 1$ if $r(x) > t$ and $d(x) = 0$ if $r(x) \leq t$—is also a decision policy that maximizes aggregate utility, so there is no conflict between doing what is best from each individual person's perspective and what is best for the population as a whole.

While our framing in terms of expected utility is suitably general, threshold policies can be simpler to interpret when we reparameterize in terms of more familiar quantities. In the diabetes screening example, if the patient does not have diabetes, the cost of action $D = 1$ over $D = 0$ is $-v(0) = c_{\text{TEST}}$, i.e., the cost (monetary and non-monetary) of the test. If the patient does have diabetes, the benefit of $D = 1$ over $D = 0$ is $v(1) = b_{\text{TREAT}} - c_{\text{TEST}}$, i.e., the benefit of treatment minus the cost of the test. Rewriting Eq. (9) in terms of these quantities gives

$$t = \frac{c_{\text{TEST}}}{b_{\text{TREAT}}}.$$

In particular, if the benefit of early treatment of diabetes is 50 times greater than the cost of performing the diagnostic test, one would ideally screen patients who have at least a 2% chance of developing the disease.

Threshold rules are a natural approach to decision making in a variety of settings. In our running medical example, a threshold rule corresponds to screening patients with a sufficiently high risk of having diabetes. A threshold rule—with the optimally chosen threshold—ensures that only the patients at highest risk of having diabetes take the test, thereby optimally balancing the costs and benefits of screening. Indeed, in many medical examples, from diagnosis to treatment, there are no significant externalities. As a result, deviating from utility-maximizing threshold policies can only force individuals to experience greater costs—in the form of unnecessary tests or untreated illness—in expectation, without compensatory benefits. We return to the problem of optimal (and equitable) decision-making in the presence of externalities in Section 4.

3.2 The Problem of Inframarginality

In the setting that we have been considering, threshold policies guarantee optimal choices are made for each individual. However, as we now show, threshold policies in general violate various versions of classification parity, such as demographic parity and equalized false positive rates. This incompatibility highlights a critical limitation of classification parity as a fairness criterion, as enforcing the definition often requires making decisions that harm individuals without any clear compensating benefits.

To help build intuition for this phenomenon, we consider the empirical distribution of diabetes risk among White and Asian patients. Following Aggarwal et al. (2022), we base our risk estimates on age, BMI, and race, using a sample of approximately 15,000 U.S. adults aged 18–70 interviewed as part of the National Health and Nutrition Survey (NHANES; Centers for Disease Control and Prevention, 2011-2018). The resulting risk distributions are shown in the left-hand panel of Figure 2. The dashed vertical lines show the group means, and indicate that the incidence of diabetes is higher among Asian Americans (11%)

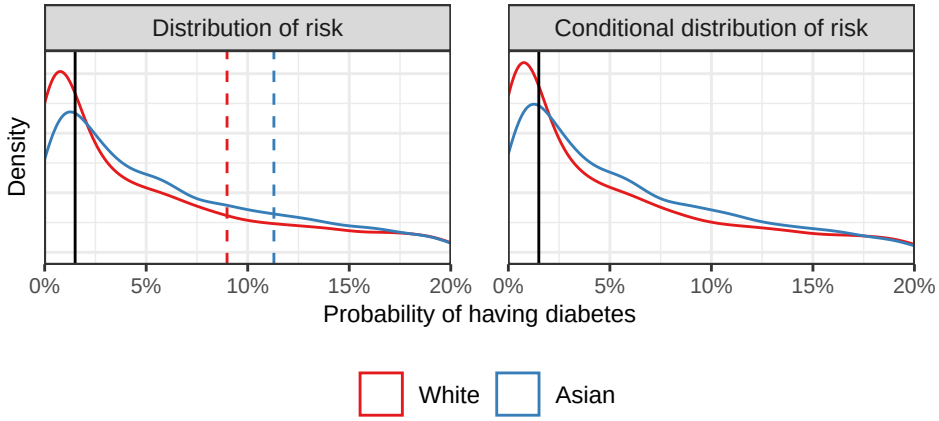


Figure 2: A graphical illustration of the incompatibility between threshold policies and classification parity, based on the National Health and Nutrition Survey. *Left:* The distribution of diabetes risk for White Americans and Asian Americans, with the dashed vertical lines corresponding to the overall incidence rate within each group. At a screening threshold of 1.5% (indicated by the solid black line), the screening rate for Asian Americans is higher than for White Americans, violating demographic parity. *Right:* The distribution of diabetes risk among individuals who do *not* have diabetes. Since the proportion of Asian Americans above the screening threshold is greater than the proportion of White Americans above the threshold, the false positive rate for Asian Americans is greater than the false positive rate for White Americans.

than among White Americans (9%).[16] This difference in base rates is also reflected in the heavier tail of the risk distribution among Asian individuals.

Drawing on recommendations from the United States Preventative Screening Task Force, Aggarwal et al. (2022) suggest screening patients with at least a 1.5% risk of diabetes, irrespective of race. We depict this risk threshold by the solid black vertical line in the plot. Based on that recommendation, 81% of Asian Americans and 69% of White Americans are to the right of the threshold and should be screened—violating demographic parity. If, hypothetically, we were to raise the screening threshold to 2.2% for Asian Americans and lower the threshold to 1% for White Americans, 75% of people in both groups would be screened, satisfying demographic parity.[17] The cost of doing so, however, would be failing

16. The precise shapes of the risk distributions depend on the set of covariates used to estimate outcomes, but the means of the distributions correspond to the overall incidence of diabetes in each group, and, in particular, are unaffected by the choice of covariates. It is thus necessarily the case that the risk distributions will differ across groups in this example, regardless of which covariates are used.
17. Corbett-Davies et al. (2017) show that group-specific threshold policies are utility-maximizing under the constraint of satisfying various notions of classification parity, including demographic parity and equality of false positive rates.

to screen some Asian Americans who have a relatively high risk of diabetes, and subjecting some relatively low-risk White Americans to a procedure that is medically inadvisable given their low likelihood of having diabetes. In an effort to satisfy demographic parity, we would have harmed members from both groups.

This example illustrates a similar incompatibility between threshold policies and equalized false positive rates. In our setting, the false positive rate for a group is the screening rate among those in the group who do not in reality have diabetes. To visualize the race-specific false positive rates, the right-hand panel of Figure 2 shows the distribution of diabetes risk among those individuals who do not have diabetes. (Because the overall prevalence of diabetes is low, the conditional distribution displayed in the right-hand panel is nearly identical to the unconditional distribution displayed in the left-hand panel.) The false positive rate for each group is the proportion of people in the group falling to the right of the 1.5% screening threshold. In this case, the false positive rate is 79% for Asian Americans and 67% for White Americans—violating equalized false positive rates. As before, we could alter the screening guidelines to equalize false positive rates, but doing so requires deviating from our threshold policy, in which case we would end up screening some individuals who are relatively low-risk and not screening others who are relatively high-risk.

In this example, the incompatibility between threshold policies and classification parity stems from the fact that the risk distributions differ across groups. This general phenomenon is known as the problem of *inframarginality* in the economics and statistics literature, and has long been known to plague tests of discrimination in human decisions (Anwar and Fang, 2006; Ayres, 2002; Carr and Megbolugbe, 1993; Engel and Tillyer, 2008; Galster, 1993; Knowles et al., 2001; Pierson et al., 2018; Simoiu et al., 2017). Common legal and economic understandings of fairness are concerned with what happens at the *margin* (e.g., whether the same standard is applied to all individuals)—a point we return to in Section 5. What happens at the margin also determines whether decisions maximize social welfare, with the optimal threshold set at the point where the marginal benefits equal marginal costs. However, popular error metrics assess behavior away from the margin, hence they are called *infra*-marginal statistics. As a result, when risk distributions differ, standard error metrics are often poor proxies for individual equity or social well-being.

In general, we expect any two non-random subgroups of a population to differ on a variety of social and economic dimensions, which in turn is likely to yield risk distributions that differ across groups. As a result, as our running diabetes example shows, the optimal decision policy—which maximizes each patient's own well-being—will likely violate various measures of classification parity. Thus, to the extent that formal measures of fairness are violated, that tells us more about the shapes of the risk distributions than about the quality of decisions or the utility delivered to members of any group. This intuition can be made precise, in the sense that for *almost every* risk distribution, the optimal decision policy violates the various notions of classification parity considered here.

The notion of *almost every* distribution that we use here was formalized by Christensen (1972), Hunt et al. (1992), Anderson and Zame (2001), and others (cf. Ott and 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.)

Theorem 9 *Let t be the optimal decision threshold, as in Eq. (9). If $0 < t < 1$, then for almost every collection of group-specific risk distributions which have densities on $[0, 1]$, no utility-maximizing decision policy satisfies demographic parity or equalized false positive rates.*

The proof of Theorem 9, which formalizes the informal discussion above, is given in Appendix F.6. At a high level, the constraints of classification parity are sensitive to even small perturbations in the underlying risk distributions. As a result, any particular collection of risk distributions is unlikely to satisfy the constraints. For simplicity, we have been considering settings in which the decision D does not impact the outcome Y. However, this basic style of argument extends to causal settings, showing that threshold policies are almost surely, in the measure theoretic sense, incompatible with counterfactual predictive parity, counterfactual equalized odds, and conditional principal fairness—definitions of fairness that we consider in depth in Section 4, in the more complex setting of having a budget $b < 1$.

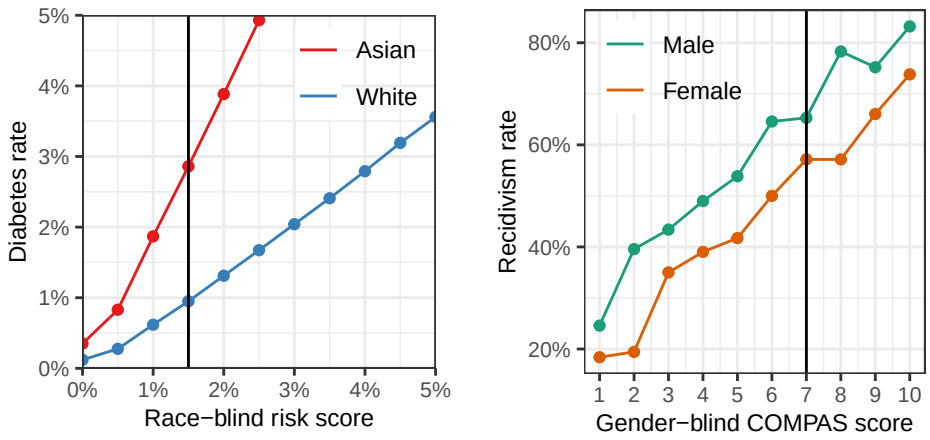
3.3 The Problem with Fairness through Unawareness

We now consider notions of fairness, both causal and non-causal, that aim to limit the effects of attributes on decisions. As above, we show the inherent incompatibility of these definitions with optimal decision making. We note, though, that while blinding can lead to suboptimal decisions—and, in some cases, harm marginalized groups—the legal, political, and social benefits of, for example, race-blind and gender-blind algorithms may outweigh their costs in certain instances (Cerdeña et al., 2020; Coots et al., 2023).

3.3.1 Blinding

A common starting point for designing an ostensibly fair algorithm is to exclude protected characteristics from the statistical model. This strategy ensures that decisions have no explicit dependence on group membership. For instance, in the case of estimating diabetes risk, one could use only BMI and age—rather than including race, as we did above. However, excluding race from models of diabetes risk can ultimately harm both White and Asian patients.

In Figure 3a, we compare the actual diabetes rate to estimated diabetes risk resulting from the race-blind risk model. Aggarwal et al. (2022) showed that Asian patients have higher incidence of diabetes than White patients with comparable age and BMI. As a result, the race-blind model systematically underestimates risk for Asian patients and systematically overestimates risk for White individuals. In particular, applying a nominal 1.5% screening threshold under the race-blind model amounts to effectively applying a 1%



(a) Diabetes

(b) Recidivism

Figure 3: Calibration plots showing the effect of removing protected attributes from risk models when estimating the risk of diabetes (left) and recidivism (right). Because Asian patients with the same BMI and age have a higher rate of diabetes, the race-blind model underestimates their risk of having diabetes. Similarly, because women reoffend at lower rates than men with similar criminal histories, the gender-blind COMPAS score overstates the recidivism risk for women.

screening threshold to White patients and a 3% screening threshold to Asian patients. Thus, by using race-blind risk scores, we subject relatively low-risk White patients to screening, and fail to screen Asian patients who have a relatively high risk for having diabetes. A race-aware model would ensure that nominal risk thresholds correspond to observed incidence rates across race groups.

This phenomenon—which we call miscalibration across subgroups—is not unique to diabetes screening. Consider, for instance, the case of pretrial recidivism predictions. Shortly after an individual is arrested in the United States, a judge must often determine conditions of release pending future court proceedings. In many jurisdictions across the country, these pretrial decisions are informed by statistical risk estimates of the likelihood the individual would be arrested or convicted of a future crime if released. After adjusting for factors such as criminal history, age, and substance use, women have been found to reoffend less often than men in many jurisdictions (DeMichele et al., 2018; Skeem et al., 2016). Consequently, gender-blind risk assessments are miscalibrated, meaning that they tend to overstate the recidivism risk of women and understate the recidivism risk of men.

Figure 3b illustrates this point, plotting the observed recidivism rate for men and women in Broward County, Florida as a function of their gender-blind COMPAS risk scores—a commonly used risk assessment tool (Bao et al., 2021). In particular, women with a COMPAS score of seven recidivate about 55% of the time, whereas men with the same score recidivate about 65% of the time. Said differently, women with a score of seven

recidivate approximately as often as men with a score of five, and this two-point differential persists across the range of scores. By acknowledging the predictive value of gender in this setting, one could create a decision rule that detains fewer people (particularly women) while achieving the same public safety benefits. Conversely, by ignoring this information and basing decisions solely on the gender-blind risk assessments, one would effectively be subjecting women to a more stringent risk standard—and potentially harsher penalties—than men.

As in the case of classification parity, one cannot typically remove protected attributes from the risk predictions without decreasing utility (cf. Manski et al., 2022); however, the reduction in utility is not always as large as one might expect (Coots et al., 2023). In concrete terms, in our running diabetes example, basing decisions on race-blind risk estimates necessarily means screening some patients who would have preferred not to be screened had they been given race-aware risk estimates, and, conversely, not screening some patients who would have preferred to be screened had they been given the more complete estimates. We state this result formally below.

Theorem 10 *Suppose* $0 < t < 1$*, where* t *is the optimal decision threshold on the risk scale, as in Eq.* (9)*. Let* $\pi : \mathcal{X}_u \times \mathcal{A} \to \mathcal{X}_u$ *denote restriction to the unprotected covariates. Let* $\rho(x) = \Pr(Y = 1 \mid \pi(X) = \pi(x))$ *denote the risk estimated using the blinded covariates. Suppose that* $r(x)$ *and* $\rho(x)$ *have densities on* $[0, 1]$ *that are positive in a neighborhood of* t*. Further suppose that there exists* $\epsilon > 0$ *such that the conditional variance* $\textsc{Var}(r(X) \mid \rho(X)) > \epsilon$ *a.s., where* $r(x)$ *is the risk estimated from the full set of covariates. Then no blind policy is utility-maximizing.*

The proof of Theorem 10 is given in Appendix D. In short, when race, gender, or other protected traits add predictive value—a condition codified in our assumption that the conditional variance be greater than ϵ—excluding these attributes will in general decrease utility, both for individuals and in the aggregate.

Basing decisions on blinded risk scores can harm individuals and communities, for example by failing to flag relatively high-risk Asian patients for diabetes screening. But it is also important to consider potential harms stemming from the use of race- and gender-specific risk tools. In medicine, for instance, one might worry that race-specific risk assessments could encourage doctors and the public-at-large to advance spurious and pernicious arguments about inherent differences between race groups. In reality, the differences in diabetes risk we see are likely due to a complex mix of factors, both environmental and genetic, and should not be misinterpreted as indicating any causal effects of race. Indeed, even "race" itself is a thorny, socially influenced, concept, that elides easy definition. Similarly, the use of gender-specific recidivism estimates could reduce trust in the criminal justice system, giving the impression that individuals are held to different standards based on their gender. (Though, as we have seen above, *blinded* risk assessments can likewise—and perhaps more persuasively—be said to subject individuals to different standards based on their race and gender.) In some circumstances, race- and gender-specific risk estimates are even prohibited by law—a topic we return to in Section 5.1. For these reasons, risk assessments in medicine, criminal justice, and beyond have generally avoided using race, gender, and other sensitive demographic attributes. Ultimately, when constructing risk assessment tools, it is

important to acknowledge and carefully balance both the costs and benefits of blinding in any given circumstance.

3.3.2 Counterfactual and Path-Specific Fairness

As discussed in Section 2, counterfactual and path-specific fairness are generalizations of simple blinding that attempt to account for both the direct and indirect effects of protected attributes on decisions. Because the constraints are more stringent, the resulting decrease in utility is proportionally greater. In particular, 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 admissions example, path-specific fairness requires admitting all applicants with the same probability, irrespective of academic preparation or group membership.

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 $a \in \mathcal{A}$,

$$\begin{aligned} d(x^*) &= \mathbb{E}[D_{\Pi,A,a} \mid X = x^*] \\ &= \sum_{x \in \alpha^{-1}(a)} d(x) \cdot \Pr(X_{\Pi,A,a} = x \mid X = x^*). \end{aligned} \tag{10}$$

That is, the probability of an individual with covariates x^* receiving a positive decision must be the average probability of the individuals with covariates x in group a receiving a positive decision, weighted by the probability that an individual with covariates x^* in the real world *would* have covariates x counterfactually.

Next, we suppose that 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. (10) 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 11 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 11 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 11 *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* $\alpha(x') = a'$, $\alpha(x) \neq a'$, *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 11 holds for any reduced set of covariates Z that is not causally affected by changes in A (e.g., Z is not a descendant 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 admissions example, this result shows that, under mild conditions, causally-fair policies require admitting all applicants with equal probability. In particular, suppose that 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, for instance, we might expect if the individual-level causal effects are drawn from an (appropriately discretized) normal distribution. In this case, the empty set of reduced covariates—formally encoded by setting ζ to a constant function—satisfies the conditions of Theorem 11. The theorem then implies that under any Π-fair policy, every applicant is admitted with equal probability. (We motivated our admissions example by assuming that only a fraction $b < 1$ of applicants could be admitted; however, Theorem 11 holds irrespective of the budget, and, in particular, when $b = 1$, and so we discuss this result together with our others on unconstrained decision making as a natural extension of blinding.)

Even when decisions are not perfectly uniform lotteries, Theorem 11 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 11 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 11 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 and Shpitser, 2018; Zhang and Bareinboim, 2018), in which case Theorem 11 does not apply. Although, in that case, path-specific fairness is still typically incompatible with optimal decision-making, as shown in Theorem 17.

4. Equitable Decisions in the Presence of Externalities

We have thus far considered cases where there is largely agreement on the utility of different decision policies. In that setting, we showed that maximizing utility is at odds with various mathematical formalizations of fairness. We further argued that these results illustrate weaknesses in the formalizations themselves, since deviating from utility-maximizing polices in that setting can harm both individuals and groups—as seen in our diabetes screening example.

Agreement on the utility, however, is perhaps the exception rather than the rule. One could indeed argue that the value of mathematical formalizations of fairness is their ability to arbitrate between competing definitions of utility. Here we critically examine that perspective. We show, in analog to our previous results, that even when it is unclear how to balance competing priorities, enforcing existing fairness constraints typically leads to worse outcomes on each dimension. For instance, in our running college admissions example, policies constrained to satisfy various fairness constraints will typically require admitting a student body that is both less academically prepared and less diverse, relative to alternative policies that violate these mathematical fairness definitions.

We start, in Section 4.1, by examining our college admissions example in detail, illustrating in geometric terms how existing fairness definitions can lead to problematic admissions policies. Then, in Section 4.2, we develop our formal theory of equitable decision making in the presence of externalities. The mathematics necessary to establish our key results are significantly deeper than what we have needed thus far, but our high-level message is the same: enforcing several formal notions of fairness leads to policies that can paradoxically harm the very groups that they were designed to protect.

4.1 The Geometry of Fair Decision Making

To build intuition about the limitations of popular definitions of fairness, we return to our running example on college admissions. In that setting, we imagine an admissions committee debating the merits of different admissions policies. In particular, we imagine disagreement within the committee over how best to balance two competing objectives: academic preparation (operationalized, e.g., in terms of the high school grades and standardized test scores of admitted students) and class diversity (e.g., the number of admitted applicants from marginalized groups).

We assume that our hypothetical committee members all agree that more (total) academic preparedness and more class diversity are better. Thus, in the absence of any resource constraints (with $b = 1$, as is approximated in some online courses), the university could admit all applicants, maximizing both the number of admitted students from marginalized groups and also the total academic preparedness of the admitted class. But given limits on the number of students who can be admitted (i.e., $b < 1$), one must make difficult choices on whom to admit, with reasonable and expected disagreement on how much to trade one dimension for another. The trade-offs in decision making are most acute when the budget $b < 1$, and for this reason we focus here on that case.

In light of these trade-offs, one might turn to the myriad formal fairness criteria we have discussed to ensure admissions decisions are equitable. Many of the fairness definitions we consider make reference to a distinguished outcome Y. In our example, we can imagine

this outcome corresponds to college degree attainment, an *ex post* measure of academic preparedness. In the case of causal fairness definitions, we could take $Y(1)$ to mean degree attainment if the student were admitted, and $Y(0)$ to be degree attainment if the student were not admitted, with the understanding that a student who is not admitted could potentially attend and graduate from another university. For example, satisfying counterfactual predictive parity requires that among rejected applicants, the proportion who would have attained a college degree, had they been accepted, is equal across race groups. In these cases, we imagine academic preparedness is some student-level measure that connects observables X—upon which the committee must make their admissions decisions—to (potential) outcomes Y. For example, the "academic index" $m(x)$ might be a prediction of $Y(1)$ given X based on historical data, or, more generally, could encode committee preferences for both academic preparation and participation in extracurricular activities, among other factors.

The key point of our informal discussion thus far is that we assume committee members would like to enact an admissions policy d that balances two competing objectives. First, they would like a policy that leads to large $m(x)$, i.e., they would like $\mathbb{E}[m(X) \cdot d(X)]$ to be big, where $m(x)$ is some quantity that may, for example, encode academic preparedness and other preferences. Second, the committee would like large diversity, i.e., they would like $\mathbb{E}[\mathbb{1}_{\alpha(X)=a_1} \cdot d(X)]$ to be big, where a_1 corresponds to some target group of interest. All committee members would like more of each dimension, but, given the budget constraint, it is in general impossible to maximize both dimensions simultaneously, leading to the inherent trade-offs we consider in this section.

We now explore the consequences of imposing additional fairness constraints on our college admissions example, as given by the causal DAG in Figure 1, via a simulation study of one million hypothetical applicants, for one quarter of whom ($b = \frac{1}{4}$) seats are allocated. In particular, in the hypothetical pool of 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. We define the "academic index" $m(x)$ of applicants to be the estimated probability that an applicant will graduate if admitted, based on their observed test score and race. See Appendix C for additional details, including the specific structural equations we use in the simulation.

Each of the panels in Figure 4 illustrates the geometry of fairness constraints for five different formal notions of fairness described in Section 2: counterfactual fairness, path-specific fairness, principal fairness, counterfactual equalized odds, and counterfactual predictive parity. The vertical axes of each panel correspond to aggregate academic index and the horizontal axes to the number of admitted applicants from the target group. The purple lines trace out the boundary of the set of feasible policies, with points on or below the curves achievable by policies that adhere to the budget constraint. Policies lying strictly below the purple curves (or, similarly, on the dashed segments of the purple curves) are "Pareto dominated," meaning that one can find feasible alternatives that are larger on both of the depicted axes (i.e., academic index and diversity). Since we have assumed committee members prefer higher values on each dimension, their effective choice set consists of those policies on the solid purple segments—the "Pareto frontier." Committee members may still disagree over which policy on the frontier to adopt. But for any policy not on the frontier,

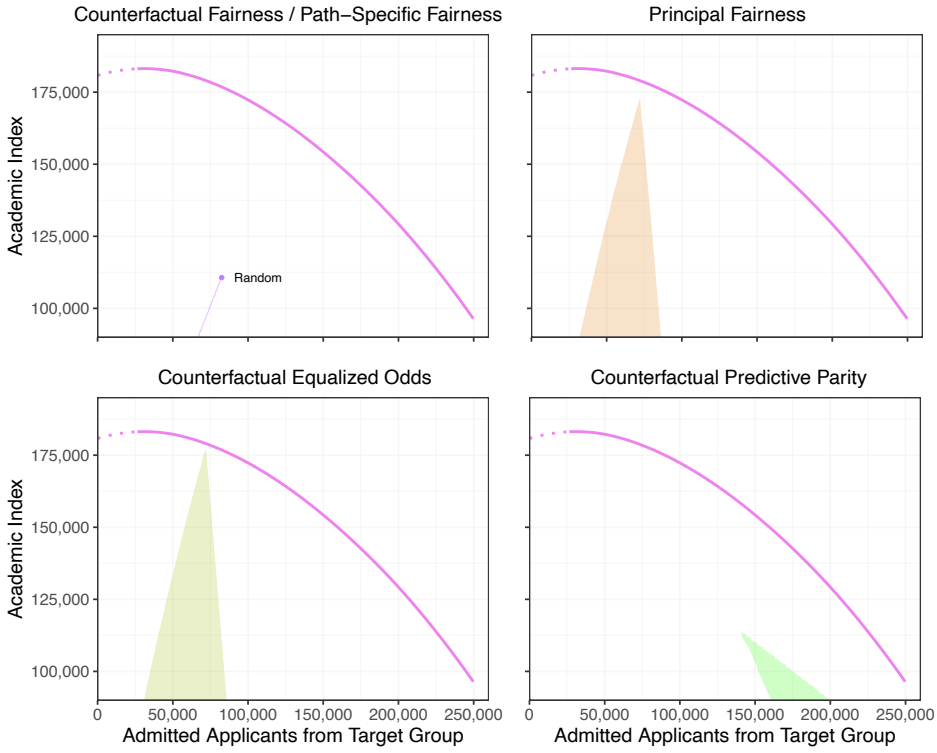


Figure 4: The geometry of fairness constraints, in an illustrative example of college admissions. Points under the purple curve correspond to all feasible policies—given the budget constraint—whereas the shaded regions correspond to feasible policies that satisfy various formal definitions of fairness. (For path-specific fairness, we set Π equal to the single path $A \to E \to T \to D$ highlighted in Figure 1, and set $W = X$.) For each definition, the constrained policies lie strictly under the purple curve, meaning there are alternative, unconstrained, feasible policies that simultaneously achieve greater student-body diversity (indicated by the x-axis) and greater academic preparedness (indicated by the y-axis). The solid segments of the purple lines correspond to policies on the Pareto frontier—for which one cannot simultaneously increase both diversity and academic preparedness. The points labeled "random" in the upper left-hand corner correspond to policies generated by random lotteries in which each individual is admitted with equal probability, in accordance with Theorem 11.

there is a feasible policy above and to the right of it, which is thus preferred by every member of the committee.

Finally, the shaded regions indicate the set of feasible policies constrained to satisfy each of the fairness definitions. (In Appendix B, we show that these feasibility regions can be computed by solving a series of linear programs.) In each case, the constrained regions do not intersect the Pareto frontier, and so there is an alternative, unconstrained feasible policy

that simultaneously achieves more student-body diversity and an overall higher academic index. For example, in the case of policies satisfying counterfactual or path-specific fairness, shown in the upper left panel, the set of feasible policies lie on a single line segment. That structure follows from Theorem 11, since the only policies satisfying either of these notions of fairness in our setting are ones that admit all students with a constant probability, irrespective of their covariates. While not as extreme, the other fairness definitions similarly restrict the space of feasible policies in severe ways, as shown in the remaining panels. These results illustrate that constraining decision-making algorithms to satisfy popular definitions of fairness can have unintended consequences, and may even harm the very groups they were ostensibly designed to help.

Our discussion in this section aimed to highlight the geometry of "fair" decision policies and their consequences in the context of a simple motivating example. We next show that these qualitative findings are guaranteed to hold much more generally.

4.2 A Formal Theory of Fairness in the Presence of Externalities

Our simulation above showed that policies satisfying one of the mentioned fairness definitions are suboptimal, in the sense that they constrain one to a portion of the feasible region in which policies could be improved along both dimensions of interest. As was the case in the absence of trade-offs in Section 3.2, the phenomenon occurring in our simulation is true much more generally. To understand why, we begin by isolating and formalizing the relevant mathematical properties of our example. To generalize our setting in Section 3, we consider arbitrary utility functions of the form $u : \mathcal{X} \to \mathbb{R}$. As before, for a function u and decision policy d, we write $u(d) = \mathbb{E}[d(X) \cdot u(X)]$ to denote the expected utility of decision policy $d(x)$ under the utility u. An important constraint on the admissions committee was the fact that their admissions decisions could not, in expectation, exceed the budget.

Definition 12 *For a budget b, we say a decision policy $d(x)$ is* feasible *if $\mathbb{E}[d(X)] \leq b$.*

A key feature of the college admissions example is that despite some level of uncertainty regarding the "true" utility—i.e., exactly how to trade off between its objectives—the committee knows what its objectives are: to increase the academic index and diversity of the incoming class. One way to encode this kind of uncertainty is to consider a set $\mathcal{U}$ consisting of all "reasonable" ways of trading off between the objectives. While the utilities need not be the same, they should be consistent, in the sense that conditional on an applicant's group membership, all of the utilities should "agree" that a higher academic index is better.

Definition 13 *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 $u'(x_1) > u'(x_2)$.*

A second relevant feature of the admissions problem is that certain policies were strictly better from the admissions committee's perspective, despite their uncertainty about the exact form of their utility. The notion that one policy is better than another *regardless* of the exact form of the utility is formalized by Pareto dominance.

Definition 14 *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.*

As discussed above and in Section 3.1, in the absence of trade-offs, optimal decision policies take the simple form of threshold policies. The existence of trade-offs broadens the range of forms a Pareto efficient policy can take. Even so, for consistent collections of utilities, the Pareto efficient policies take a closely related form.

Proposition 15 *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.:*

$$d(x) = \begin{cases} 1 & u(x) > t_{\alpha(x)}, \\ 0 & u(x) < t_{\alpha(x)}. \end{cases} \tag{11}$$

The proof of Proposition 15 is in the Appendix.[18]

4.2.1 Fairness Definitions with Many Constraints

All of the definitions we study in this section prominently feature causal quantities, but the important quality driving our analysis in this section is that each definition imposes many constraints. For instance, counterfactual equalized odds requires that

$$\Pr(D = 1 \mid A = a, Y(1) = y) = \Pr(D = 1 \mid Y(1) = y)$$

for every outcome y.

Theorem 17 shows that for *almost every* joint distribution of X, $Y(0)$, and $Y(1)$ such that $u(X)$ has a density, any feasible 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 feasible policies satisfying path-specific fairness—including counterfactual fairness—are Pareto dominated. (The analogous statements for counterfactual predictive parity, equalized false positive rates, and demographic parity are not true; we return to this point in Section 4.2.2.) That is, we show that, for a typical joint distribution, any feasible policy satisfying the fairness definitions enumerated above cannot have the form of a multiple-threshold policy. 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 16.

18. 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 multiple-threshold policy d, we can construct a standardized multiple-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 multiple-threshold policies so that applicants at the threshold are admitted with the same group-specific probability.

Definition 16 *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 $g \in \mathcal{G}$.*

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 Proposition 72 in the Appendix for details.

Theorem 17 *Suppose $\mathcal{U}$ is a set of utilities consistent modulo α. Further suppose that for all $a \in \mathcal{A}$ there exist a $\mathcal{U}$-fine distribution of X and a utility $u \in \mathcal{U}$ such that $\Pr(u(X) > 0, A = a) > 0$, where $A = \alpha(X)$. Then,*

- *For almost every $\mathcal{U}$-fine distribution of X and $Y(1)$, any feasible 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 feasible decision policy satisfying conditional principal fairness is strongly Pareto dominated.*
- *If $|\textsc{Img}(\omega)| < \infty$ and there exists a $\mathcal{U}$-fine distribution of X such that $\Pr(A = a, W = w_i) > 0$ 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 feasible decision policy satisfying path-specific fairness is strongly Pareto dominated.*[19]

The proof of Theorem 17 is given in the Appendix. At a high level, the proof proceeds in three steps, which we outline below using the example of counterfactual equalized odds. First, we show that for almost every fixed $\mathcal{U}$-fine joint distribution μ of X and $Y(1)$ there is at most one policy $d^*(x)$ satisfying counterfactual equalized odds that is not strongly Pareto dominated. To see why, note that for any specific y_0, since counterfactual equalized odds requires that $D \perp\!\!\!\perp A \mid Y(1) = y_0$, setting the threshold for one group determines the thresholds for all the others; the budget constraint then can be used to fix the threshold for the original group. Second, we construct a "slice" around μ such that for any distribution ν in the slice, $d^*(x)$ is still the only policy that can potentially lie on the Pareto frontier while satisfying counterfactual equalized odds. We create the slice by strategically perturbing μ only where $Y(1) = y_1$, for some $y_1 \neq y_0$. This perturbation moves mass from one side of the thresholds of $d^*(x)$ to the other. Due to inframarginality, this perturbation typically breaks the balance requirement $D \perp\!\!\!\perp A \mid Y(1) = y_1$ for almost every ν in the slice. Finally, we appeal to the notion of prevalence to stitch the slices together, showing that for almost every distribution, any policy satisfying counterfactual equalized odds is strongly Pareto dominated. Analogous versions of this general argument apply to the cases of conditional principal fairness and path-specific fairness.[20] We note that the conditions of Theorem 17

19. 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 $u \in \mathcal{U}$, i.e., component-wise application of u to elements of $\mathcal{X}^{\mathcal{A}}$. In other words, the requirement is that the joint distribution of the $u(X_{\Pi,A,a})$ has a density.

20. This argument does not depend in an essential way on the definitions being causal. In Corollary 70 in the Appendix, we show an analogous result for the non-counterfactual version of equalized odds.

are sufficient, rather than necessary, meaning that the conclusion of the theorem may—and, indeed, we expect will—hold even in some cases where the conditions are not satisfied. In particular, we note that this proof technique prevents the conditions of Theorem 17 from holding when A factors through W and, in particular, when $W = X$. Although, when $X = W$, Theorem 11 shows that under slightly different conditions, a much stronger result holds.

To bring our discussion full circle, we now map Theorem 17 onto the motivation offered in Section 4.1. Recall that the admissions committee knew that given the opportunity, it preferred policies that increased both the overall academic index of its admitted class, and policies that resulted in more students being admitted from the target group. In other words, we imagine that members of the admissions committee have utilities u^* of the form[21]

$$u^*(d) = v\left(\mathbb{E}[m(X) \cdot d(X)], \mathbb{E}[\mathbb{1}_{\alpha(X)=a_1} \cdot d(X)]\right), \tag{12}$$

where, as above, $m(x)$ denotes the academic index of an applicant with covariates $X = x$, and v increases in both coordinates. Corollary 18 establishes the inherent incompatibility of such preferences with the formal fairness criteria we have been considering.

Corollary 18 *Consider a utility of the form given in Eq.* (12)*, where v is monotonically increasing in both coordinates and $m(x) \geq 0$. Then, under the same hypotheses as in Theorem 17,*[22] *for almost every joint distribution, no utility-maximizing decision-policy satisfies counterfactual equalized odds, conditional principal fairness, or path-specific fairness.*

Lastly, while, in general, one's decision policy can depend only on the covariates known at the time of the decision, in some cases, the restriction that $u(x)$ be a function of $x \in \mathcal{X}$ alone may be too restrictive; the connection between an individual having covariates $X = x$ and our utility may depend also on the relationship between X and Y. For instance, in the admissions example, the admissions committee may value high test scores and extracurriculars not, e.g., as *per se* measures of academic merit, but rather instrumentally insofar as they are connected to whether an applicant will eventually graduate. However, allowing u to depend on both x and y greatly complicates the underlying geometry of the problem. Proving Theorem 17 in this more general setting remains an open problem. However, intuition from finite-dimensions—where more powerful measure-theoretic tools are available—suggests that the result remains true in the more general setting. For example, Proposition 19 presents a version of this result over a natural, finite-dimensional family of distributions.

Proposition 19 *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 $r(x) = \mathbb{E}[Y(1) \mid X = x]$. For almost every $(\alpha_0, \beta_0, \alpha_1, \beta_1) \in \mathbb{R}^4_{>0}$, if the conditional distributions of $r(X)$ given A are beta distributed with

$$r(X) \mid A = a_i \sim \textsc{Beta}(\alpha_i, \beta_i),$$

21. Strictly speaking, we are saying that members of the admissions committee, rather than having an aggregate utility—which, as we have considered so far, has the form $\mathbb{E}[u(X) \cdot d(X)]$—has a utility on aggregate *outcomes*.

22. The full statement is given in Appendix F.7.

then any policy satisfying counterfactual equalized odds is strongly Pareto dominated.

4.2.2 Fairness Definitions with Few Constraints

We conclude this analysis by considering equalized false positive rates, demographic parity, and counterfactual predictive parity. These fairness notions are less demanding than the notions considered above, in that they introduce only "one" additional constraint, e.g., that $Y(1) \perp\!\!\!\perp A \mid D = 0$, in the case of counterfactual predictive parity. Since the budget introduces a second constraint, and the form of a multiple-threshold policy allows for a degree of freedom in each group, the number of constraints and the number of degrees of freedom are equal—as opposed to the causal fairness definitions covered by Theorem 17, in which the constraints outnumber the degrees of freedom. As such, it is possible in some instances to have a policy on the Pareto frontier that satisfies these conditions; though see Section 5.3 for discussion about why such policies are still often at odds with broader goals.

However, it is not *always* possible to find a point on the Pareto frontier satisfying these definitions. In Proposition 20, we show that counterfactual predictive parity cannot lie on the Pareto frontier 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 Appendix H.3.

Proposition 20 *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 $\mu_{a_0} > \mu_{a_1}$ and $v > 0$.[23] *Then any policy satisfying counterfactual predictive parity is strongly Pareto dominated.*

5. A Path Forward

We have thus far worked to clarify some of the statistical limitations of existing mathematical definitions of fairness. We have argued that in many cases of interest, these definitions can ultimately do more harm than good, hurting even those individuals that these notions of fairness were ostensibly designed to help.

We end on a more optimistic note, charting out a potential path toward designing more equitable algorithms. To do so, we start, in Section 5.1 by reviewing conceptions of discrimination in law and economics, and, in particular, we contrast process-oriented and outcome-oriented notions of fairness. Whereas the computer science literature is dominated by process-oriented, *deontological* definitions of fairness, we see more promise in adopting

23. 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 $v = \alpha + \beta$.

an outcome-oriented, *consequentialist* approach represented by the utilitarian analysis we have described above. In Section 5.2, we enumerate and discuss four issues that we feel are critical in developing equitable algorithms: (1) balancing inherent trade-offs in decision problems; (2) assessing calibration; (3) selecting the inputs and targets of prediction; and (4) designing data collection strategies. Finally, in Section 5.3, we illustrate how to grapple with these considerations in a case study of complex medical care, motivated by work from Obermeyer et al. (2019).

5.1 Competing Notions of Ethical Decision Making: Process vs. Outcomes

There are many distinct but related understandings of ethical decision making in law, economics, philosophy, and beyond. One key dimension on which we organize these notions is the extent to which they consider the *process* through which decisions are made versus the *outcomes* that those decisions render.

The dominant legal doctrine of discrimination in the United States treats explicit race- and gender-based decisions with heightened scrutiny. The Equal Protection Clause of the U.S. Constitution's Fourteenth Amendment restricts government agencies from adopting policies that explicitly reference legally protected categories, and myriad federal and state *disparate treatment* statutes similarly constrain a variety of private actors. Conversely, policies that do not explicitly consider legally protected traits—or obvious proxies—are generally deemed not to violate disparate treatment principles. Formally, it is lawful to use legally protected attributes in a limited way to further a compelling government interest, but, in practice, such exceptions are few and far between. Until recently, the prime example of a race-conscious policy passing legal muster was affirmative action in college admissions (Fisher v. University of Texas, 2016). However, in 2023, the U.S. Supreme Court barred the explicit consideration of race in admissions decisions (SFFA v. Harvard, 2023).

Disparate treatment doctrine has evolved over time, and reflects ongoing debates about the role of *classification* (use of protected traits, a process-oriented, deontological notion) versus *subordination* (subjugation of disadvantaged groups, an outcome-oriented notion) in discrimination cases (Fiss, 1976). Some legal scholars have argued that courts, even when formally applying anti-classification criteria, are often sympathetic to the potential effects of judgments on social stratification, indicating tacit concern for anti-subordination (Balkin and Siegel, 2003; Colker, 1986; Siegel, 2003). Others, though, have noted that such judicial support for anti-subordination appears to be waning (Nurse, 2014). At a high level, we thus view modern disparate treatment law as primarily interested in process over outcomes, though these debates illustrate that the two concepts cannot be perfectly separated.

In contrast to process-oriented disparate treatment principles, the economics literature distinguishes between two outcome-focused, consequentialist rationales for explicitly considering race, gender, and other protected traits: taste-based and statistical. With *taste-based discrimination* (Becker, 1957), decision makers act as if they have a preference or "taste" for bias, sacrificing profit to avoid certain transactions. This includes, for example, an employer who forfeits financial gain by failing to hire exceptionally qualified minority applicants. But, in contrast to legal reasoning, the economic argument against taste-based discrimination is not that decisions are based on race *per se*, but rather because consideration of race leads to worse outcomes: a loss of profit. With *statistical discrimination* (Arrow, 1973; Phelps,

1972), decision makers explicitly consider protected attributes in order to optimally achieve some non-prejudicial goal. For example, profit-maximizing auto insurers may charge a premium to male drivers to account for gender differences in accident rates. Despite their differing outcome-based justifications, both taste-based and statistical discrimination are often considered legally problematic as they explicitly consider race, in violation of disparate treatment laws.[24]

As the above insurance example and our running diabetes example illustrate, one might consider it acceptable to base decisions in part on legally protected traits when doing so leads to good outcomes. Conversely, whereas process-oriented disparate treatment principles generally deem race-blind policies acceptable, one might declare such blind policies problematic if they lead to bad outcomes. Indeed, under the statutory *disparate impact* standard, a practice may be deemed discriminatory if it has an unjustified adverse effect on legally protected groups, even in the absence of explicit categorization (Barocas and Selbst, 2016).[25] The disparate impact doctrine was formalized in *Griggs v. Duke Power Co.*, a 1971 U.S. Supreme Court case. In 1955, the Duke Power Company mandated that employees have a high school diploma to be considered for promotion, which, in practice, severely limited the eligibility of Black employees. The Court found that this facially race-neutral requirement had little relation to job performance, and accordingly deemed it to have an unjustified—and illegal—disparate impact. The Court noted that the employer's motivation for instituting the policy was irrelevant to its decision; even if enacted without discriminatory purpose, the policy was deemed discriminatory in its effects and hence illegal. However, disparate impact law does not prohibit *all* group differences produced by a policy—the law only prohibits *unjustified* disparities. For example, if, hypothetically, the high-school diploma requirement in *Griggs* were shown to be necessary for job success, the resulting disparities would be legal.

On the spectrum from process- to outcome-based understandings of discrimination, we view the formal, axiomatic fairness definitions described in Section 2 as reflecting a largely process-based orientation. Blinding and its more stringent causal variants—counterfactual fairness and path-specific fairness—can be viewed as descendants of disparate treatment considerations, as they seek to remove the effects of race and other protected attributes on decisions. The remaining definitions—for example, those that aim to equalize error rates across groups—do explicitly reference an outcome Y, but they do so in a way that seems largely disconnected from the consequences one might naturally consider. As we have

24. In the case of auto insurance specifically, some states (including California, Hawaii, Massachusetts, Maine, Michigan, North Carolina, and Pennsylvania)—though not all—have barred the use of gender in pricing policies.

25. The legal doctrine of disparate impact stems largely from federal statutes, not constitutional law, and applies only in certain contexts, such as employment (via Title VII of the 1964 Civil Rights Act) and housing (via the Fair Housing Act of 1968). Apart from federal statutes, some states have passed more expansive disparate impact laws, including Illinois and California. The distinction between statutory and constitutional rules is particularly relevant here, as there is debate among scholars over whether disparate impact laws violate the Equal Protection Clause and are thus unconstitutional (Primus, 2003). There is also debate over whether disparate impact law is motivated primarily by an interest in banning bad *outcomes*, or seeks to provide an alternative pathway for ferreting out bad *intent*, when actors may mask animus with race-neutral policies (Watson v. Fort Worth, 1988). But, regardless of its underlying justification, disparate impact law is formally focused on outcomes, not intent or classification, and so we view it as an outcomes-focused principle.

argued, whether error rates are equal across groups has more to do with the structure of group-specific risk distributions than with whether decisions lead to good or bad outcomes for group members. For example, in our college admissions example, enforcing various formal notions of fairness would, in theory, typically lead to student bodies that are both less diverse and less academically prepared than those resulting from feasible alternatives not constrained to satisfy these notions.

One might, on principle, favor certain process-based understandings of discrimination over outcome-based notions. One might even adopt a meta-consequentialist position, and argue that procedural considerations (e.g., ensuring medical decisions are blind to race) engender trust and in turn bring about better downstream outcomes. In many cases, though, the ethical underpinnings of popular mathematical definitions of fairness have not been clearly articulated. Absent such justification, we advocate for an approach that more directly engages with the real-world costs and benefits of different decision policies, a perspective that we outline in more detail in the remaining sections.

5.2 Designing Equitable Algorithms

A key advantage of the dominant axiomatic approach to algorithmic fairness is that it can be readily applied across contexts, with little domain-specific knowledge. One can build automated tests to check whether any predictive algorithm satisfies various formal fairness desiderata, and even automatically modify algorithms to ensure that they do satisfy a specific fairness criterion (e.g., Cotter et al., 2019; Weerts et al., 2023). But, as we have argued, this approach is often at odds with improving well-being, including for disadvantaged groups. A particularly pernicious risk of automated, axiomatic approaches is that they can make invisible the cost to well-being: automatically constraining algorithms to be "fair" can lead one to overlook unconstrained alternatives that are clearly preferable. We have instead called for a more careful analysis of the consequences, good and bad, of different decision policies, selecting the appropriate course of action based on the specific context. This is admittedly hard to do—and does not easily scale—but there are general principles that we believe are helpful to keep in mind when navigating this terrain. Below we enumerate and discuss four of them.

5.2.1 Contending with Inherent Trade-Offs

There are inherent trade-offs in many important decision problems. For instance, in our college admissions example, one must balance academic preparedness with student-body diversity. Although one cannot generally circumvent these trade-offs, we believe it is useful to explicitly enumerate the primary dimensions of interest and to acknowledge the trade-offs between them. In some cases, like our stylized admissions example, one might be able to explicitly calculate the Pareto frontier shown in Figure 4, in which case it often makes sense to focus on those policies lying on the frontier. In many cases, it won't be possible to compute the frontier. Still, by listing and discussing trade-offs, even informally, one can reduce the risk of adopting clearly problematic policies, like those that typically result from uncritically constraining decisions to satisfy formal fairness criteria.

In this sense, designing equitable algorithms is akin to designing equitable policy writ large. One might accordingly adapt democratic mechanisms used to draft and enact leg-

islation to algorithm design. For example, adopting such a policy-oriented perspective, Chohlas-Wood et al. (2023a) and Koenecke et al. (2023) surveyed a diverse sample of Americans to elicit preferences on how best to balance competing objectives in programs that algorithmically allocate government benefits.

5.2.2 Assessing Calibration

When designing or auditing a risk assessment algorithm, it is important to check whether predictions are *calibrated*, meaning that risk scores correspond to the same observed level of risk across groups. In general, the relationship between predictors and outcomes may plausibly differ across groups, leading to miscalibrated risk estimates—what Ayres (2002) calls the problem of *subgroup validity*. Figure 3 shows instances of risk scores for diabetes and recidivism that are miscalibrated across race and gender, respectively. For example, a nominal 1.5% diabetes risk—based on age and BMI—corresponds to an actual, observed diabetes rate of approximately 1% among White patients and 3% among Asian patients. Similarly, among individuals receiving a COMPAS risk score of 7—based on criminal history and related factors—about 55% of women recidivate, compared to 65% of men. These miscalibrated risk scores can result in inequitable decisions. For instance, a policy to screen patients with a nominal diabetes risk of 1.5% or above—in line with existing medical recommendations (Aggarwal et al., 2022)—would overscreen White patients and underscreen Asian patients, harming individuals in both groups.

Calibration can often be visually assessed by plotting predicted risk against average outcomes, as in Figure 6 (cf. Arrieta-Ibarra et al., 2022). For a simple, more quantitative, measure, we recommend regressing observed outcomes against risk estimates and group membership. A coefficient of approximately zero on group membership suggests risk estimates correspond to similar average outcomes across groups, with deviations from zero indicating the degree of miscalibration.

In practice, miscalibration can often be rectified by training group-specific risk models, or, roughly equivalently, including group membership in a single risk model fit across groups. For example, diabetes risk models that include race, and recidivism risk models that include gender, are approximately calibrated. The relatively new literature on multicalibration has introduced new computational techniques to ensure predictions are simultaneously calibrated across many different subgroups (Hébert-Johnson et al., 2018). Of course, including protected traits in risk models raises additional legal and ethical challenges. In some cases, it may be possible to reduce or eliminate miscalibration by incorporating additional, non-protected covariates. Regardless, we believe it is important to check the calibration of risk scores to make informed decisions about if and how to address any observed disparities.

Calibration is an important necessary condition to ensure risk estimates correspond to actually observable levels of risk across groups. But it is not sufficient. Indeed, even calibrated risk scores can encode and reinforce deeply discriminatory policies. To see this, imagine a bank that wants to discriminate against Black applicants. Further suppose that: (1) within ZIP code, White and Black applicants have similar default rates; and (2) Black applicants live in ZIP codes with relatively high default rates. Then the bank can surreptitiously discriminate against Black borrowers by basing estimates of default risk only on an applicant's ZIP code, ignoring all other relevant information. Such scores would be

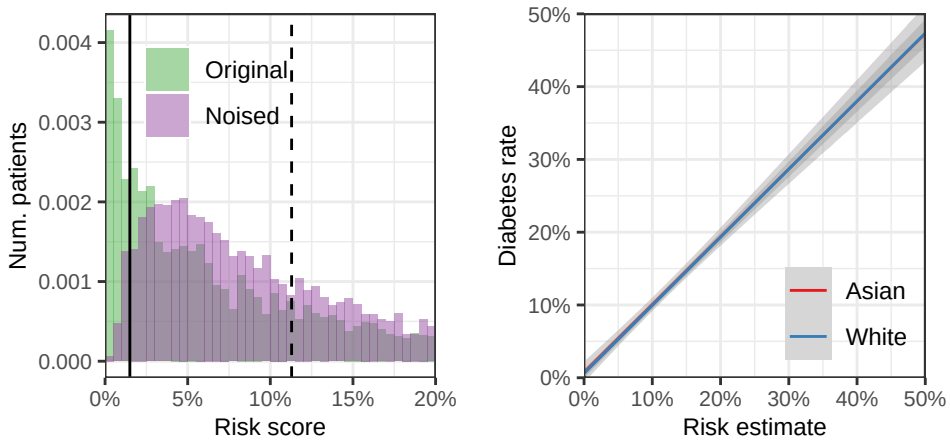


Figure 5: Calibration is insufficient to prevent discrimination. *Left:* The distribution in green shows diabetes risk for Asian patients based on accurately collected age and BMI, and the distribution in purple shows estimates when the risk model is trained on noisy inputs. Estimates under the noisy model concentrate around the mean (dashed vertical line), pushing more Asian patients above the screening threshold (solid vertical line). *Right:* A calibration plot comparing noisy risk estimates for Asian patients and accurate risk estimates for White patients. The calibrated risk scores can mask both intentional discrimination and inadvertent errors.

calibrated (White and Black applicants with the same score would default equally often), and the bank could use these scores to justify denying loans to nearly all Black applicants. The bank, however, would be sacrificing profit by refusing loans to creditworthy Black applicants,[26] and thus engaging in taste-based discrimination. This discriminatory lending strategy is indeed closely related to the historical (and illegal) practice of redlining, and illustrates the limitations of calibration as a measure of equity.

Figure 5 shows another example of calibrated scores masking disparities. In the left-hand panel, we plot in green the distribution of diabetes risk for Asian patients, as estimated from age, BMI, and race. In purple, we plot the distribution of estimated risk when age and BMI are imperfectly measured for Asian patients in the training data. Training the risk model on noisy features pushes risk estimates toward the mean. As a result, based on the noisy risk model, 97% of Asian patients are above the 1.5% screening threshold, compared to 81% of Asian patients under the more accurately estimated model—leading to more medically unnecessary screening under the noisy model. Importantly, however, the noisy model is still calibrated, as shown in the right-hand panel of Figure 5, where we compare risk scores for Asian patients estimated from the noisy predictors and risk scores for White patients estimated from the accurately estimated information. In theory, a

26. These applicants are creditworthy in the sense that they would have been issued a loan had the bank used all the information it had available to determine their risk.

malicious algorithm designer could generate such calibrated but inaccurate scores to intentionally harm Asian patients. In practice, this pattern could equally arise from negligence rather than malice. These examples illustrate the importance of considering all available data when constructing statistical risk estimates; assessments that either intentionally or inadvertently ignore predictive information may facilitate discriminatory decisions while satisfying calibration—though, as we discuss below, even this intuitive heuristic of "use all the data" has its limitations.

5.2.3 Selecting the Target of Prediction

In constructing algorithmic risk scores, a key ingredient is the target of prediction. In practice, though, there is often a mismatch between our true outcome of interest and the available data—an occurrence we call *label bias* (Zanger-Tishler et al., 2023). As with the other issues we discuss, there is typically no perfect solution to this problem, but there are ways to mitigate it.

For example, in pretrial risk assessment, we would often like to estimate the likelihood a defendant would commit a crime if released. But there are two key difficulties with this goal. First, though we might want to measure crime conducted by defendants awaiting trial, we typically only observe crime that results in a conviction or an arrest. These observable outcomes, however, are imperfect proxies for the underlying criminal act. Further, heavier policing in communities of color might lead to Black and Hispanic defendants being arrested, and later convicted, more often than White defendants who commit the same offense (Lum and Isaac, 2016). Poor outcome data might thus cause one to systematically underestimate the risk posed by White defendants. The second, related, issue is that our target of interest is a *counterfactual* outcome; it corresponds to what would have happened had a defendant been released. In reality, we only observe what actually happened conditional on the judge's actual detention decision.

One way to reduce label bias in this case is to adjust the target of interest. For example, criminologists have found that arrests for violent crime—as opposed to drug crime—may suffer from less racial bias.[27] In particular, Skeem and Lowenkamp (2016) note that the racial distribution of individuals arrested for violent offenses is in line with the racial distribution of offenders inferred from victim reports, and is also in line with self-reported offending data. In other cases, like lending, where one may seek to estimate default rates, the measured outcome (e.g., failure to pay) corresponds more closely to the event of interest. The problem of estimating counterfactuals can likewise be partially addressed in some applications. In the pretrial setting, Angwin et al. (2016) measure recidivism rates in

27. D'Alessio and Stolzenberg (2003) find evidence that White offenders are even somewhat more likely than Black offenders to be arrested for certain categories of crime, including robbery, simple assault, and aggravated assault. Measurements of minor criminal activity, like drug offenses, are more problematic. For example, there is evidence that drug arrests in the United States are biased against Black and Hispanic individuals, with racial minorities who commit drug crimes substantially more likely to be arrested than White individuals who commit the same offenses (Ramchand et al., 2006). Although this pattern is well known, many existing risk assessment tools still consider arrests or convictions for *any* new criminal activity—including drug crimes—which may lead to biased estimates. As another example of label bias, auto insurance rates are determined in part by a driver's record of receiving speeding tickets, but disparities in police enforcement mean that tickets are biased proxies of dangerous driving behavior (Cai et al., 2022b).

the first two-year period during which a defendant is not incarcerated; this is not identical to the desired counterfactual outcome—since the initial detention may be criminogenic, for example—but it seems like a reasonable estimation strategy. Further, unaided human decisions often exhibit considerable randomness, a fact that can be exploited to facilitate statistical estimation of counterfactual outcomes (Jung et al., 2020b; Kleinberg et al., 2017a). More generally, a spate of recent work at the intersection of machine learning and causal inference (Hill, 2011; Jung et al., 2020c; Mullainathan and Spiess, 2017) offers hope for more gains in counterfactual estimation.

5.2.4 Collecting Training Data

A final issue we discuss is collecting suitable training data for risk assessment algorithms to mitigate the effects of sample bias. Ideally, one would train algorithms on data sets that are broadly representative of the populations on which they are ultimately applied—though there are subtleties to this heuristic that we describe below. While often challenging in practice, failure to train on representative data can lead to unintended, and potentially discriminatory, consequences. For example, Buolamwini and Gebru (2018) found that commercial facial analysis tools struggle to correctly classify the gender of dark-skinned individuals—and of dark-skinned women in particular—a disparity likely attributable to the relative dearth of dark-skinned faces in facial analysis data sets. Similarly, Koenecke et al. (2020) found that several popular automated speech recognition systems were significantly worse at transcribing Black speakers than White speakers, likely due to insufficient data from speakers of African American Vernacular English (AAVE), a variety of English spoken by many Black Americans.

The problems of non-representative data can be even more acute in the case of risk assessment algorithms, especially when the target of interest is a causal quantity. For instance, in our running college admissions example, we seek to estimate how a student would (counterfactually) perform if admitted. Historical data are typically the result of past, potentially biased, decisions, and so may not fully generalize. Imagine, for example, that predictions of college performance are informed by where an applicant went to high school, and that, historically, only applicants from certain high schools were accepted—and we consequently only see outcomes for students from those high schools. Then we would expect less accurate predictions for students from those absent high schools. In general, regression to the mean could attenuate estimates for high achieving students who differ from those previously accepted, potentially reinforcing existing admissions practices.

As a general heuristic, we believe it is advisable to train models on representative data, but, as Cai et al. (2022a) note, the optimal sampling strategy depends on the statistical structure of the problem and the group-specific costs of collecting data. Interestingly, the value of representative data collection strategies depends in part on the degree to which race, gender, and other protected attributes are predictive. In theory, if protected attributes are not predictive, one could build an accurate risk model using only examples from one particular group (e.g., White men). Given enough examples of White men, the model would learn the relationship between features and risk, which by our assumption would generalize to the entire population. This phenomenon highlights a tension in informal discussions of fairness, with some advocating both for representative training data and for the exclusion of

protected attributes. However, representative data are often most important precisely when protected attributes add information, in which case their use is arguably more justified. Even if protected attributes are not predictive, representative data can still help in two additional ways. First, a representative sample ensures that the full support of features is present at training time, as it is possible that the distribution of features varies across groups, even if the connection between features and outcomes does not. We note, though, that one might have adequate support even without a representative sample in many real-world settings, particularly when models are trained on large data sets and the feature space is relatively low dimensional. Second, a representative sample can help with model validation, allowing one to assess the potential effects of group imbalance on model fit. In particular, without a representative sample, it can be difficult to determine whether a model trained on a single group generalizes to the entire population.

In many settings, one may be able to gather better training data with greater investment of time and money. For example, in our diabetes example one could aim to collect more complete medical records, a process that may be both costly and logistically difficult. In theory, this additional information may lead to welfare gains, and policymakers must accordingly evaluate the relative costs and benefits to all groups of exerting this extra effort when designing algorithms. Fortunately, in practice, there are often diminishing returns to information, with a relatively short list of key features providing most of the predictive power (Jung et al., 2020b), at least partially mitigating this concern. Carefully documenting data provenance and content—including what features the data contain, how observations were selected into the sample, potential missingness or measurement error, and the representation of protected classes in the data—with the use of datasheets or other structured tools can both aid in determining the extent of possible issues of non-representative data and in alerting future analysts to potential issues (Arnold et al., 2019; Chmielinski et al., 2022; Gebru et al., 2021; Holland et al., 2020; Stoyanovich and Howe, 2019).

As with the other issues we have discussed, there is no universal solution to data collection. It might, for example, simply be prohibitive in the short run to train models on the data set one would ideally like to use. Nevertheless, as in all situations, one must carefully weigh the potential costs and benefits of adopting a necessarily imperfect risk assessment algorithm relative to the other possible options. In particular, even an imperfect algorithm may in some circumstances be better than leaving decisions to similarly imperfect humans who have their own biases.

5.3 Case Study: An Algorithm to Allocate Limited Medical Resources

We conclude by illustrating the principles discussed above with a real-world example: an algorithm for referring patients into a "high-risk care management" program, previously considered by Obermeyer et al. (2019). The care management program more effectively aids patients with complex medical needs, in principle both improving outcomes for patients and reducing costs to the medical system for patients who are enrolled. But the program has limited capacity, which we formalize by assuming that only 2% of patients can be enrolled

(i.e., we set $b = \frac{1}{50}$). For our analysis, we use the data released by Obermeyer et al.,[28] which contain demographic variables, cost information, comorbidities, biomarker and medication details, and health outcomes for a population of approximately 43,000 White and 5,600 Black primary care patients at an academic hospital from 2013–2015.

As a first step toward identifying patients to enroll in the program, one could train a model predicting the healthcare resources a patient is likely to require over the next year. Sicker patients require more care and, consequently, incur greater healthcare costs. Thus, our initial approach is to predict how likely a patient is to be "high cost"—which we operationalize as being in the top decile of healthcare expenditures in a given year—based on the available information. (One of the main contributions of Obermeyer et al. (2019) is highlighting that healthcare costs is a problematic outcome due to label bias, a point we return to shortly.)

As discussed above, it is useful to assess the calibration of risk assessment algorithms across groups. In particular, while calibration across race is largely guaranteed if race is included as a predictor in the statistical model, race-blind models are often preferred, particularly in healthcare, in part to avoid perceptions of bias. As a result, we assess the calibration of a race-blind model trained on all available information except for race, shown in Figure 6. Unlike in the diabetes screening example considered in Section 3.2, the race-blind healthcare cost predictions are calibrated across race groups, meaning that risk estimates largely match observed costs across groups. It accordingly appears that race provides little marginal predictive power in this example, assuaging potential concerns with its omission.

We turn next to assessing the effects of applying formal fairness criteria to our enrollment decisions. Often, in healthcare contexts, a false negative—e.g., failing to screen for a disease when it is present—is more consequential than a false positive—e.g., screening for a disease when it is not present. For this reason, one might seek to ensure enrollment decisions are fair by mandating false negative rates be equal across race groups (e.g., Seyyed-Kalantari et al., 2021),[29] i.e., requiring that $A \perp\!\!\!\perp D \mid Y = 1$. In our example, equalizing false negative rates means that among patients who ultimately incur high medical costs ($Y = 1$), the same proportion ($\mathbb{E}[D]$) are referred into the program across race groups (A).

In our setting, approximately 1,000 patients can be referred into the care management program—2% of the roughly 50,000 patients in our data set. When we equalize false-negative rates, 747 of the enrolled patients ultimately incur high costs, and 113 enrolled patients are Black.[30] However, an unconstrained decision rule (i.e., one that enrolls the patients most likely to incur high costs) enrolls both more high-cost patients (758) and more Black patients (205). In this example, we end up providing worse care to Black patients when we constrain our algorithm to satisfy the formal, mathematical fairness criterion.

28. Obermeyer et al. released a synthetic data set closely mirroring the real data set, available at: `https://gitlab.com/labsysmed/dissecting-bias`. In contrast to $b = 2\%$, which we adopt to better illustrate some of the statistical phenomena we discuss in this section, Obermeyer et al. use a budget of $b = 3\%$.

29. The authors describe their metric of concern as the "false positive rate," where the positive prediction is understood as a "no finding" label, e.g., not having a disease. In our example, we follow the more common convention that a "positive" classification is assigned to the rare event (i.e., being high cost), and so we call this metric the "false negative rate."

30. Following Corbett-Davies et al. (2017), we equalize false-negative rates in a manner that maximizes the number of enrolled patients who ultimately incur high costs.

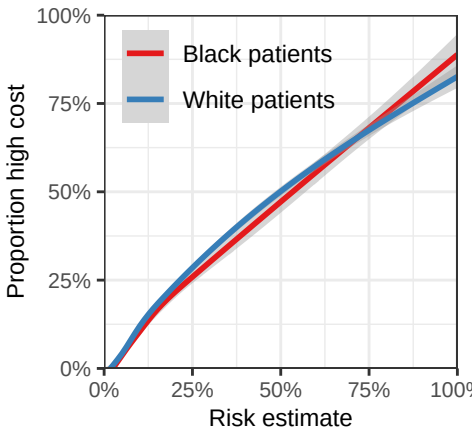
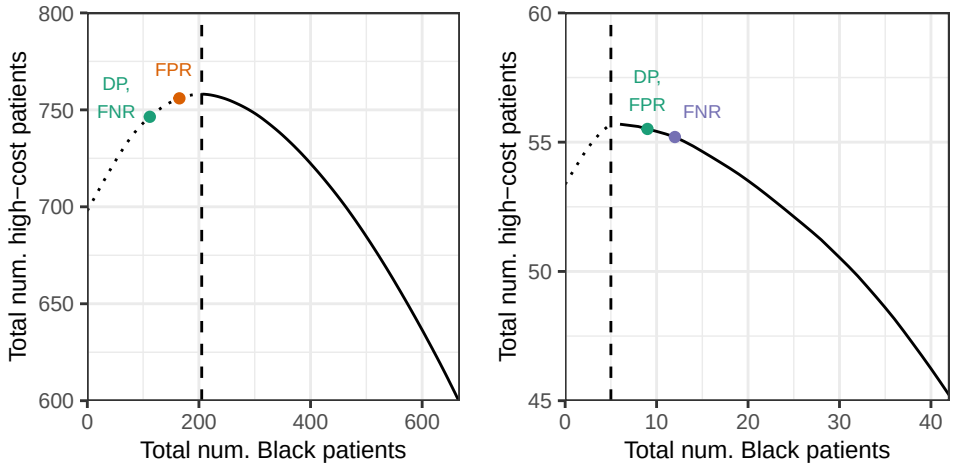


Figure 6: Calibration of a race-blind model predicting whether a patient is "high-cost". The lack of a gap between the proportion of patients who are actually high-cost and the proportion predicted to be high-cost for both groups indicates that the model is well calibrated across race groups.

Instead of applying such mathematical fairness criteria, we advocate for directly weighing the costs and benefits of different decision policies. In Figure 7a, we show the Pareto frontier for our example, tracing out policies that optimally trade off the demographic composition of the enrolled population with the number of enrolled patients who in reality incur high costs. The green point corresponds to equalizing false negative rates, and is to the left of the dashed vertical line that corresponds to the unconstrained decision rule—visually illustrating how constraining our algorithm leads to fewer resources for Black patients. Also shown on the plot are points corresponding to demographic parity and equal false positive rates, both of which likewise lead to fewer resources for Black patients.[31]

The result in Figure 7a stems from the false negative rate for Black patients being lower than the false negative rate for White patients in the unconstrained algorithm—a pattern we expect since Black patients are more likely than White patients to incur high medical costs in our data. Equalizing false negative rates thus means raising the enrollment bar for Black patients and lowering the bar for White patients. In light of this example, one might argue for applying formal fairness criteria only when error rates for racial minorities are higher than for White individuals. Figure 7b repeats our analysis above for the subset of women between the ages of 25 and 34, a subpopulation in which Black patients have higher error rates than White patients. In this case, equalizing false positive rates or false negative rates, or enforcing demographic parity all result in more Black patients being admitted

31. As discussed in Section 4, with the exception of counterfactual predictive parity, all of the remaining fairness definitions given in Section 2 are known *a priori* to restrict one to enrollment policies that will lower *both* the number of Black patients enrolled as well as the number of truly high-cost patients enrolled.



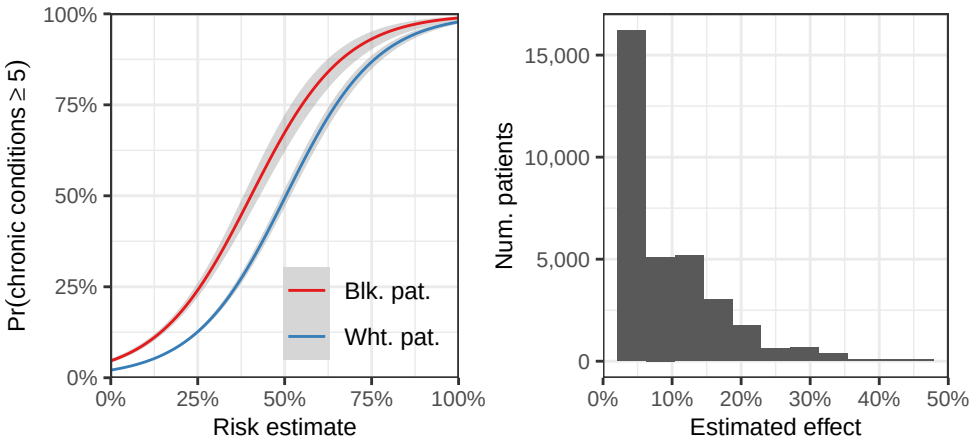
(a) Full population

(b) Women, aged 25–34

Figure 7: Enforcing formal fairness criteria can harm marginalized groups. Feasible regions for admissions policies to a high-risk care management program, where the dashed line indicates the number of Black patients admitted by the policy admitting the maximal number of high-cost patients. *Left:* The Pareto frontier for all patients in the population. Because more Black patients incur high medical costs in the population as a whole, equalizing false negative rates (FNR)—as well as false positive rates (FPR), or enforcing demographic parity (DP)—results in fewer Black patients being admitted than under the policy that maximizes the total number of high-cost patients admitted. (Equalized false negative rates and demographic parity are achieved at the same point.) *Right:* The Pareto frontier for the subpopulation of women between the ages of 25 and 34. Within this subpopulation, Black patients incur lower medical costs, and so equalizing false positive rates, false negative rates, or achieving demographic parity all result in more Black patients being admitted to the high-risk management program than the policy that admits the maximum number of high-cost patients.

into the program. It is, however, unclear why one should adopt those particular error-rate equalizing policies over any other. The policies on the Pareto frontier (i.e., on the curve to the right of the dashed line) are all arguably reasonable to consider. It is admittedly difficult to determine which of these policies to adopt, but we believe it is best to confront this challenge head on, recognizing the hard trade-offs inherent to the problem.

We conclude our case study by considering label bias, the primary concern identified by Obermeyer et al. (2019) in this context. As those authors noted, medical cost is a poor proxy for medical need, and so allocating healthcare resources to minimize anticipated costs can lead to severe disparities. Replicating an analysis by Obermeyer et al., Figure 8a shows that among patients with similar likelihood of incurring high-cost medical care, Black patients are considerably more likely than White patients to have complex medical needs,



(a) Label bias

(b) Heterogeneous treatment effects

Figure 8: The target of prediction impacts equity. *Left:* The probability of having complex medical needs (i.e., at least five active chronic conditions) for Black and White patients as a function of their estimated likelihood of incurring high medical costs, reproducing an analysis by Obermeyer et al. (2019). The large gap across groups indicates that Black patients have greater medical need than White patients with similar anticipated healthcare costs. *Right:* Distribution of the estimated change in the probability that an individual will have complex medical needs after enrolling in the care management program, showing that the extent to which enrollment reduces complex medical needs varies considerably across individuals.

operationalized as having five or more active chronic conditions. This gap is likely a consequence of worse access to healthcare among Black patients, due to a mix of socioeconomic factors and discrimination. To the extent that care management programs aim to aid the sickest patients—as opposed to simply reducing costs—targeting resources based on anticipated costs can lead to inefficient and inequitable outcomes. Obermeyer et al. accordingly suggest switching the target of prediction from medical costs to health status.

It is possible to achieve further gains by recognizing resource allocation as an inherently causal problem. In particular, one may seek to enroll patients in the program in a manner that maximizes $\mathbb{E}[Y(D)]$, where $Y(D)$ is the potential health outcome under the enrollment decision. To do so, we could prioritize patients by their estimated treatment effect $\hat{Y}_i(1) - \hat{Y}_i(0)$, rather than an estimate $\hat{Y}_i$ of their future health status that ignores the causal effect of the program.[32] A proper causal analysis is, in general, a complex topic requiring careful

32. In many analyses that do not explicitly grapple with the causal effects of interventions, the estimand Y is further corrupted by the fact that some patients are expected to be enrolled in the program. As a result, some of the sickest patients may not be prioritized for care, since their expected outcome already incorporates the fact that they would have been enrolled, leading to a prediction paradox. To avoid this situation, one could explicitly estimate and prioritize patients by $Y_i(0)$, the potential outcome in the absence of care. In this case, however, the allocation decisions do not necessarily lead to the largest

treatment beyond the scope of this article. Nonetheless, Figure 8b shows the distribution of $\hat{Y}_i(1) - \hat{Y}_i(0)$, as estimated with a simple regression model. The plot suggests that there is considerable predictable heterogeneity in the extent to which enrollment in the care management program causally improves health. In particular, we find that the estimated treatment effect is only weakly correlated with the number of chronic conditions a patient currently exhibits ($r = 0.05$). Consequently, directly targeting resources to those most likely to benefit could yield large health improvements. Once the types of label bias we have discussed above have been identified, it may be possible to re-train predictive models to better align decision-making algorithms with policy goals.

6. Conclusion

From medicine to criminal justice, practitioners are increasingly turning to statistical risk assessments to help guide and improve human decisions. Algorithms can avoid many of the implicit and explicit biases of human decisions makers, but they can also exacerbate historical inequities if not developed with care. Policymakers, in response, have rightly demanded that these high-stakes decision systems be designed and audited to ensure outcomes are equitable. The research community has responded to the challenge, coalescing around several formal mathematical definitions of fairness. However, as we have aimed to articulate, these popular measures of fairness suffer from significant statistical limitations. Indeed, adopting these measures as algorithmic design principles can often harm the groups that these measures were designed to protect.

In contrast to the dominant axiomatic approach to algorithmic fairness, we advocate for a more consequentialist orientation (Cai et al., 2022a; Chohlas-Wood et al., 2023a; Liang et al., 2022; Nyarko et al., 2021). Most importantly, we stress the importance of grounding technical and policy discussions of fairness in terms of real-world quantities. For example, in the pretrial domain, one might consider a risk assessment's short and long-term impacts on public safety and the size of the incarcerated population, as well as a tool's alignment with principles of due process. In lending, one could similarly consider a risk assessment's immediate and equilibrium effects on community development and the sustainability of a loan program. Formal mathematical measures of fairness only indirectly address such issues, and can inadvertently lead discussions astray. Of course, it is not always clear how best to quantify or to balance the relevant costs and benefits of proposed algorithmic interventions. In some cases, it may be possible to conduct randomized controlled trials; in other cases, the best one can do is hypothesize about an algorithm's potential effects. Regardless, we believe a more explicit focus on consequences is necessary to make progress.

We further recommend decoupling the statistical problem of risk assessment from the policy problem of designing interventions. At their best, predictive algorithms estimate the likelihood of events under different scenarios; they cannot dictate policy. An algorithm might (correctly) infer that a defendant has a 20% chance of committing a violent crime if released, but that fact does not, in and of itself, determine a course of action. For example, detention is not the only alternative to release, as one could take any number of rehabilitative interventions (Barabas et al., 2018). Even if detention is deemed an appropriate

health gains, as the patients likely to be sickest in the absence of care are not typically the same as those likely to benefit the most from the program.

intervention, one must still determine what threshold would appropriately balance public safety with the social and financial costs of detention. One might even decide that society's goals are best achieved by setting different thresholds for different groups. For example, a policymaker might reason that, all else being equal, the social costs of detaining a single parent are higher than the social costs of detaining an individual without children, and thus decide to apply different thresholds to the two groups. When policymakers consider these options and others, we believe the primary role of a risk assessment tool is, as its name suggests, to estimate risk. This view, however, is at odds with requiring that algorithms satisfy popular fairness criteria. Such constrained algorithms typically do not reflect the best available estimates of risk, and thus implicitly conflate the statistical and policy problems.

Fair machine learning still has much left to accomplish and there are several important avenues of research that could benefit from new statistical and computational insights. From mitigating measurement error and sample bias, to understanding externalities and equilibrium effects, to eliciting and aggregating preferences to arbitrate between competing algorithms, there is much work to be done. But the benefits are equally large. When carefully designed and evaluated, statistical algorithms have the potential to dramatically improve both the efficacy and equity of consequential decisions. As these algorithms are increasingly deployed in all walks of life, it will become ever more important to ensure they are fair.

Acknowledgments

We thank Guillaume Basse, Sander Beckers, Hana Chockler, Alex Chohlas-Wood, Madison Coots, Avi Feller, Josh Grossman, Joe Halpern, Jennifer Hill, Aziz Huq, David Kent, Keren Ladin, Julian Nyarko, Emma Pierson, Ravi Sojitra, and Michael Zanger-Tishler for helpful conversations. This paper is based on work by Corbett-Davies and Goel (2018) and Nilforoshan et al. (2022). H.N was supported by a Stanford Knight-Hennessy Scholarship and an NSF 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. S.G. was supported by a grant from the Harvard Data Science Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or Amazon. Reproduction materials are available at `https://github.com/jgaeb/measure-mismeasure`.

Appendix 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**
        $V_k^\dagger \leftarrow \overline{V}_k^*$
      else
        $V_k^\dagger \leftarrow V_k^*$
      end
    end
    $\wp(V_j)^\dagger \leftarrow \{V_\ell^\dagger \mid V_\ell \in \wp(V_j)\}$
    $\overline{V}_j^* \leftarrow f_{V_j}(\wp(V_j)^\dagger, U_j^*)$
  end
end

$Z^*_{\Pi,a,a'} \leftarrow \overline{Z}^*$

Appendix B. Constructing Causally Fair Policies

Our aim is to identify the feasible region of expected outcomes attainable via policies which are constructed to satisfy various causal fairness constraints.

First, consider the problem of finding decision policies that maximize expected utility, subject to satisfying a given definition of causal fairness, as well as the outcome and budget constraints. 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

$$\begin{aligned} d^* \in \arg\max_{d \in \mathcal{C}} \quad & \mathbb{E}[d(X) \cdot u(X)] \\ \text{s.t.} \quad & o_1 - \epsilon \leq \mathbb{E}[d(X) \cdot \mathbb{1}_{\alpha(X)=a_1}] \leq o_1 + \epsilon \\ \text{s.t.} \quad & o_2 - \epsilon \leq \mathbb{E}[d(X) \cdot \mathbb{E}[Y(1) \mid X]] \leq o_2 + \epsilon \\ \text{s.t.} \quad & \mathbb{E}[d(X)] \leq b. \end{aligned} \tag{13}$$

We prove that this optimization problem 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 21 *Consider the optimization problem given in Eq.* (13)*.*

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.* (13) *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 $\mathcal{C}$.*

Before moving on to the proof of Theorem 21, we note that since the constraints of the linear programs are convex, the feasible regions in Figure 4 can be determined by solving the convex feasibility problem where we impose the additional convex constraint that the expected outcomes—in our admissions example, the aggregate academic index and number of admitted applicants from the target group—lie within some distance ϵ of a given point. Performing a grid search over all points then determines the feasible regions.

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, as well as the outcome and budget constraints are all linear in the decision variables.

1. The objective function $\mathbb{E}[d(X) \cdot u(X)]$ equals $\sum_{i=1}^m d_i \cdot u(x_i) \cdot p_i$
2. The budget constraint $\mathbb{E}[d(X)] \leq b$. constraint equals $\sum_{i=1}^m d_i \cdot p_i \leq b$
3. The first outcome constraint $o_1 - \epsilon \leq \mathbb{E}[d(X) \cdot \mathbb{1}_{\alpha(X)=a_1}] \leq o_1 + \epsilon$ equals $o_1 - \epsilon \leq \sum_{i=1}^m \mathbb{1}_{\alpha(x_i)=a_1} \cdot d_i \cdot p_i] \leq o_1 + \epsilon$
4. The second outcome constraint $o_2 - \epsilon \leq \mathbb{E}[d(X) \cdot \mathbb{E}[Y(1) \mid X]] \leq o_2 + \epsilon$ equals $o_2 - \epsilon \leq \sum_{i=1}^m \mathbb{E}[Y(1) \mid X = x_i] \cdot d_i \cdot p_i \leq o_2 + \epsilon$

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 21 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 21 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

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

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

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

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 21 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

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

Similarly, expanding the denominator yields

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

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

$$\sum_{i=1}^{m} [1 - d_i] \cdot \Pr(Y(1) = y, A = a, X = x_i) = C_y \cdot \sum_{i=1}^{m} [1 - d_i] \cdot \Pr(Y(1) = y, X = x_i), \quad (14)$$

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, outcome constraint, and budget constraint as before, together with additional constraints for each $a \in \mathcal{A}$ as in Eq. (14), where $C_{y_i} = v_i$ for $i = 1, \ldots, k$.

By assumption, there exists a feasible solution to the optimization problem in Eq. (13), so the solution to at least one program in $\mathcal{L}$ is a utility-maximizing policy that satisfies counterfactual predictive parity. ■

Appendix C. A Stylized Example of College Admissions

In the example that we consider in Section 4.1, the exogenous variables in the DAG, $\mathcal{U} = \{u_A, u_D, u_E, u_M, u_T, u_Y\}$, are independently distributed as follows:

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

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:

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

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 $b = \frac{1}{4}$.

Appendix D. Proof of Theorem 10

We begin with the following simple lemma.

Lemma 22 *Suppose $\mathcal{F}$ is a sub–σ-algebra of measurable sets, and suppose X is a non-negative bounded random variable with $X \leq b$ a.s. Then*

$$\frac{\textsc{Var}(X \mid \mathcal{F})}{b} \leq \mathbb{E}[X \mid \mathcal{F}].$$

Proof Note that since $0 \leq X \leq b$ a.s.,

$$\mathbb{E}[X^2 \mid \mathcal{F}] \leq \mathbb{E}[X \mid \mathcal{F}] \cdot b.$$

By Jensen's inequality, $\textsc{Var}(X \mid \mathcal{F}) \leq \mathbb{E}[X^2 \mid \mathcal{F}]$, and so, it follows that, a.s.,

$$\frac{\textsc{Var}(X \mid \mathcal{F})}{b} \leq \mathbb{E}[X \mid \mathcal{F}],$$

as desired. ■

Ignoring the conditioning, Lemma 22 can be interpreted as saying that the minimum of a bounded random variable cannot be too close to the mean. This fact enables us to prove Theorem 10.

Proof of Theorem 10 We wish to show that no event E of the form $\{r(X) \geq t\} \subseteq E \subseteq \{r(X) > t\}$ is π-measurable. To that end, it suffices to show that $\mathbb{E}[E \mid \pi(X)] \notin \{0, 1\}$ with positive probability.

Since $\Pr(r(X) = t) = 0$, $\mathbb{1}_{r(X) \geq t} = \mathbb{1}_{r(X) > t}$ a.s., and so it suffices to show that $\mathbb{E}[\mathbb{1}_{r(X)>t} \mid \pi(X)] \notin \{0, 1\}$ with positive probability.

Consider the set of covariates $x = (x_u, a)$ such that $\rho(x)$ lies in the interval $(t, t + \epsilon)$, where, without loss of generality, we assume $t + \epsilon < 1$. Since $\mathbb{E}[r(X) \mid X_u = x_u] = \rho(x) > t$, it follows immediately that for these x, $\Pr(r(X) > t \mid X_u) > 0$. Let $I(x)$ denote the essential infimum of the conditional distribution of $r(X) \mid \rho(X) = \rho(x)$.

Then, we can apply Lemma 22 to $r(X) - I(X)$, using the fact that $0 \leq r(X) - I(X) \leq 1$ a.s., to obtain that $\rho(X) - I(X) > \epsilon$ a.s. It follows that the event

$$\{\rho(X) \in (t, t + \epsilon), \Pr(r(X) < t \mid \rho(X)) > 0\} \tag{15}$$

has positive probability. We can conclude from this that the related event

$$\{\rho(X) \in (t, t + \epsilon), \Pr(r(X) < t \mid \pi(X)) > 0\}$$

cannot have probability zero, since, by the tower law, we would consequently have that, a.s.,

$$0 = \mathbb{1}_{\rho(X) \in (t, t+\epsilon)} \cdot \Pr(r(X) < t \mid \pi(X)),$$

and so, taking conditional expectations with respect to $\rho(X)$, a.s.,

$$0 = \mathbb{1}_{\rho(X) \in (t, t+\epsilon)} \cdot \Pr(r(X) < t \mid \rho(X)),$$

which contradicts Eq. (15).

Therefore, for x such that $\rho(x) \in (t, t + \epsilon)$, $\Pr(r(X) < t \mid X_u = x_u) > 0$ and $\Pr(r(X) > t \mid X_u = x_u) > 0$. Since $\Pr(\rho(X) \in (t, t + \epsilon)) > 0$, it follows that $\mathbb{E}[\mathbb{1}_{r(X)>t} \mid \pi(X)] \notin \{0, 1\}$ with positive probability, as desired. ■

Appendix E. Proof of Proposition 15

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 23 *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 24 *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 25 *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. 15.

Lemma 26 *Let $\mathcal{U}$ be a collection of utilities consistent modulo α. Let $d(x)$ be a feasible 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 $m^{\text{Up}}(t') > 0$, and so by hypothesis $m^{\text{Lo}}(t') = 0$. 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 $m^{\text{Up}} = m^{\text{Up}}(t^*)$ and $m^{\text{Lo}} = m^{\text{Lo}}(t^*)$ and consider the alternative policy

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

Then it follows by construction that

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

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

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

and so

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

Therefore

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

It remains to show 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 $u'(x) \geq t' \geq u'(x')$, and, moreover, that at least one of the inequalities is strict. Without loss of generality, assume $u'(x) > t' \geq u'(x')$. Then, we have that $u(x) > t^*$ if and only if $u'(x) > t'$. Therefore, 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

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

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 25, 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 results, which draw on Lemma 26, are useful in the proof of Theorem 17.

Definition 27 *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 28 *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 29 *Let $\mathcal{U}$ be a collection of utilities consistent modulo α. If $\tau(x)$ is a feasible policy that is not a budget-exhausting multiple threshold policy with non-negative thresholds, then $\tau(x)$ is strongly Pareto dominated.*

Proof Suppose $\tau(x)$ is not strongly Pareto dominated. By Lemma 26, 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

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

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 30 *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

$$t_a = \inf\{s \in \mathbb{R} : \Pr(u(X) > s) < q_a\}. \tag{16}$$

Then define

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

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. (16). ■

We can further refine Cor. 29 and Lemma 30 as follows:

Lemma 31 *Let u be a utility. Then a feasible 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. 29, 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 30, such a policy exists, and so the maximum is attained. ■

Appendix F. Prevalence and the Proof of Theorem 17

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 dominated 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 26, 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, 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 F.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 and Yorke (2005). In Section F.2 we provide a road map of the proof itself. In Section F.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 F.4, we establish a number of technical lemmata used in the proof of Theorem 17, and provide a proof of the theorem itself in Section F.5. In Section F.8, 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.

F.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 and 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 32 (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 compact* $C \subseteq V$ *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 33 (Hunt et al. (1992)) *Let* V *be a completely metrizable topological vector space. Then:*

- *Any 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* $\lambda_n[\mathbb{R}^n \setminus G] = 0$.

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. 33 holds.

Definition 34 (Anderson and 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 *or* shy relative to 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 35 (Anderson and Zame (2001)) *If* E *is shy at some point* $c \in C$, *then* E *is shy at every point in* C *and hence is shy in* C.

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

Proposition 36 (Anderson and 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:*

- *Any 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 non-empty 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 17.

Definition 37 (Anderson and Zame (2001)) *A universally measurable subset E of a convex and completely metrizable set 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 Lebesgue 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 Lebesgue 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.[33] *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 $\lambda_W[E + v] = 0$ for all $v \in V$—then E itself is small in V, i.e., E has λ_V-measure zero.

Proposition 38 (Anderson and Zame (2001)) *Every k-shy set in C is shy in C.*

F.2 Outline

To aid the reader in following the application of the theory in Section F.1 to the proof of Theorem 17, we provide the following outline of the argument.

In **Section F.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 $\mathbf{Q} \subseteq \mathbf{K}$, the set of *$\mathcal{U}$-fine joint probability distributions* of, respectively, X and $Y(1)$;

33. 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.

X, $Y(0)$, $Y(1)$; or A and the $X_{\Pi,A,a}$. Within $\mathbf{Q}$, 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 17 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. 36 holds—Definition 34 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 42**, we observe that $\mathbb{K}$ is a complete topological vector space under the total variation norm, and so is a Banach space. We extend this in **Cor. 47**, showing that $\mathbf{K}$ is also a Banach space. We use this fact in **Lemma 50** to show that $\mathbf{Q}$ is a completely metrizable subset of $\mathbf{K}$, as well as convex. Lastly, in **Lemma 56**, we show that the set $\mathbf{E}$ is closed, and therefore universally measurable.

In **Section F.4**, we develop the machinery needed to construct a probe $\mathbf{W}$ for the proof of Theorem 17 and prove several lemmata simplifying the eventual proof of the theorem. To build the probe, it is necessary to construct measures $\mu_{\max,a}$ with maximal support on the utility scale. This ensures that if any two threshold policies produce different decisions on *any* $\mu \in \mathbf{K}$, they will produce different decisions on typical perturbations. The construction of the $\mu_{\max,a}$, is carried out in **Lemma 58** and **Cor. 59**. Next, we introduce the basic style of argument used to show that a subset of $\mathbf{Q}$ is shy in **Lemma 62** and **Lemma 63**, in particular, by showing that the set of $\mu \in \mathbf{Q}$ that give positive probability to an event E is either prevalent or empty. We use then use a technical lemma, **Lemma 64**, 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. 66**, we introduce the concept of overlapping and splitting utilities, and show in **Lemma 67** that this property is generic in $\mathbf{Q}$ unless there exists a ω-stratum that contains no positive-utility observables x. Lastly, in **Lemma 68**, we provide a mild simplification of the characterization of finitely shy sets that makes the proof of Theorem 17 more straightforward.

Finally, in **Section F.5**, we give the proof of Theorem 17. 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. (4) 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. (4) can only hold for a $\lambda_{\mathbf{W}}$-null set of perturbations. In the final section, we lay out modifications that can be made to the proof given for counterfactual equalized odds in the first two parts that 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. (5) and (8) does not affect the argument.

F.3 Convexity, Complete Metrizability, and Universal Measurability

In this section, we establish the background requirements of Prop. 38 for the setting of Theorem 17. 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 $\mathcal{U}$-fine probability 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 closed, and therefore universally measurable.

F.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 17 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 $\mathbb{R}^n$.[34]

We recall the definition of totally bounded measures.

Definition 39 *Let $\mathcal{M}$ be a σ-algebra on V, and let μ be a countably additive $(V, \mathcal{M})$-measure. Then, we define*

$$|\mu|[E] = \sup \sum_{i=1}^{\infty} |\mu[E_i]| \tag{17}$$

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 $|\mu|[V]$.*

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

Lemma 40 *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

34. In the case of path-specific fairness, we can equivalently think of $\mathcal{A}$ as a set of integers indexing the groups.

E, $\Pr_\mu(E) = \mu[E]$.[35] 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 41 *Let g be a Borel function on $\mathcal{K}$, and suppose $\Pr_\mu(g = c) \neq 0$ for some constant $c \in \mathbb{R}$. Then, we have that μ-a.s., for any Borel function f,*

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

See Rao (2005) for proof.

With these notational conventions in place, we turn to establishing the background conditions of Prop. 38.

Lemma 42 *The set of totally bounded measures on a measure space $(V, \mathcal{M})$ form a complete topological vector space under the total variation norm, and hence a Banach space.*

See, e.g., Steele (2019) for proof. It follows from this that $\mathbb{K}$ is a Banach space.

Remark 43 *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 38 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.

Definition 44 *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 45 *Consider the space of totally bounded measures on a measure space $(V, \mathcal{M})$ and fix μ. The set of ν such that $\nu \ll \mu$ is closed.*

35. 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.

Proof Let $\{\nu_i\}_{i\in\mathbb{N}}$ be a convergent sequence of measures absolutely continuous with respect to μ. Let the limit of the ν_i be ν. 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 46 *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 $\mu \circ f^{-1}$ *on V' by the map $E' \mapsto \mu[f^{-1}(E')]$ for $E' \in \mathcal{M}'$.*

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

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

In the case of path-specific fairness, we require the joint distributions of the counterfactual utilities to have a joint density. That is, we define the subspace $\mathbf{K}$ to be the set of totally bounded measures μ on $\mathcal{K}$ such that the pushforward measure $\mu \circ (u^{\mathcal{A}})^{-1}$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{\mathcal{A}}$ for all $u \in \mathcal{U}$. Here, we recall that

$$u^{\mathcal{A}} : (a, (x_{a'})_{a'\in\mathcal{A}}) \mapsto (u(x_{a'}))_{a'\in\mathcal{A}}.$$

As before, there exists a corresponding density $f_\mu \in L^1(\mathbb{R}^{\mathcal{A}})$.

We therefore see that $\mathbf{K}$ extends in a natural way 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 $\Pr_\mu(u(X) \in E, A = a) \le \Pr_\mu(u(X) \in E)$, it also follows that, for $a \in \mathcal{A}$ such that $\Pr_\mu(A = a) > 0$, 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 45.

Corollary 47 *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 45, it is complete, and therefore a Banach space. ■

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

Lemma 48 *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 $|\mu - \mu'|[\mathcal{K}] < \epsilon$. Then, let

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

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

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

where the second equality follows by the definition of pushforward measures and the inequality follows from Lemma 40. 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 defined wholly by their thresholds for distributions in $\mathbf{K}$ and $\mathbf{Q}$. Importantly, this observation does not hold when there are atoms on the utility scale—which measures in $\mathbf{K}$ lack—which can in turn lead to counterexamples to Theorem 17; see Appendix F.8.

Lemma 49 *Let $\tau_0(x)$ and $\tau_1(x)$ be two multiple threshold policies. If $\tau_0(x)$ and $\tau_1(x)$ have the same thresholds, then for any $\mu \in \mathbf{K}$, $\tau_0(X) = \tau_1(X)$ μ-a.s. Similarly, for $\mu \in \mathbf{Q}$, if*

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

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

Moreover, for $\mu \in \mathbf{K}$ in the case of path-specific fairness, if $\tau_0(x)$ and $\tau_1(x)$ have the same thresholds, then $\tau_0(X_{\Pi,A,a}) = \tau_1(X_{\Pi,A,a})$ μ-a.s. for any $a \in \mathcal{A}$. Similarly, for $\mu \in \mathbf{Q}$ in the case of path-specific fairness, if

$$\mathbb{E}_\mu[\tau_0(X_{\Pi,A,a})] = \mathbb{E}_\mu[\tau_1(X_{\Pi,A,a})]$$

then $\tau_0(X_{\Pi,A,a}) = \tau_1(X_{\Pi,A,a})$ μ-a.s. as well.

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,

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

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

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

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. ■

F.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. 38. To do so, we must show that $\mathbf{Q}$ is a convex and completely metrizable subset of $\mathbf{K}$.

Lemma 50 *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. 47, 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 51 *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. (7), 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 55, 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 56.

Definition 52 *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 $C = \{g : g = f \ \mu\text{-a.e.}\}$.*[36] *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 53 *It is straightforward to see that for $f \in L^\infty(\mu)$, a standard version always exists with $C = \|f\|_\infty$.*

Remark 54 *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*

$$\|\mathbb{E}_{\mu}[f \mid \mathcal{F}] - \mathbb{E}_{\mu'}[f \mid \mathcal{F}]\|_{L^1(\mu)} \tag{18}$$

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. (18) *for any standard version of $\mathbb{E}_{\mu'}[f \mid \mathcal{F}]$.*

Lemma 55 *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

$$\int_V |\mathbb{E}_{\mu}[f \mid \mathcal{F}] - g| \, \mathrm{d}\mu \leq 4C \cdot |\mu - \mu'|[V]. \tag{19}$$

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

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

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

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

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

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 f is everywhere bounded by C, applying Lemma 40 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 56 *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 notational 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 48, $f_{\mu_i,a} \to f_{\mu,a}$ in $L^1(\mathbb{R})$. Moreover, by Lemma 26, 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

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

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 30, 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 48, $\tau_{n_i}(X) \to \tau(X)$ in $L^1(\mu)$.

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

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

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

$$\mathbb{E}_{\mu_{n_i}}[\tau_{n_i}(X) \mid A, Y(1)] = \mathbb{E}_{\mu_{n_i}}[\tau_{n_i}(X) \mid Y(1)]. \tag{20}$$

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. (20), $g_i(x)$ 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 55, we have that

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

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.[37] Further recall that $D = \mathbb{1}_{U_D \le \tau(X)}$, 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,

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

37. For a proof of this fact see, e.g., Brozius (2019).

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. ■

F.4 Shy Sets and Probes

We require a number of additional technical lemmata for the proof of Theorem 17. The probe must be constructed carefully, so that, on the utility scale, an arbitrary element of $\mathbf{Q}$ 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 17. For instance, Lemma 63 is used in Theorem 17 to show that a certain conditional expectation is generically well-defined, avoiding the need to separately treat certain corner cases.

Cor. 59 concerns the construction of the probe used in the proof of Theorem 17. Lemmata 64 to 68 use Cor. 59 to provide additional simplifications to the proof of Theorem 17.

F.4.1 Maximal Support

First, to construct the probe used in the proof of Theorem 17, we require elements $\mu \in \mathbf{Q}$ such that the densities f_μ have "maximal" support. To produce such distributions, we use the following measure-theoretic construction.

Definition 57 *Let $\{E_\alpha\}_{\gamma\in\Gamma}$ 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 $\{E_\gamma\}_{\gamma\in\Gamma}$ if $\mu[E_\gamma \setminus E] = 0$ for all $\gamma \in \Gamma$ and $E = \bigcup_{i=1}^{\infty} E_{\gamma_i}$ for some countable subcollection $\{\gamma_i\} \subseteq \mathbb{N}$.*

While measure-theoretic unions themselves are known (cf. Silva (2008), Rudin (1991)), for completeness, we include a proof of their existence, which, to the best of our knowledge, is not found in the literature.

Lemma 58 *Let μ be a finite positive measure on a measure space $(V, \mathcal{M})$. Then an arbitrary collection $\{E_\gamma\}_{\gamma\in\Gamma}$ of μ-measurable sets has a measure-theoretic union.*

Proof For each countable subcollection $\Gamma' \subseteq \Gamma$, consider the "error term"

$$r(\Gamma') = \sup_{\gamma\in\Gamma} \mu\left[E_\gamma \setminus \bigcup_{\gamma'\in\Gamma'} E_{\gamma'}\right]$$

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

For, toward a contradiction, suppose it were greater than or equal to $\epsilon > 0$. Choose any set E_{γ_1} such that $\mu[E_{\gamma_1}] \geq \epsilon$. Such a set must exist, since otherwise $r(\emptyset) < \epsilon$. Choose E_{γ_2} such that $\mu[E_{\gamma_2} \setminus E_{\gamma_1}] > \epsilon$. Again, some such set must exist, since otherwise $r(\{\gamma_1\}) < \epsilon$. Continuing in this way, we construct a countable collection $\{E_{\gamma_i}\}_{i\in\mathbb{N}}$.

Therefore, we see that

$$\mu[V] \geq \mu\left[\bigcup_{i=1}^{n} E_{\gamma_i}\right] = \sum_{i=1}^{n} \mu\left[E_{\gamma_i} \setminus \bigcup_{j=1}^{i} E_{\gamma_j}\right].$$

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

Therefore, there exist countable collections $\{\Gamma_n\}_{n\in\mathbb{N}}$ such that $r(\Gamma_n) < \frac{1}{n}$. It follows immediately that for all n

$$r\left(\bigcup_{n\in\mathbb{N}} \Gamma_n\right) \leq r(\Gamma_k)$$

for any fixed $k \in \mathbb{N}$. Consequently,

$$r\left(\bigcup_{n\in\mathbb{N}} \Gamma_n\right) = 0,$$

and $\bigcup_{n\in\mathbb{N}} \Gamma_n$ is countable. ■

The construction of the "maximal" elements used to construct the probe in the proof of Theorem 17 follows as a corollary of Lemma 58

Corollary 59 *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 $\{\textsc{Supp}(f_{\mu,a})\}_{\mu\in\mathbf{K}}$. By Lemma 58, there exists a countable collection of measures $\{\mu_i\}_{i\in\mathbb{N}}$ such that for any $\mu \in \mathbf{K}$,

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

where, without loss of generality, we may assume that $\lambda[\textsc{Supp}(f_{\mu_i,a})] > 0$ for all $i \in \mathbb{N}$. Such a sequence must exist, since, by the first hypothesis of Theorem 17, for every $a \in \mathcal{A}$, there exists $\mu \in \mathbf{Q}$ such that $\Pr_\mu(A = a) > 0$. Therefore, we can define the probability measure $\mu_{\max,a}$, where

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

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 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 60 *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 as Cor. 59, 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 for all $w \in \mathcal{W}$, there is some $\mu \in \mathbf{K}$ such that $\Pr_\mu(W = w) > 0$.

Remark 61 *Because their support is maximal, the hypotheses of Theorem 17, in addition to implying that $\mu_{\max,a}$ is well-defined for all $a \in \mathcal{A}$, also imply that $\Pr_{\mu_{\max,a}}(u(X) > 0) > 0$. In the case of conditional principal fairness, they further imply that $\Pr_{\mu_{\max,a}}(W = w) > 0$ for all $w \in \mathcal{W}$ and $a \in \mathcal{A}$. Likewise, in the case of path-specific fairness, they further imply that $\Pr_{\mu_{\max,a}}(W = w_i) > 0$ for $i = 0, 1$ and some $a \in \mathcal{A}$.*

F.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 68, which allows us to use these generic properties to simplify the proof of Theorem 17.

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

Lemma 62 *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 $\mu \in \mathbf{K}$ such that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$.*

Proof Set

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

where $\nu_1, \ldots, \nu_n$ form a basis of $\mathbf{W}$. Then, if $|\beta_i| \leq \frac{1}{|\nu_i|[\mathcal{K}]}$, it follows that

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

Since

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

it follows that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$. ■

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 65 alone in principle suffices for the proof of Theorem 17, we include Lemma 63 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 17 to show that a set is shy relative to $\mathbf{Q}$.

Lemma 63 *For a Borel set $E \subseteq \mathcal{K}$, suppose there exists $\nu \in \mathbf{Q}$ such that $\nu[E] > 0$. Then the set of $\mu \in \mathbf{Q}$ such that $\mu[E] > 0$ is prevalent.*

Proof First, we note that the set of $\mu \in \mathbf{Q}$ such that $\mu[E] = 0$ 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

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

Now, if $\mu[E] > 0$ for all $\mu \in \mathbf{Q}$, there is nothing to prove. Therefore, suppose that there exists $\nu' \in \mathbf{Q}$ such that $\nu'[E] = 0$.

Next, consider the measure $\tilde{\nu} = \nu' - \nu$. Then, let $\mathbf{W} = \textsc{Span}(\tilde{\nu})$. Since $\tilde{\nu} \neq 0$ and

$$\tilde{\nu}[\mathcal{K}] = \nu'[\mathcal{K}] - \nu[\mathcal{K}] = 0,$$

it follows by Lemma 62 that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$ for some μ.

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

$$\mu[E] - \beta \cdot \nu[E] = 0$$

i.e.,

$$\beta = \frac{\mu[E]}{\nu[E]}.$$

A singleton has null Lebesgue measure, and so the set of $\nu \in \mathbf{W}$ such that $(\mu + \nu)[E] = 0$ is $\lambda_{\mathbf{W}}$-null. Therefore, by Prop. 38, the set of $\mu \in \mathbf{Q}$ such that $\mu[E] = 0$ is shy relative to $\mathbf{Q}$, as desired. ■

While Lemma 63 shows that a typical element of $\mathbf{Q}$ "sees" individual events, in the proof of Theorem 17, 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 64 *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 $\beta, \beta' \in \mathbb{R}$,

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

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 if $\{Z_{\beta_i}\}_{i\in\mathbb{N}}$ is a countable collection of distinct elements of $\mathbb{R}$, then

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

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,

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

is countable. ∎

We now apply Lemma 64 to the proof of 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 64 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 65 *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 65 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, we 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

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

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

$$\tilde{\nu}[E'] = \int_{E'} \mathbb{1}_{u^{-1}(S_0)} - \mathbb{1}_{u^{-1}(S_1)} \, \mathrm{d}\nu.$$

Take $\mathbf{W} = \textsc{Span}(\tilde{\nu})$. Since $\tilde{\nu} \neq 0$ and $\tilde{\nu}[\mathcal{K}] = 0$, we have by Lemma 62 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 non-zero $(\mu \circ u^{-1})$-a.e. Therefore, by Lemma 64, 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. 38. ■

The following definition and technical lemma are needed to extend Theorem 17 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 17 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 receives a positive or negative decision). Definition 66 formalizes this pathology, and Lemma 67 shows that this issue—under a mild hypothesis—does not arise for a generic element of $\mathbf{Q}$.

Definition 66 *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 67.

Lemma 67 *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 $\mathbf{E}_\Gamma$ of measures $\mu \in \mathbf{Q}$ such that there exists a splitting policy where the thresholds fall below $w \in \Gamma \subseteq \mathcal{W}$ and above $w \notin \Gamma$ is Borel, before constructing a probe that shows that it is shy. Then, we argue that $\mathbf{E}' = \bigcup_{\Gamma \subseteq \mathcal{W}} \mathbf{E}_\Gamma$, 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 $\Gamma \subseteq \mathcal{W}$, the sets

$$F_\Gamma^{\mathrm{Up}} = \{x \in \mathbb{R} \times \mathbb{R}^{\mathcal{W}} : x_0 \geq b, b = \sum_{w \in \Gamma} x_w\},$$
$$F_\Gamma^{\mathrm{Lo}} = \{x \in \mathbb{R} \times \mathbb{R}^{\mathcal{W}} : x_0 \leq b, x_0 = \sum_{w \in \Gamma} x_w\},$$

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

$$\mathbf{E}_\Gamma = \mathbf{Q} \cap \Phi^{-1}\left(\bigcup_{\Gamma \subseteq \mathcal{W}} F_\Gamma^{\mathrm{Up}} \cup F_\Gamma^{\mathrm{Lo}}\right) \tag{21}$$

is closed, and therefore universally measurable.

Note that by our hypothesis and Cor. 60, 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 $\mathbf{E}_\Gamma$ 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 62.

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

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

Now, since $0 < b < 1$ and ω partitions $\mathcal{X}$, it follows that

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

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

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

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

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

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

only if

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

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

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

for all $\mu \in \mathbf{K}$. Therefore, by Prop. 38, $\mathbf{E}_\Gamma$ is shy in $\mathbf{Q}$. Taking the union over $\Gamma \subseteq \mathcal{W}$, it follows by Prop. 36 that $\bigcup_{\Gamma \subseteq \mathcal{W}} \mathbf{E}_\Gamma$ is shy.

Now, we must show that $\mathbf{E}' = \bigcup_{\Gamma \subseteq \mathcal{W}} \mathbf{E}_\Gamma$. By construction, $\mathbf{E}_\Gamma \subseteq \mathbf{E}'$, since the policy $\tau(x) = \mathbb{1}_{\omega(x) \in \Gamma}$ 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

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

Since $\tau(x)$ is budget-exhausting, it follows immediately that $\mu \in \mathbf{E}_{\Gamma_\mu}$. Therefore, $\mathbf{E}' = \bigcup_{\Gamma \subseteq \mathcal{W}} \mathbf{E}_\Gamma$, 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. 38.

Lemma 68 *Suppose E is a universally measurable subset of a convex, completely metrizable set C in a topological vector space V. Suppose that for some finite-dimensional subspace V', $\lambda_{V'}[C + v_0] > 0$ for some $v_0 \in V$. If, in addition, for all $v \in E$,*

$$\lambda_{V'}[\{v' \in V' : v + v' \in E\}] = 0, \tag{22}$$

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. (22).

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

F.5 Proof of Theorem 17

Using the lemmata above, we can prove Theorem 17. We briefly summarize what has been established so far:

- **Lemma 42:** The set $\mathbf{K}$ of $\mathcal{U}$-fine distributions on $\mathcal{K}$ is a Banach space;
- **Lemma 50:** The subset $\mathbf{Q}$ of $\mathcal{U}$-fine probability measures on $\mathcal{K}$ is a convex, completely metrizable subset of $\mathbf{K}$;
- **Lemma 56:** 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. 38, it follows that what remains is 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. 29, 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. (4). In the second case, we treat the possibility that the candidate budget-exhausting multiple threshold policy has a non-trivial positive threshold for at least one group. 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 $\nu \in \mathbf{W}$,

$$\int_{\mathcal{K}} 1 \, \mathrm{d}\nu = 0, \tag{23}$$

and

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)>0} \, \mathrm{d}\nu = 0. \tag{24}$$

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 $\nu \in \mathbf{W}$,

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)>0, A=a} \, \mathrm{d}\nu = 0. \tag{25}$$

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

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)<0, Y=y_i, A=a} \, \mathrm{d}\nu^{\mathrm{Lo}} \neq 0. \tag{26}$$

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

$$\int_{\mathcal{K}} \mathbb{1}_{u(X)>t, A=a} \, \mathrm{d}\nu^{\mathrm{Up}} \neq 0. \tag{27}$$

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

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

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

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

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

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

then it is easy to see that Eqs. (23) to (26) will hold. The only non-trivial case is Eq. (27). However, by Cor. 59, 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 $t \leq r_a$ or $t > r_a$. First, assume $t \leq r_a$; then, it follows by the construction of ν_a^{Up} that

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

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. (27) holds.

Since $\mathbf{W}$ is non-trivial[38] and $\nu[\mathcal{K}] = 0$ for all $\nu \in \mathbf{W}$, it follows by Lemma 62 that $\lambda_{\mathbf{W}}[\mathbf{Q} - \mu] > 0$ for some $\mu \in \mathbf{K}$.

Shyness. Recall that, by Prop. 36, a set E is shy if and only if, for an arbitrary shy set E', $E \setminus E'$ is shy. By Lemma 63, 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 65, a generic $\mu \in \mathbf{Q}$ satisfies $\nu_a^{\mathrm{Up}} \circ u^{-1} \ll (\mu \restriction_{\mathcal{X} \times \{y_1\}}) \circ u^{-1}$. Therefore, to simplify our task and recalling Remark 61, we may instead demonstrate the shyness of the set of $\mu \in \mathbf{Q}$ such that:

- There exists a budget-exhausting multiple threshold policy $\tau(x)$ with non-negative thresholds satisfying counterfactual equalized odds over μ;

38. 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 61, 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 $\mathbf{W} \neq \{0\}$.

- For $i = 0, 1$,
$$\Pr_\mu(u(X) > 0, A = a, Y(1) = y_i) > 0; \tag{28}$$
- For all $a \in \mathcal{A}$,
$$\nu_a^{\mathrm{Up}} \circ u^{-1} \ll (\mu \restriction_{\alpha^{-1}(a) \times \{y_1\}}) \circ u^{-1}. \tag{29}$$

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. (28) is not affected by perturbation by $\nu \in \mathbf{W}$, and Eq. (29) is not affected by perturbation by $\nu^{\mathrm{Lo}} \in \mathbf{W}$.

In particular, by Lemma 68, 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.

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 $a \in \mathcal{A}$,

$$0 < \mathbb{E}_\mu[\tau'(X) \mid A = a] < 1. \tag{30}$$

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 so,

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

contradicting the fact that $\tau'(x)$ satisfies counterfactual equalized odds over μ. Therefore, in what follows, we can assume that Eq. (30) 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. (30) would not hold for $\tau(x)$.

Next, we note that by Eq. (24), 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 equalized odds holds with decision policy $\tau(x) = \mathbb{1}_{u(x)>0}$, then, by Eq. (7) and Lemma 41, 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 $a, a' \in \mathcal{A}$.[39]

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. (28),

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

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 41, that

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

where

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

39. To ensure that both quantities are well-defined, here and throughout the remainder of the proof we use the fact that by Eqs. (25) and (28), $\Pr_{\mu+\nu}(u(X) > 0, A = a, Y(1) = y_1) > 0$.

since

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

Here, the equality follows by the fact that ν^{Lo} is supported on $S_a^{\mathrm{Lo}} \times \{y_0, y_1\}$ and the inequality from Eq. (26).

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

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

since, as noted above, $\rho \neq 0$ by Eq. (26). 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. (28), this is impossible.

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

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

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 $\nu^{\mathrm{Lo}} \in \mathbf{W}^{\mathrm{Lo}}$, 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 $\mu + \nu^{\mathrm{Lo}} + \nu^{\mathrm{Up}}$ for any $\nu^{\mathrm{Up}} \in \mathbf{W}^{\mathrm{Up}}$. Then, we show that the set of ν^{Up} such that $\tilde{\tau}(X)$ satisfies counterfactual equalized odds has $\lambda_{\mathbf{W}^{\mathrm{Up}}}$ measure zero. Finally, we argue 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 $\lambda_{\mathbf{W}^{\mathrm{Up}}}[\mathbf{E} - (\mu + \nu^{\mathrm{Lo}})] = 0$. To begin, we note that since $\nu_{a,i}^{\mathrm{Up}}$ concentrates on $\{y_1\} \times \mathcal{X}$ for all $a \in \mathcal{A}$, it follows that

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

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

Now, suppose there exists some $\nu^{\mathrm{Up}} \in \mathbf{W}^{\mathrm{Up}}$ 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 $\mu+\nu^{\mathrm{Lo}}+\nu^{\mathrm{Up}}$. (If not, then we are done and $\lambda_{\mathbf{W}^{\mathrm{Up}}}[\mathbf{E}-(\mu+\nu^{\mathrm{Lo}})] = 0$, as the measure of the empty set is zero.) Let

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

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. (28), $\Pr_{\mu+\nu^{\text{Lo}}}(A = a_0, Y(1) = y_0) > 0$, 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 fact that ν^{Lo} is supported on $u^{-1}((-\infty, 0])$, the preceding discussion, and Lemma 49, it follows that

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

By Eq. (29), 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. (24), 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. (30), 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. (27) 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_a^{Lo}, ν_a^{Up} 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 41, that

$$\begin{aligned} &\Pr_{\mu+\nu}(u(X) > \tilde{t}_{a^*} \mid A = a^*, Y(1) = y_1) \\ &\qquad = \frac{\eta + \beta_a^{\mathrm{Up}} \cdot \gamma}{\pi} \end{aligned} \tag{31}$$

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

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

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

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

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

$$\mathbf{W}^{\mathrm{Up}} = \textsc{Span}(\nu^{\mathrm{Up}}_{a,w}),$$
$$\mathbf{W}^{\mathrm{Lo}} = \textsc{Span}(\nu^{\mathrm{Lo}}_{a,w}),$$

almost as before.

The proof otherwise proceeds virtually identically, except for two points. First, recalling Remark 61, we use the fact that a generic element of $\mathbf{Q}$ satisfies $\Pr_\mu(A = a, W = w) > 0$ in place of $\Pr_\mu(A = a) > 0$ throughout. Second, we use the fact that ω overlaps utility in place of Eq. (30). In particular, If ω does not overlap utilities for a generic $\mu \in \mathbf{Q}$, then, by Lemma 67, 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. (30).

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

$$S^{\mathrm{Lo}}_{a,w} = S_{a,w} \cap (-\infty, r_{a,w}),$$
$$S^{\mathrm{Up}}_{a,w} = S_{a,w} \cap [r_{a,w}, \infty),$$

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

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

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

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

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

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

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 $\nu_{a,w} \circ \pi_X^{-1} = 0$, 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. (23) and (24) hold. Moreover, the ν_a satisfy the following strengthening of Eq. (27). 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

$$\int_{\mathcal{K}} \mathbb{1}_{u(X_{\Pi,A,a})>t} \, \mathrm{d}\nu_a \neq 0. \tag{32}$$

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. (30), 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 61 and Lemma 63, this conditioning is well-defined for a generic element of $\mathbf{Q}$. ■

We have focused on causal definitions of fairness, but the thrust of our analysis applies to non-causal conceptions of fairness as well. Below we show that policies constrained to satisfy (non-counterfactual) equalized odds (Hardt et al., 2016) are generically strongly Pareto dominated, a result that follows immediately from our proof above.

Definition 69 Equalized odds *holds for a decision policy* $d(x)$ *when*

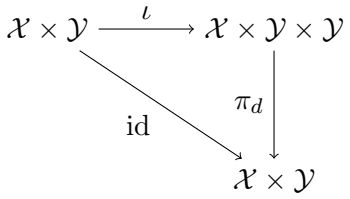
$$d(X) \perp\!\!\!\perp A \mid Y. \tag{33}$$

We note that Y in Eq. (33) does *not* depend on our choice of $d(x)$, but rather represents the realized value of Y, e.g., under some *status quo* decision making rule.

Corollary 70 *Suppose* $\mathcal{U}$ *is a set of utilities consistent modulo* α*. Further suppose that for all* $a \in \mathcal{A}$ *there exist a* $\mathcal{U}$*-fine distribution of* X *and a utility* $u \in \mathcal{U}$ *such that* $\Pr(u(X) > 0, A = a) > 0$*, where* $A = \alpha(X)$*. Then, for almost every* $\mathcal{U}$*-fine distribution of* X *and* Y *on* $\mathcal{X} \times \mathcal{Y}$*, any decision policy* $d(x)$ *satisfying equalized odds is strongly Pareto dominated.*

Proof Consider the following maps. Distributions of X and $Y(1)$, i.e., probability measures on $\mathcal{X} \times \mathcal{Y}$, can be embedded in the space of joint distributions on X, $Y(0)$, and $Y(1)$ via pushing forward by the map ι, where $\iota : (x, y) \mapsto (x, y, y)$. Likewise, given a fixed decision policy $D = d(X)$, joint distributions of X, $Y(0)$, and $Y(1)$ can be projected onto the space of joint distributions of X and Y by pushing forward by the map $\pi_d : (x, y_0, y_1) \mapsto (x, y_{d(x)})$. Lastly, we see that the composition of ι and π_d—regardless of our choice of $d(x)$—is the

identity, as shown in the diagram below.



We note also that counterfactual equalized odds holds for μ exactly when equalized odds holds for $\mu \circ (\pi_d \circ \iota)^{-1}$. The result follows immediately from this and Theorem 17. ■

F.6 Proof of Theorem 9

The proof of Theorem 9 is simpler than the proof of Theorem 17, but uses some of the same machinery. As before, let $\mathcal{K} = \mathcal{A} \times [0,1]$ denote the state space, and $\mathbb{K}$ denote the set of measures on $[0,1] \times \mathcal{A}$. Let $\mathbf{K}$ denote the measures $\mu \in \mathbb{K}$ that are absolutely continuous with respect to $\lambda \times \delta$, where λ is Lebesgue measure and δ is the counting measure on $\mathcal{A}$—i.e., measures such that the restriction to $[0,1] \times \{a\}$ has a density for all $a \in \mathcal{A}$. Applying Cor. 47 with $\mathcal{U} = \{\pi\}$, where $\pi : (u, a) \mapsto u$, shows that $\mathbf{K}$ is a Banach space. As before, we let $\mathbf{Q} \subseteq \mathbf{K}$ denote the probability simplex, i.e., the set of all $\mu \in \mathbf{K}$ such that $\mu[\mathcal{K}] = 1$ and $\mu[E] \geq 0$ for all Borel sets E.

Let $R = r(X)$ denote risk.[40] By Lemma 31, we have that a policy is utility maximizing if and only if $\mathbb{1}_{R>t} \leq d(X) \leq \mathbb{1}_{R \geq t}$ a.s. By Since $\Pr_\mu(R = t) = 0$ for any absolutely continuous measure μ, we see that, in fact, a policy is utility maximizing if and only if $\mathbb{1}_{R>t} = d(X)$ μ-a.s.

Consider the sets

$$\mathbf{E}_{\text{FP}} = \left\{ \mu \in \mathcal{K} : (\forall a \in \mathcal{A}) \, \frac{\mathbb{E}_\mu[(1-R) \cdot \mathbb{1}_{A=a,R>t}]}{\mathbb{E}_\mu[(1-R) \cdot \mathbb{1}_{A=a}]} = \frac{\mathbb{E}_\mu[(1-R) \cdot \mathbb{1}_{R>t}]}{\mathbb{E}_\mu[1-R]} \right\},$$

and

$$\mathbf{E}_{\text{DP}} = \{ \mu \in \mathcal{K} : (\forall a \in \mathcal{A}) \, \Pr_\mu(R > t \mid A = a) = \Pr_\mu(R > t) \}.$$

We note that since the mapping $\mu \mapsto \Pr_\mu(E)$ is a continuous function on $\mathcal{K}$, the set of measures such that $\Pr_\mu(A = a, R > t) = 0$ or $\Pr_\mu(R > t) = 0$ is closed and shy. It follows that $\mathbf{E}_{\text{FP}}$ and $\mathbf{E}_{\text{DP}}$ are Borel. We note that, by definition, $\mathbf{E}_{\text{DP}}$ is the set of risk distributions such that there exists a utility-maximizing policy satisfying demographic parity (when the condition is well defined). Likewise, an application of the law of iterated expectations yields that $\mathbf{E}_{\text{FP}}$ is the set of distributions satisfying equalized false positive rates (when the condition is well-defined).

With this context in place, we can move on to the proof of Theorem 9.

Proof of Theorem 9 First we construct the probe. Choose distinct $a_0, a_1 \in \mathcal{A}$ arbitrarily, and let $\tilde{\nu}$ be defined by

$$\tilde{\nu}[E] = \frac{\lambda[E_{a_0} \cap [0,t)]}{t} - \frac{\lambda[E_{a_1} \cap [0,t)]}{t},$$

40. Here, since the measures are on the risk scale, rather than on X, we write R for notational simplicity.

where $E_a = E \cap \{A = a\}$. Then, by Lemma 62, there exists μ_0 such that $\lambda_{\mathbf{W}}[\mathcal{K} + \nu_0] > 0$, where $\mathbf{K} = \textsc{Span}(\tilde{\nu})$.

We see that

$$\begin{aligned} &\Pr_{\tilde{\nu}}(R < t, A = a_0) = 1, && \Pr_{\tilde{\nu}}(R > t, A = a_0) = 0, \\ &\Pr_{\tilde{\nu}}(R < t, A = a_1) = -1, && \Pr_{\tilde{\nu}}(R > t, A = a_1) = 0. \end{aligned}$$

It follows that

$$\Pr_{\mu+\beta\cdot\tilde{\nu}}(R > t \mid A = a_0) = \frac{e_0}{p_0 + \beta}, \qquad \Pr_{\mu+\beta\cdot\tilde{\nu}}(R > t, \mid A = a_1) = \frac{e_1}{p_1 - \beta},$$

where

$$\begin{aligned} &e_0 = \Pr_{\mu}(R > t, A = a_0), && e_1 = \Pr_{\mu}(R > t, A = a_1), \\ &p_0 = \Pr_{\mu}(A = a_0), && p_1 = \Pr_{\mu}(A = a_1). \end{aligned}$$

Note that by Lemmata 65 and 68, we can assume that $e_0, e_1 > 0$. It follows, rearranging terms, that

$$\Pr_{\mu+\beta\cdot\tilde{\nu}}(R > t \mid A = a_0) = \Pr_{\mu+\beta\cdot\tilde{\nu}}(R > t \mid A = a_1)$$

if and only if

$$\beta = \frac{e_0 \cdot p_1 - e_1 \cdot p_0}{e_0 + e_1},$$

which is a measure-zero subset of $\beta \in \mathbb{R}$. Therefore $\lambda_{\mathbf{W}}[\mathbf{E}_{\mathrm{DP}} + \mu] = 0$.

In the same way, we observe that

$$\begin{aligned} &\mathbb{E}_{\tilde{\nu}}[R \cdot \mathbb{1}_{A=a_0, R<t}] = \frac{t}{2}, && \mathbb{E}_{\tilde{\nu}}[R \cdot \mathbb{1}_{A=a_0, R>t}] = 0 \\ &\mathbb{E}_{\tilde{\nu}}[R \cdot \mathbb{1}_{A=a_1, R<t}] = -\frac{t}{2}, && \mathbb{E}_{\tilde{\nu}}[R \cdot \mathbb{1}_{A=a_1, R>t}] = 0, \end{aligned}$$

and so

$$\frac{\mathbb{E}_{\mu+\beta\cdot\tilde{\nu}}[(1-R)\cdot\mathbb{1}_{A=a_0,R>t}]}{\mathbb{E}_{\mu+\beta\cdot\tilde{\nu}}[(1-R)\cdot\mathbb{1}_{A=a_0}]} = \frac{e_0'}{p_0' + \beta\cdot\frac{t}{2}}, \qquad \frac{\mathbb{E}_{\mu+\beta\cdot\tilde{\nu}}[(1-R)\cdot\mathbb{1}_{A=a_1,R>t}]}{\mathbb{E}_{\mu+\beta\cdot\tilde{\nu}}[(1-R)\cdot\mathbb{1}_{A=a_1}]} = \frac{e_1'}{p_1' - \beta\cdot\frac{t}{2}},$$

where

$$\begin{aligned} &e_0' = \mathbb{E}_{\mu}[(1-R)\cdot\mathbb{1}_{A=a_0,R>t}], && e_1' = \mathbb{E}_{\mu}[(1-R)\cdot\mathbb{1}_{A=a_a,R>t}], \\ &p_0' = \mathbb{E}_{\mu}[(1-R)\cdot\mathbb{1}_{A=a_0}], && p_1' = \mathbb{E}_{\mu}[(1-R)\cdot\mathbb{1}_{A=a_0}]. \end{aligned}$$

As before, $\mu + \beta\cdot\tilde{\nu} \in \mathbf{E}_{\mathrm{FP}}$ if and only if

$$\beta = \frac{2}{t}\cdot\frac{e_0\cdot p_1 - e_1\cdot p_0}{e_0 + e_1},$$

which is again a measure-zero subset of $\beta \in \mathbb{R}$. Therefore $\lambda_{\mathbf{W}}[\mathbf{E}_{\mathrm{FP}} + \mu] = 0$.

Therefore it follows that both $\mathbf{E}_{\mathrm{FP}}$ and $\mathbf{E}_{\mathrm{DP}}$ are shy. ■

F.7 Proof of Corollary 18

The proof of Corollary 18—restated in full detail as Corollary 71 below—is a straightforward application of Theorem 17. We begin by giving the complete statement.

Corollary 71 *Consider a utility of the form given in Eq.* (12)*, where v is monotonically increasing in both coordinates and $m(x) \geq 0$. Suppose that for all $a \in \mathcal{A}$ there exists an $\{m\}$-fine distribution of X such that $\Pr(m(X) > 0, A = a) > 0$, where $A = \alpha(X)$. Then,*

- *For almost every $\{m\}$-fine distribution of X and $Y(1)$, no utility-maximizing decision policy satisfies counterfactual equalized odds.*
- *If $|\textsc{Img}(\omega)| < \infty$ and there exists an $\{m\}$-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 $\{m\}$-fine joint distribution of X, $Y(0)$, and $Y(1)$, no utility-maximizing decision policy satisfies conditional principal fairness.*
- *If $|\textsc{Img}(\omega)| < \infty$ and there exists a $\{m\}$-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 $\{m\}^{\mathcal{A}}$-fine joint distributions of A and the counterfactuals $X_{\Pi,A,a'}$, no utility-maximizing decision policy satisfies path-specific fairness.*

Recall that for notational simplicity, we refer to the distinguished utility as u^*, where

$$u^*(d) = v\left(\mathbb{E}[m(X) \cdot d(X)], \mathbb{E}[\mathbb{1}_{\alpha(X)=a_1} \cdot d(X)]\right).$$

Proof of Corollary 18 Consider the subset S of $\mathbb{R}^2$ consisting of all pairs

$$(\mathbb{E}[m(X) \cdot d(X)], \mathbb{E}[\mathbb{1}_{\alpha(X)=a_1} \cdot d(X)]),$$

where d ranges over feasible policies. We note that, for $\theta \in [0, 1]$,

$$\theta \cdot \mathbb{E}[m(X) \cdot d_0(X)] + [1 - \theta] \cdot \mathbb{E}[m(X) \cdot d_1(X)] = \mathbb{E}[m(X) \cdot (\theta \cdot d_0(X) + [1 - \theta] \cdot d_1(X))],$$

and similarly for $\mathbb{E}[\mathbb{1}_{\alpha(X)=a_1} \cdot d(X)]$. Likewise,

$$\theta \cdot \mathbb{E}[d_0(X)] + [1 - \theta] \cdot \mathbb{E}[d_1(X)] = \mathbb{E}[\theta \cdot d_0(X) + (1 - \theta) \cdot d_1(X)],$$

and so convex combinations of feasible policies are feasible. It follows that S is convex.

Now, consider a point (x_0, y_0) at which $v(x, y)$ is maximized on S. Since v is monotonically increasing in both coordinates, we must have that

$$S \cap \left((x_0, y_0) + \mathbb{R}^2_{\geq 0}\right) = \{(x_0, y_0)\},$$

and so, by the separating hyperplane theorem, there exists $(h_0, h_1) \in \mathbb{R}^2$ and $t \in \mathbb{R}$ such that

$$\begin{aligned}
(h_0, h_1)^\top (x_0, y_0) &= t, \\
(h_0, h_1)^\top \mathbf{c} &> t, \quad \left(\mathbf{c} \in (x_0, y_0) + \mathbb{R}^2_{\geq 0}, \mathbf{c} \neq (x_0, y_0)\right) \\
(h_0, h_1)^\top \mathbf{s} &< t. \quad (\mathbf{s} \in S, \mathbf{s} \neq (x_0, y_0))
\end{aligned}$$

We note that it follows that both h_0 and h_1 are positive, since, otherwise, without loss of generality, if $h_0 = 0$, then

$$(h_0, h_1)^\top (x_0 + \epsilon, y_0) = h_1 \cdot y_0 = (h_0, h_1)^\top (x_0, y_0) = 0,$$

contrary to assumption, since

$$(x_0 + \epsilon, y_0) \in (x_0, y_0) + \mathbb{R}^2_{\geq 0}, (x_0 + \epsilon, y_0) \neq (x_0, y_0).$$

Let $\lambda_* = h_1/h_0$, and consider the collection of utilities

$$\mathcal{U} = \{m(x) + \lambda \cdot \mathbb{1}_{\alpha(x)=a_1}\}_{\lambda > 0}.$$

Since $m(x) \geq 0$, $\mathcal{U}$ is consistent modulo α.

We need the further claim that any policy that is utility maximizing for some $u(x) \in \mathcal{U}$ is Pareto efficient. For, suppose that $d_0(x)$ were utility-maximizing for $u_0(x)$ but not Pareto efficient, i.e., there existed $d_1(x)$ and $u_1(x)$ such that $u_0(d_1) = u_0(d_0)$ but $u_1(d_1) > u_1(d_0)$. Then, we would have for any $u \in \mathcal{U}$ that

$$\begin{aligned} u(d_1) - u(d_0) &= u(d_1) - u(d_0) - (u_0(d_1) - u_0(d_0)) \\ &= (\lambda - \lambda_0) \cdot \left(\mathbb{E}[d_1(X) \cdot \mathbb{1}_{\alpha(X)=a_1}] - \mathbb{E}[d_0(X) \cdot \mathbb{1}_{\alpha(X)=a_1}] \right). \end{aligned}$$

First suppose that $\lambda_1 > \lambda_0$. Then, it follows from the fact that $u_1(d_1) > u_1(d_0)$ that

$$\mathbb{E}[d_1(X) \cdot \mathbb{1}_{\alpha(X)=a_1}] > \mathbb{E}[d_0(X) \cdot \mathbb{1}_{\alpha(X)=a_1}].$$

Now, if $0 < \lambda_2 < \lambda_1$ and $u(x) = m(x) + \lambda_2 \cdot \mathbb{1}_{\alpha(x)=a_1}$, then we have that $u_2(d_1) < u_2(d_0)$. In the same way, if $\lambda_1 < \lambda_0$, we could choose $\lambda_2 > \lambda_1$ such that $u_2(d_1) < u_2(d_0)$. Therefore d_1 does not Pareto dominate d_0, contrary to hypothesis. Therefore, any policy that is utility maximizing for some $u(x) \in \mathcal{U}$ is Pareto efficient.

In particular, it follows that the policy that maximizes $u(x) = m(x) + \lambda_* \cdot \mathbb{1}_{\alpha(x)=a_1}$ in expectation is Pareto efficient. By the construction of the separating hyperplane, this is also the policy that maximizes u^*, and so the policy that maximizes u^* is Pareto efficient. Therefore, under the hypotheses of Theorem 17, for almost every joint distribution, the utility maximizing policy does not satisfy counterfactual equalized odds, conditional principal fairness, or path-specific fairness. ■

F.8 General Measures on $\mathcal{K}$

Theorem 17 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 can 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 ensure causal fairness notions, such as counterfactual equalized odds, hold in a locally robust way. In particular, the generalization of Theorem 17 beyond $\mathcal{U}$-fine measures to all totally bounded measures on the state space is false, as illustrated by the following proposition.

Proposition 72 *Consider the set $\mathbf{E}' \subseteq \mathbb{K}$ 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 *relatively shy.*

Proof We adopt the notational conventions of Section F.3. We note that by Prop. 36, 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 $0 \le a \le b < 1$.

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 $\mathbb{E}[\tau(X)] = \frac{3}{4} = b$. Since $\tau(x)$ is a threshold policy and exhausts the budget, it is utility maximizing by Lemma 31. Moreover, if $D = \mathbb{1}_{U_D \le \tau(X)}$ for some $U_D \sim \textsc{Unif}(0, 1)$ independent of X and $Y(1)$, then $D \perp\!\!\!\perp A \mid Y(1)$. Since $u(X) \perp\!\!\!\perp A, Y(1)$, it follows that

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

Therefore Eq. (4) 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 policies. 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

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

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

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

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

$$\begin{aligned} &\pi_{a_1,0} \cdot (q_{a_0,0} + p_{a_0,0} \cdot r_{a_0,0}) \\ &\qquad = \pi_{a_0,0} \cdot (q_{a_1,0} + p_{a_1,0} \cdot r_{a_1,0}), \\ &\pi_{a_1,1} \cdot (q_{a_0,1} + p_{a_0,1} \cdot r_{a_0,1}) \\ &\qquad = \pi_{a_0,1} \cdot (q_{a_1,1} + p_{a_1,1} \cdot r_{a_1,1}), \end{aligned} \tag{34}$$

since, as above,

$$\Pr(D = 1 \mid A = a, Y(1) = y) = \mathbb{E}[\tau'(X) \mid A = a, Y(1) = y].$$

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. (34) 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 31. 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. ■

Appendix G. Theorem 11 and Related Results

We first prove a variant of Theorem 11 for general, continuous covariates $\mathcal{X}$. Then, we extend and generalize Theorem 11 using the theory of finite Markov chains, offering a proof of the theorem different from the sketch included in the main text.

G.1 Extension to Continuous Covariates

Here we follow the proof sketch in the main text for Theorem 11, 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 73 *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,

$$\mathbb{E}[f(X,U) \mid X] = \int f(X,u)\,\mathrm{d}F_U(u), \tag{35}$$

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. (35), 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. (35), we have that for all $a' \in \mathcal{A}$,

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

This is exactly Eq. (8), and so the result follows. ■

We are now ready to prove a continuous variant of Theorem 11. 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 74 *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 $0 \leq p \leq 1$).

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 $a \in \mathcal{A}$,

$$\Pr(d(X) = d_{\max} \mid A = a) = 0. \tag{36}$$

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

$$\Pr(d_{\max} - d(X) < \epsilon \mid A = a_0) > 0 \tag{37}$$

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. (8). But, under the assumption in Eq. (36), 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 73. Rearranging, it follows that for any $\rho > 0$, a.s.,

$$\Pr(d_{\max} - d(X_{\Pi,A,a_1}) < \rho \mid X) \geq 1 - \frac{d_{\max} - d(X)}{\rho}. \tag{38}$$

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. (36), 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. (38), 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 73. 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 73, establishing the result. ■

G.2 A Markov Chain Perspective

The theory of Markov chains illuminates—and allows us to extend—the proof of Theorem 11. Suppose $\mathcal{X} = \{x_1, \ldots, x_n\}$.[41] For any $a' \in \mathcal{A}$, let $P_{a'} = [p^{a'}_{i,j}]$ where $p^{a'}_{i,j} = \Pr(X_{\Pi,A,a'} = x_j \mid X = x_i)$. 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 73 can be recast as stating that when $W = X$,

41. 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.

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 75 *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*

$$d_i = \sum_{k=1}^{m} \left[\lim_{n\to\infty} \sum_{j \in S_k} P_{ij}^n \right] \cdot p_k. \tag{39}$$

Remark 76 *We note that $\lim_{n\to\infty} \sum_{j\in S_k} P_{i,j}^n$ 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 $M^{\textsc{Abs}}$ where the set of states of $M^{\textsc{Abs}}$ 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 $P^{\textsc{Abs}}$ associated with $M^{\textsc{Abs}}$ is given by

$$P^{\textsc{Abs}} = \begin{bmatrix} I & \\ R & Q \end{bmatrix}$$

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 $P^{\textsc{Abs}}$, 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 and 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 77 *Suppose $\mathcal{X}$ is finite, and define the stochastic matrix $P = \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} P_{a'}$. If $d(x)$ is a Π-fair policy then it is constant on the recurrent classes of P.*

Proof By Lemma 73, d is Π-fair if and only if $P_{a'}d = d$ for all $a' \in \mathcal{A}$. Therefore,

$$\frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} P_{a'}d = \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} d = d, \tag{40}$$

and so d is a 1-eigenvector of P. Therefore it is constant on the recurrent classes of P by Lemma 75. ■

We note that Theorem 11 follows immediately from this.

Proof of Theorem 11 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. ■

Appendix H. Proofs of Propositions 19 and 20

The proofs of Proposition 19 and Proposition 20 rely on certain shared theory about beta distributions. We begin by reviewing this theory before moving onto the proofs of the respective propositions.

H.1 Beta Distributions and Stochastic Dominance

We begin by introducing incomplete beta functions and distributions.

Definition 78 *The* incomplete beta function $I_t(\alpha, \beta)$ *for* $t \in (0, 1]$ *is given by*

$$I_t(\alpha, \beta) = \int_0^t x^{\alpha-1}(1-x)^{\beta-1}.$$

A random variable X *is said to be distributed as* $\textsc{Beta}_t(\alpha, \beta)$ *if* $X \sim Y \mid Y < t$ *for* $Y \sim \textsc{Beta}(\alpha, \beta)$*; equivalently, if* X *has PDF*

$$\mathbb{1}_{x\in(0,t)} \cdot \frac{x^{\alpha-1}(1-x)^{\beta-1}}{I_t(\alpha, \beta)}.$$

An important property relating different beta distributions is stochastic dominance.

Definition 79 *Let* X *and* Y *be random variables with CDFs* $F_X(t)$ *and* $F_Y(t)$*, respectively. We say that* X stochastically dominates Y*, written* $Y \leq_{\text{st}} X$*, if* $F_X(t) \leq F_Y(t)$ *for all* $t \in \mathbb{R}$*. We say that* X strictly stochastically dominates Y*, i.e.,* $Y <_{\text{st}} X$*, if* $F_X(t) = F_Y(t)$ *implies that either* $F_X(t) = F_Y(t) = 0$ *or* $F_X(t) = F_Y(t) = 1$*.*

Stochastic dominance has the following useful property.

Lemma 80 *Suppose* $Y \leq_{\text{st}} X$*. Then, for any monotonically non-decreasing* $\varphi : \mathbb{R} \to \mathbb{R}$*, we have that* $\mathbb{E}[\varphi(Y)] \leq \mathbb{E}[\varphi(X)]$*.*

For proof, see (1.A.7) in Shaked and Shanthikumar (2007). We will need to improve the inequality to a strict inequality. We begin with the simplest case.

Lemma 81 *Suppose* $Y <_{\text{st}} X$*, where* X *and* Y *are positive and* $F_X(t)$*,* $F_Y(t)$ *are continuous. Then,* $\mathbb{E}[Y \cdot \mathbb{1}_{Y>t}] < \mathbb{E}[X \cdot \mathbb{1}_{X>t}]$ *for any* $t > 0$ *such that there exists* $t' \geq t$ *with* $F_X(t') < F_Y(t')$*.*

Proof Since since $F_Y(x) \geq F_X(x)$ for all x and, in particular, $F_Y(x) > F_X(x)$ on an open interval containing t', we have that

$$\begin{aligned}
\mathbb{E}[Y \cdot \mathbb{1}_{Y>t}] &= t \cdot [1 - F_Y(t)] + \int_t^\infty 1 - F_Y(x)\,\mathrm{d}x \\
&< t \cdot [1 - F_X(t)] + \int_t^\infty 1 - F_X(x)\,\mathrm{d}x \\
&= \mathbb{E}[X \cdot \mathbb{1}_{X>t}]
\end{aligned}$$

where we have applied the Darth Vader rule to calculate the expectations (Muldowney et al., 2012). ■

This leads to the following modest technical generalization.

Lemma 82 *Suppose* $Y <_{\text{st}} X$*, where* X *and* Y *are positive and* $F_X(t)$ *and* $F_Y(t)$ *are continuous. Suppose that* $t > 0$ *is such that there exists* $t' < t$ *with* $F_X(t') < F_Y(t')$*, and* $f : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ *is monotonically decreasing and continuous. Then*

$$\mathbb{E}[f(Y) \cdot \mathbb{1}_{Y<t}] > \mathbb{E}[f(X) \cdot \mathbb{1}_{X<t}].$$

Proof Consider the transformed variables $f(X)$ and $f(Y)$. Since f is monotonically decreasing and continuous, it is invertible, and in particular

$$\Pr(f(X) < x) = \Pr(X > f^{-1}(x)) = 1 - F_X(f^{-1}(x)).$$

It follows that the CDFs of $f(X)$ and $f(Y)$ are $1 - F_X(f^{-1}(x))$ and $1 - F_Y(f^{-1}(x))$, respectively. Then, observe that since $F_X(x) \leq F_Y(x)$ for all $x \in \mathbb{R}$,

$$1 - F_X(x) \geq 1 - F_Y(x)$$

for all $x \in \mathbb{R}$, i.e., $f(X) \leq_{\text{st}} f(Y)$. In particular, since $F_X(x) = 0$ if and only if $1 - F_X(x) = 1$ and *vice versa*, it follows from the invertibility of f that $f(X) <_{\text{st}} f(Y)$. Therefore, after noting that $X < t$ if and only if $f(X) > f(t)$ and similarly for Y, the result follows from Lemma 81. ■

It is relatively straightforward to characterize the stochastic dominance relationships between various (incomplete) beta distributions according to α and β; see Arab et al. (2021) for full details. Here, we require only the following result, which closely follows the proof of Theorem 1 there.

Lemma 83 *If $X \sim \text{Beta}_t(\alpha_0, \beta_0)$, $Y \sim \text{Beta}_t(\alpha_1, \beta_1)$, where $\alpha_0 \geq \alpha_1$ and $\beta_0 \leq \beta_1$, then $Y \leq_{\text{st}} X$. If, in addition, either $\alpha_0 > \alpha_1$ or $\beta_0 < \beta_1$, then $Y <_{\text{st}} X$.*

Proof Consider the CDFs $F_X(s)$ and $F_Y(s)$. We will use the difference $G(s) = F_X(s) - F_Y(s)$ to demonstrate the result. The case where $\alpha_0 = \alpha_1$ and $\beta_0 = \beta_1$ is trivial, so we restrict our attention to the case where one of the inequalities is strict. For simplicity, we assume that $\alpha_0 > \alpha_1$; the case where $\beta_0 < \beta_1$ is virtually identical.

In particular, observe that $G(0) = G(t) = 0$, and that for $s \in (0, t)$,

$$\begin{aligned} G'(s) &= \frac{s^{\alpha_0 - 1}(1-s)^{\beta_0 - 1}}{I_t(\alpha_0, \beta_0)} - \frac{s^{\alpha_1 - 1}(1-s)^{\beta_1 - 1}}{I_t(\alpha_1, \beta_1)} \\ &= \frac{s^{\alpha_0 - 1}(1-s)^{\beta_0 - 1}}{I_t(\alpha_1, \beta_1)} \cdot \left[s^{\alpha_1 - \alpha_0}(1-s)^{\beta_1 - \beta_0} - \frac{I_t(\alpha_1, \beta_1)}{I_t(\alpha_0, \beta_0)} \right]. \end{aligned}$$

We consider the two multiplicands in the final expression. We note that the first is greater than zero for all $s \in (0, t)$. Therefore, for $s \in (0, t)$, $G'(s) = 0$ if and only if

$$s^{\alpha_1 - \alpha_0}(1-s)^{\beta_1 - \beta_0} = \frac{I_t(\alpha_1, \beta_1)}{I_t(\alpha_0, \beta_0)}.$$

Now, since $\alpha_0 > \alpha_1$ and $\beta_0 \leq \beta_1$, it follows that the left-hand side of the previous expression is strictly decreasing. In particular, $G'(s) = 0$ for at most one $s \in (0, t)$. Since $G(0) = G(t) = 0$ and $G(t)$ is non-constant, it follows that $G'(s) = 0$ for exactly one $s \in (0, t)$ by Rolle's theorem. In particular, either $G(s) > 0$ for all $s \in (0, t)$ or $G(s) < 0$ for all $t \in (0, t)$.

Consequently,

$$s^{\alpha_1 - \alpha_0}(1-s)^{\beta_1 - \beta_0} - \frac{I_t(\alpha_1, \beta_1)}{I_t(\alpha_0, \beta_0)}$$

changes sign for some $s_0 \in (0,t)$. Since the minuend is strictly decreasing, it follows that $G'(s) > 0$ for $s \in (0, s_0)$. Therefore, in particular, $G(s) > 0$ for all $s \in (0, s_0)$, and hence for $s \in (0,t)$. Therefore $F_X(s) > F_Y(s)$ for $s \in (0,t)$, and so $Y <_{\text{st}} X$. ■

H.2 Proof of Proposition 19

Our proof is a relatively straightforward application of Sard's theorem. Intuitively, counterfactual predictive parity imposes three constraints on α_0, β_0, α_1, β_1, and the group-specific thresholds t_0 and t_1. These constraints are sufficiently smooth that the zero locus should take the form of a 3-manifold, by the inverse function theorem. Projecting onto the first four coordinates (i.e., eliminating t_0 and t_1) gives rise to a set of measure zero by Sard's theorem.

Proof of Proposition 19 We begin by noting that the risk distributions, conditional on $Y(1) = i$, for $i = 0, 1$, take a particularly simple form.

$$\begin{aligned} r(X) \mid A = a_i, Y(1) = 1 &\sim \textsc{Beta}(\alpha_i + 1, \beta_i), \\ r(X) \mid A = a_i, Y(1) = 0 &\sim \textsc{Beta}(\alpha_i, \beta_i + 1). \end{aligned}$$

This follows upon noting that

$$\begin{aligned} \Pr(r(X) < t, Y(1) = 1 \mid A = a_i) &= \frac{1}{B(\alpha_i, \beta_i)} \int_0^1 x \cdot \mathbb{1}_{x<t} \cdot x^{\alpha_i - 1}(1-x)^{\beta_i - 1} \, \mathrm{d}x \\ &= \frac{I_t(\alpha_i + 1, \beta_i)}{B(\alpha_i, \beta_i)}, \\ \Pr(r(X) < t, Y(1) = 0 \mid A = a_i) &= \frac{1}{B(\alpha_i, \beta_i)} \int_0^1 (1-x) \cdot \mathbb{1}_{x<t} \cdot x^{\alpha_i - 1}(1-x)^{\beta_i - 1} \, \mathrm{d}x \\ &= \frac{I_t(\alpha_i, \beta_i + 1)}{B(\alpha_i, \beta_i)}. \end{aligned}$$

By Cor. 29, if there exists a Pareto efficient policy satisfying counterfactual equalized odds, then it must correspond to some multiple threshold policy. In particular, by Eq. (4) and the fact that the policy is budget exhausting, and using Prop. 15 with $\lambda = 0$, there must be t_0, t_1 such that

$$\begin{aligned} \frac{I_{t_0}(\alpha_0, \beta_0)}{B(\alpha_0, \beta_0)} + \frac{I_{t_1}(\alpha_1, \beta_1)}{B(\alpha_1, \beta_1)} &= 2 - b \\ \frac{I_{t_0}(\alpha_0 + 1, \beta_0)}{B(\alpha_0 + 1, \beta_0)} &= \frac{I_{t_1}(\alpha_1 + 1, \beta_1)}{B(\alpha_1 + 1, \beta_1)} \\ \frac{I_{t_0}(\alpha_0, \beta_0 + 1)}{B(\alpha_0, \beta_0 + 1)} &= \frac{I_{t_1}(\alpha_1, \beta_1 + 1)}{B(\alpha_1, \beta_1 + 1)}. \end{aligned}$$

Here, the first equality encodes budget exhaustion, the second the fact that $\Pr(D = 1 \mid A = a_0, Y(1) = 1) = \Pr(D = 1 \mid A = a_1, Y(1) = 1)$, and the third the fact that $\Pr(D = 1 \mid A = a_0, Y(1) = 0) = \Pr(D = 1 \mid A = a_1, Y(1) = 0)$.

Let $\mathbf{f} : \mathbb{R}^6 \to \mathbb{R}^3$ be given by

$$f_0(\alpha_0, \beta_0, t_0, \alpha_1, \beta_1, t_1) = \frac{I_{t_0}(\alpha_0, \beta_0)}{B(\alpha_0, \beta_0)} + \frac{I_{t_1}(\alpha_1, \beta_1)}{B(\alpha_1, \beta_1)} - (2 - b),$$
$$f_1(\alpha_0, \beta_0, t_0, \alpha_1, \beta_1, t_1) = \frac{I_{t_0}(\alpha_0 + 1, \beta_0)}{B(\alpha_0 + 1, \beta_0)} - \frac{I_{t_1}(\alpha_1 + 1, \beta_1)}{B(\alpha_1 + 1, \beta_1)},$$
$$f_2(\alpha_0, \beta_0, t_0, \alpha_1, \beta_1, t_1) = \frac{I_{t_0}(\alpha_0, \beta_0 + 1)}{B(\alpha_0, \beta_0 + 1)} - \frac{I_{t_1}(\alpha_1, \beta_1 + 1)}{B(\alpha_1, \beta_1 + 1)}.$$

Then, given α_0, β_0, α_1 and β_1, there exists a Pareto optimal policy satisfying counterfactual equalized odds only if there exist t_0 and t_1 such that

$$\mathbf{f}(\alpha_0, \beta_0, t_0, \alpha_1, \beta_1, t_1) = \mathbf{0}.$$

Let $D\mathbf{f}$ denote the Jacobian of $\mathbf{f}$. If we can show that $\mathbf{f}$ is smooth and $D\mathbf{f}$ has full rank, then the proof is complete. For, by Theorem 5.12 in Lee (2013), it follows that $\mathbf{f}^{-1}(\mathbf{0}) \subseteq \mathbb{R}^6$ is a smooth 3-manifold. The restriction of the map

$$\pi : (\alpha_0, \beta_0, t_0, \alpha_1, \beta_1, t_1) \mapsto (\alpha_0, \beta_0, \alpha_1, \beta_1)$$

to $\mathbf{f}^{-1}(\mathbf{0})$ is smooth, and so by Sard's theorem (see Theorem 6.10 in Lee (2013)), since $\mathbb{R}^4_{>0}$ is a trivially a smooth 4-manifold, the measure of the singular values of $\pi \restriction_{\mathbf{f}^{-1}(\mathbf{0})}$ is zero. However, since the maximum rank of $D\pi$ on $\mathbf{f}^{-1}(\mathbf{0})$ is three, every point of $\mathbf{f}^{-1}(\mathbf{0})$ is singular, and consequently, the whole image of π is singular, i.e., $\pi(\mathbf{f}^{-1}(\mathbf{0}))$ has measure zero. However, as argued above, the set of $(\alpha_0, \beta_0, \alpha_1, \beta_1)$ such that there exists a Pareto efficient distribution satisfying counterfactual equalized odds is a subset of $\pi(\mathbf{f}^{-1}(\mathbf{0}))$.

Therefore, it remains only to show that $\mathbf{f}$ is smooth, and that $D\mathbf{f}$ has full rank for all $(\alpha_0, \beta_0, t_0, \alpha_1, \beta_1, t_1) \in \mathbb{R}^2_{>0} \times [0, 1] \times \mathbb{R}^2_{>0} \times [0, 1]$. This verification is a routine exercise in linear algebra and multivariable calculus, and is given below.

Smoothness. We note that since smooth functions are closed under composition, it suffices to show that $I_t(\alpha, \beta)$ is a smooth function of t, α, and β. First, we consider partial derivatives with respect to α and β. If we could differentiate under the integral sign, then we would have that

$$\frac{\partial^{n+m}}{\partial \alpha^n \partial \beta^m} I_t(\alpha, \beta) = \int_0^t \log(x)^n \log(1 - x)^m x^{\alpha - 1}(1 - x)^{\beta - 1} \, \mathrm{d}x. \tag{41}$$

Recall the well-known condition for the Leibniz integral rule that if $\frac{\partial}{\partial x}\phi(x, t)|_{x=x'}$ is dominated by some integrable $g(t)$ for all x' in some neighborhood of x, then[42]

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_0^t \phi(x, t) \, \mathrm{d}t = \int_0^t \frac{\partial}{\partial x} \phi(x, t) \, \mathrm{d}t.$$

Since the integrand $\log(x)^n \log(1-x)^m x^{\alpha-1}(1-x)^{\beta-1}$ is strictly decreasing in both α and β for $x \in (0, 1)$, it suffices merely to show that $\log(x)^n \log(1-x)^m x^{\alpha-1}(1-x)^{\beta-1}$ is integrable on $(0, 1)$ for all $\alpha > 0$ and $\beta > 0$. Moreover, since

42. See, e.g., Theorem 6.28 in Klenke (2020).

- The whole integrand is bounded on $(\epsilon, 1-\epsilon)$,
- The factor $\log(1-x)^m(1-x)^{\beta-1}$ is bounded on $(0,\epsilon)$,
- The factor $\log(x)^n x^{\alpha-1}$ is bounded on $(1-\epsilon, 1)$,

it suffices to show that $\log(x)^n x^{\alpha-1}$ is integrable on $(0,\epsilon)$ and that $\log(1-x)^m(1-x)^{\beta-1}$ is integrable on $(1-\epsilon,1)$. Up to the change of variables $x \mapsto 1-x$, since n, m, α, and β are arbitrary, we see that it suffices to verify that $\log(x)^n x^{\alpha-1}$ alone is integrable on $(0,1)$. Integrating by parts, we have that

$$\int_0^t \log(x)^n x^{\alpha-1}\,\mathrm{d} = \left[\frac{\log(x)^n x^\alpha}{\alpha}\right]_0^1 - n\int_0^1 \log(x)^{n-1}x^{\alpha-1}\,\mathrm{d}x.$$

Since, by l'Hôpital's rule, $\lim_{x\to 0}\log(x)^n x^\alpha = 0$, this expression equals

$$-n\int_0^1 \log(x)^{n-1}x^{\alpha-1}\,\mathrm{d}x.$$

Since $x^{\alpha-1}$ is integrable on $0,1$ we see inductively that so is $\log(x)^n x^{\alpha-1}$. Therefore, we can differentiate under the integral sign, and Eq. (41) holds.

Now, taking derivatives with respect to t yields that

$$\frac{\partial^{n+m+1}}{\partial t\partial\alpha^n\partial\beta^m} I_t(\alpha,\beta) = \log(t)^n\log(1-t)^m t^{\alpha-1}(1-t)^{\beta-1},$$

which is a polynomial in smooth functions of t, and hence is smooth in t. Therefore

$$\frac{\partial^{n+m+k}}{\partial t^k\partial\alpha^n\partial\beta^m} I_t(\alpha,\beta)$$

exists and is a polynomial in t for all $k > 1$.

By continuity, the orders of the partial derivatives can be switched arbitrarily (see, e.g., Theorem 9.40 in Rudin (1976)), and so it follows that $\mathbf{f}$ is smooth.

Full rank. By the rank-nullity theorem, it suffices to show that the column rank of $D\mathbf{f}$ is three. However, letting $B_{\alpha,\beta} \sim \textsc{Beta}(\alpha,\beta)$ we have, by the results of the previous section, that the first three columns (i.e., partial derivatives with respect to α_0, β_0, and t_0, respectively) of $D\mathbf{f}$ are given by

$$\begin{bmatrix} \mathbb{E}[\log(B_{\alpha_0,\beta_0})\cdot \mathbb{1}_{B_{\alpha_0,\beta_0}<t_0}] & \mathbb{E}[\log(B_{\beta_0,\alpha_0})\cdot \mathbb{1}_{B_{\beta_0,\alpha_0}<1-t_0}] & \frac{t_0^{\alpha_0-1}\cdot(1-t_0)^{\beta_0-1}}{B(\alpha_0,\beta_0)} \\ \mathbb{E}[\log(B_{\alpha_0+1,\beta_0})\cdot \mathbb{1}_{B_{\alpha_0+1,\beta_0}<t_0}] & \mathbb{E}[\log(B_{\beta_0,\alpha_0+1})\cdot \mathbb{1}_{B_{\beta_0,\alpha_0+1}<1-t_0}] & \frac{t_0^{\alpha_0}\cdot(1-t_0)^{\beta_0-1}}{B(\alpha_0+1,\beta_0)} \\ \mathbb{E}[\log(B_{\alpha_0,\beta_0+1})\cdot \mathbb{1}_{B_{\alpha_0,\beta_0+1}<t_0}] & \mathbb{E}[\log(B_{\beta_0+1,\alpha_0})\cdot \mathbb{1}_{B_{\beta_0+1,\alpha_0}<1-t_0}] & \frac{t_0^{\alpha_0-1}\cdot(1-t_0)^{\beta_0}}{B(\alpha_0,\beta_0+1)} \end{bmatrix}.$$

It is easy to see that the first two columns are necessarily independent, since all entries are negative but, by Lemmata 82 and 83, taking $f(x) = -\log(x)$, we have that

$$\mathbb{E}[\log(B_{\alpha_0+1,\beta_0})\cdot \mathbb{1}_{B_{\alpha_0+1,\beta_0}<t_0}] > \mathbb{E}[\log(B_{\alpha_0,\beta_0+1})\cdot \mathbb{1}_{B_{\alpha_0,\beta_0+1}<t_0}]$$

while

$$\mathbb{E}[\log(B_{\beta_0,\alpha_0+1}) \cdot \mathbb{1}_{B_{\beta_0,\alpha_0+1}<1-t_0}] < \mathbb{E}[\log(B_{\beta_0+1,\alpha_0}) \cdot \mathbb{1}_{B_{\beta_0+1,\alpha_0}<1-t_0}].$$

Now observe that the final three columns of $D\mathbf{f}$—i.e., its partial derivatives with respect to α_1, β_1, and t_1, respectively—are given by

$$\begin{bmatrix} \mathbb{E}[\log(B_{\alpha_1,\beta_1}) \cdot \mathbb{1}_{B_{\alpha_1,\beta_1}<t_1}] & \mathbb{E}[\log(B_{\beta_1,\alpha_1}) \cdot \mathbb{1}_{B_{\beta_1,\alpha_1}<1-t_1}] & \frac{t_1^{\alpha_1-1}\cdot(1-t_1)^{\beta_1-1}}{B(\alpha_1,\beta_1)} \\ -\mathbb{E}[\log(B_{\alpha_1+1,\beta_1}) \cdot \mathbb{1}_{B_{\alpha_1+1,\beta_1}<t_1}] & -\mathbb{E}[\log(B_{\beta_1,\alpha_1+1}) \cdot \mathbb{1}_{B_{\beta_1,\alpha_1+1}<1-t_1}] & -\frac{t_1^{\alpha_1}\cdot(1-t_1)^{\beta_1-1}}{B(\alpha_1+1,\beta_1)} \\ -\mathbb{E}[\log(B_{\alpha_1,\beta_1+1}) \cdot \mathbb{1}_{B_{\alpha_1,\beta_1+1}<t_1}] & -\mathbb{E}[\log(B_{\beta_1+1,\alpha_1}) \cdot \mathbb{1}_{B_{\beta_1+1,\alpha_1}<1-t_1}] & -\frac{t_1^{\alpha_1-1}\cdot(1-t_1)^{\beta_1}}{B(\alpha_1,\beta_1+1)} \end{bmatrix}.$$

In particular, we notice that in the fourth column—i.e., partial derivatives with respect to α_1—the element in the first row is negative, while the elements in the second and third row are positive. Therefore, the fourth column is necessarily independent of the first two columns—i.e., the partial derivatives with respect to α_0 and β_0—in which every element is negative. Therefore, we have proven that there are three linearly independent columns, and so the rank of $D\mathbf{f}$ is full. ■

H.3 Proof of Proposition 20

To prove the proposition, we must use our characterizations of the conditional tail risks of the beta distribution proven in Appendix H.1 above. Note that in Proposition 20, for expositional clarity, 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 $v = \alpha + \beta$.

Using the theory above, we begin by proving a modest generalization of Prop. 20.

Lemma 84 *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.,

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

with $\alpha_{a_1} < \alpha_{a_0}$ and $\beta_{a_0} < \beta_{a_1}$. 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. 15, we may without loss of generality assume that there exist thresholds t_{a_0}, t_{a_1} 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,

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

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 Lemmata 80 and 83, 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. 20 follows as a corollary.

Proof of Prop. 20 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 $\beta_{a_0} < \beta_{a_1}$ and $\alpha_{a_1} < \alpha_{a_0}$, and so, by Lemma 84, the proposition follows. ■

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, 2018.

Rahul Aggarwal, Kirsten Bibbins-Domingo, Robert W. Yeh, Yang Song, Nicholas Chiu, Rishi K. Wadhera, Changyu Shen, and Dhruv S. Kazi. Diabetes screening by race and ethnicity in the United States: Equivalent body mass index and age thresholds. *Annals of Internal Medicine*, 175(6):765–773, 2022.

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

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 5 2016.

Shamena Anwar and Hanming Fang. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *The American Economic Review*, 2006.

Idir Arab, Paulo Eduardo Oliveira, and Tilo Wiklund. Convex transform order of beta distributions with some consequences. *Statistica Neerlandica*, 75(3):238–256, 2021.

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David

Piorkowski, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

Imanol Arrieta-Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 2022.

Kenneth Arrow. The theory of discrimination. In *Discrimination in labor markets*. Princeton University Press, 1973.

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

Jack M Balkin and Reva B Siegel. The American civil rights tradition: Anticlassification or antisubordination. *Issues in Legal Scholarship*, 2(1), 2003.

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76, 2018.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

Gary S Becker. *The Economics of Discrimination*. University of Chicago Press, 1957.

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).

Richard Berk. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. Springer Science & Business Media, 2012.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Marianne Bertrand and Sendhil Mullainathan. 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.

Patrick Billingsley. *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.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*, 2019.

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

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

William Cai, Johann Gaebler, Nikhil Garg, and Sharad Goel. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 22–28, 2020.

William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. Adaptive sampling strategies to construct equitable training datasets. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2022a.

William Cai, Johann Gaebler, Justin Kaashoek, Lisa Pinals, Samuel Madden, and Sharad Goel. Measuring racial and ethnic disparities in traffic enforcement with large-scale telematics data. *PNAS Nexus*, 1(4):pgac144, 2022b.

Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Alycia N Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in Big Data*, 5, 2022.

James H Carr and Isaac F Megbolugbe. *The Federal Reserve Bank of Boston study on mortgage lending revisited*. Fannie Mae Office of Housing Policy Research, 1993.

Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey data. Technical report, National Center for Health Statistics, 2011-2018. URL `https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2011`.

Jessica P Cerdeña, Marie V Plaisime, and Jennifer Tsai. From race-based to race-conscious medicine: How anti-racist uprisings call us to act. *The Lancet*, 396(10257):1125–1128, 2020.

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

Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954*, 2022.

Alex Chohlas-Wood, Joe Nudell, Keniel Yao, Zhiyuan Lin, Julian Nyarko, and Sharad Goel. Blind justice: Algorithmically masking race in charging decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 35–45, 2021.

Alex Chohlas-Wood, Madison Coots, Emma Brunskill, and Sharad Goel. Learning to be fair: A consequentialist approach to equitable decision-making. *arXiv preprint arXiv:2109.08792*, 2023a.

Alex Chohlas-Wood, Madison Coots, Sharad Goel, and Julian Nyarko. Designing equitable algorithms. *Nature Computational Science*, 3, 2023b.

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

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

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

Jens Peter Reus Christensen. On sets of Haar measure zero in Abelian Polish groups. *Israel Journal of Mathematics*, 13(3-4):255–260, 1972.

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

Ruth Colker. Anti-subordination above all: Sex, race, and equal protection. *NYUL Rev.*, 61:1003, 1986.

Madison Coots, Soroush Saghafian, David Kent, and Sharad Goel. Reevaluating the role of race and ethnicity in diabetes screening. *arXiv preprint arXiv:2306.10220*, 2023.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023v2*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 582–593, 2020.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, pages 300–332. PMLR, 2019.

Stewart J D'Alessio and Lisa Stolzenberg. Race and the probability of arrest. *Social forces*, 81(4):1381–1397, 2003.

Richard B Darlington. Another look at "cultural fairness". *Journal of Educational Measurement*, 8(2):71–82, 1971.

Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. The Public Safety Assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in Kentucky, 2018. URL `https://papers.ssrn.com/abstract=3168452`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *Proceedings of the International Conference in Learning Representations*, 2016.

Robin S Engel and Rob Tillyer. Searching for equilibrium: The tenuous nature of the outcome test. *Justice Quarterly*, 25(1):54–71, 2008.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Fisher v. University of Texas. 579 U.S., 2016.

Owen M Fiss. Groups and the Equal Protection Clause. *Philosophy & Public Affairs*, pages 107–177, 1976.

Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. A causal framework for observational studies of discrimination. *Statistics and Public Policy*, 2022.

Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Causal feature selection for algorithmic fairness. *Proceedings of the 2022 International Conference on Management of Data (SIGMOD)*, 2022.

George C Galster. The facts of lending discrimination cannot be argued away by examining default rates. *Housing Policy Debate*, 4(1):141–146, 1993.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Sharad Goel, Maya Perelman, Ravi Shroff, and David Alan Sklansky. Combatting police discrimination in the age of big data. *New Criminal Law Review: An International and Interdisciplinary Journal*, 20(2):181–232, 2017.

Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, 2000.

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

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

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

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

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. *Data Protection and Privacy*, 12(12):1, 2020.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 2020.

Brian R Hunt, Tim Sauer, and James A Yorke. Prevalence: a translation-invariant "almost every" on infinite-dimensional spaces. *Bulletin of the American Mathematical Society*, 27 (2):217–238, 1992.

Idaho H.B. 118. H.B. 118, 65th Leg., 1st Reg. Sess., 2019. `https://legislature.idaho.gov/wp-content/uploads/sessioninfo/2019/legislation/H0118.pdf`.

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

Kosuke Imai, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *arXiv preprint arXiv:2012.02845*, 2020.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 677–678, 2020a.

Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):771–800, 2020b.

Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 64–70, 2020c.

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, 2013.

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

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 656–666, 2017.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2017a.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017b.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, Cham, 2020. Third edition.

John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1), 2001.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689, 2020.

Allison Koenecke, Eric Giannella, Robb Willer, and Sharad Goel. Popular support for balancing equity and efficiency in resource allocation: A case study in online advertising to increase welfare program awareness. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 494–506, 2023.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2013.

Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 58–59, 2022.

Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

Charles F Manski, John Mullahy, and Atheendar Venkataramani. Using measures of race to make clinical predictions: Decision making, patient health, and fairness. Technical report, National Bureau of Economic Research, 2022.

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

Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

Pat Muldowney, Krzysztof Ostaszewski, and Wojciech Wojdowski. The Darth Vader rule. *Tatra Mountains Mathematical Publications*, 52(1):53–63, 2012.

Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.

Abigail Nurse. Anti-subordination in the Equal Protection Clause: A case study. *NYUL Rev.*, 89:293, 2014.

Julian Nyarko, Sharad Goel, and Roseanna Sommers. Breaking taboos in fair machine learning: An experimental study. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, 2021.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

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

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

Jessica K Paulus and David M Kent. Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digital Medicine*, 3(1):1–8, 2020.

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

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

Judea Pearl. *Causality*. Cambridge University Press, second edition, 2009b.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

Edmund S Phelps. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661, 1972.

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In *The 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and Sharad Goel. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4(7):736–745, 2020.

Richard A Primus. Equal protection and disparate impact: Round three. *Harv. L. Rev.*, 117:494, 2003.

Rajeev Ramchand, Rosalie Liccardo Pacula, and Martin Y Iguchi. Racial differences in marijuana-users' risk of arrest in the United States. *Drug and alcohol dependence*, 84(3): 264–272, 2006.

M. M. Rao. *Conditional measures and applications*, volume 271 of *Pure and Applied Mathematics (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2005.

Guy N Rothblum and Gal Yona. Decision-making under miscalibration. *arXiv preprint arXiv:2203.09852*, 2022.

Walter Rudin. *Principles of Mathematical Analysis*, volume 3. McGraw-Hill New York, 1976.

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

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

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

SFFA v. Harvard. Students for Fair Admissions, Inc., Petitioner, v. President and Fellows of Harvard College. Students for Fair Admissions, Inc., Petitioner, v. University of North Carolina, et al., 2023. https://www.supremecourt.gov/opinions/22pdf/20-1199_l6gn.pdf.

Moshe Shaked and J George Shanthikumar. *Stochastic Orders*. Springer, 2007.

Ravi Shroff. Predictive analytics for city agencies: Lessons from children's services. *Big Data*, 5(3):189–196, 2017.

Reva B Siegel. Equality talk: Antisubordination and anticlassification values in constitutional struggles over Brown. *Harv. L. Rev.*, 117:1470, 2003.

C. E. Silva. *Invitation to Ergodic Theory*, volume 42 of *Student Mathematical Library*. American Mathematical Society, Providence, RI, 2008.

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

Jennifer Skeem, John Monahan, and Christopher Lowenkamp. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40(5): 580, 2016.

Jennifer L. Skeem and Christopher T. Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

Rhys Steele. 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).

Julia Stoyanovich and Bill Howe. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3), 2019.

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

Watson v. Fort Worth. Watson v. Forth Worth Bank & Trust, 1988. 487 U.S. 977.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, 2023.

Teresa Scotton Williams. Some issues in the standardized testing of minority students. *Journal of Education*, pages 192–208, 1983.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

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

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 228–238, 2017b.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017c.

Michael Zanger-Tishler, Julian Nyarko, and Sharad Goel. Risk scores, label bias, and everything but the kitchen sink. *arXiv preprint arXiv:2305.12638*, 2023.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

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

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3929–3935, 2017.