Algorithmic Design: Fairness Versus Accuracy*

Annie Liang† Jay Lu‡ Xiaosheng Mu§

November 8, 2021

Abstract

Algorithms are increasingly used to guide decisions in high-stakes contexts, such as who should receive bail, be approved for a loan, or be hired for a job. Motivated by a growing body of empirical evidence, regulators are concerned about the possibility that the error rates of these algorithms differ sharply across subgroups of the population. What are the tradeoffs between accuracy and fairness faced by algorithms, and how do they depend on the underlying statistical properties of the algorithm's inputs? To answer these questions, we propose a model in which a designer chooses an algorithm that maps observed inputs into decisions. We characterize the accuracy-fairness Pareto frontier (corresponding to different preferences for how to trade off accuracy and fairness), and identify simple statistical properties of the algorithm's inputs that determine the shape of this frontier. These general results allow us to derive characterizations in special cases where the inputs satisfy additional structure, for example when one of the inputs is the individual's group identity. We conclude by applying our results to address a recent policy question regarding whether and when to ban use of certain kinds of inputs by predictive algorithms.

1 Introduction

In 2016, an algorithm used to guide decisions about who should receive bail was revealed to have a false positive rate (i.e., incorrectly assessing a criminal defendant as high-risk of

*We thank Nageeb Ali, Sergiu Hart, and Sendhil Mullainathan for helpful comments.

†Northwestern University

‡UCLA

§Princeton University

future criminal offense) that was twice as high for non-white defendants as for white defendants (Angwin and Larson, 2016). As algorithms are increasingly used to guide decisions in important contexts—such as who should receive bail, be approved for a loan, or receive a medical treatment—policymakers have become concerned with the possibility that algorithms create "bias" in the sense that their errors are systematically borne by members of a certain group. This paper does not attempt to resolve the difficult question of how to handle these tradeoffs between fairness and accuracy. Rather, we seek to understand when these tradeoffs emerge, and how the nature of these tradeoffs depends on underlying statistical properties of the data.

We propose a framework in which an algorithm assigns actions (e.g., whether or not to recommend bail) to individuals in a population, where the algorithm's decisions are based on observed covariates about the individual (e.g., past criminal background, psychological evaluations, social network data). We use a general loss function to evaluate the algorithm's error for a given individual, where the error depends on the action chosen and the individual's unobserved type (e.g., crime reoffense or recidivism). We then aggregate errors within members of pre-defined groups (e.g., race), where the *group error* is the expected error for members of that group. One pair of group errors *Pareto dominates* another if the former involves lower errors for both groups (greater accuracy) and also a lower difference between group errors (greater fairness). We do not privilege a specific way of trading off fairness and accuracy, and the Pareto frontier reflects those points that are optimal across a broad range of preferences for how to manage this tradeoff.

We consider two problems for the designer. In a first problem, which we call *full design*, the designer can flexibly choose from the set of all algorithms that map the observed covariates into actions. We propose also a *Bayes design* problem in which the designer is constrained to algorithms that choose the optimal action given some garbling of the available information. One interpretation of the Bayes-design problem is that it describes a setting where there may be a conflict between the designer who controls information (e.g. a regulator setting "ban-the-box" style policies) and a decision-maker who sets the algorithm (e.g. a manager setting the hiring policy). The decision-maker seeks to maximize accuracy, while the designer may have other fairness concerns. The set of feasible outcomes under Bayes-design are exactly those that the designer can implement by choosing different garblings of the available covariates to make available to the decision-maker.

Our first result characterizes the Pareto frontier for the full design problem. The shape of this frontier turns out to depend on a simple statistical property of the covariates, which is whether the algorithm that minimizes each group's error also leads that group to have a lower error than the other group. If so, we say that the covariates are *group-balanced*.

In practice, covariates can fail group balance if they are systematically more informative about one group than another (e.g., because of a larger quantity of historical data about one group over the other). In this case, we say the covariates are *group-skewed*. We show that if and only if covariates are group-skewed, then a strong kind of fairness-accuracy conflict is inevitable: Movement along one part of the Pareto frontier involves increasing *both* groups' errors (a strong form of reduction in accuracy). In contrast, if the covariates are group-balanced, then uniformly increasing both groups' errors necessarily moves off the Pareto frontier, and thus cannot be optimal regardless of the designer's fairness-accuracy preferences. These results demonstrate that certain philosophical disagreements—e.g., is it justifiable to decrease accuracy for everyone, if it involves an increase in fairness across groups?—need not be practically relevant, depending on the statistical properties of the available covariates.

Next, we characterize the Pareto frontier for the Bayes design problem. We show that the only constraint implied by Bayes design is that the algorithm must improve the aggregate error rate (averaged over groups) relative to no information. Thus the Bayes design Pareto frontier is simply that part of the full design Pareto frontier that belongs to the halfspace of error pairs that improve upon the prior. This result provides a second, complementary, perspective on the Pareto frontier already derived: every point on the full design Pareto frontier that satisfies the weak condition of improving upon the prior is implementable by restricting the data available to the algorithm. Importantly, this implies that although the designer may have no control over the algorithm set by the decision-maker, he can still (in many cases) obtain his most preferred outcome with the use of informational constraints.

These general characterizations also allow us to derive more specific results for covariates satisfying certain statistical properties. We consider three settings. In the first, we consider covariates that perfectly reveal group identity. This includes the case where group identity is a direct input into the algorithm, as well as the case where observed covariates perfectly proxy for group identity. We show that the Pareto frontier in this case is "Rawlsian": Everywhere along the Pareto frontier, the disadvantaged group (i.e., the group with the higher error) receives its minimal feasible error. Thus the only content of fairness-accuracy preferences is their implications for the advantaged group.

Next, we consider the setting where there is a strong form of independence between the group identity and the other unknowns (covariates and type). We show that under this strong independence condition, the Pareto frontier consists of a single point, so that fairness-accuracy preferences become irrelevant for determination of optimal policy.

The final setting we consider is a generalization of the previous two where the measured covariates capture all of the information in group identity that is relevant towards predicting

the type. Formally, we suppose that group identity and type are independent conditional on the observed covariates. We show that in this case the Pareto frontier consists only of strong fairness-accuracy conflicts, i.e. the only way to increase fairness along the frontier is to decrease accuracy across all groups.

Finally, we conclude by applying our results to address the recent policy question of whether and when to ban use of certain covariates. Formally, we study how the Pareto frontier changes when the designer gains access to a new covariate. We provide conditions under which access to the covariate leads the Pareto frontier to become everywhere better, implying that the covariate leads to a strictly better outcome regardless of the policymaker's preference. The conditions for this uniform improvement turn out to be quite weak. For example, adding the group label as a covariate leads to a uniform improvement in the Pareto frontier whenever the other covariates are group-balanced. These results clarify that banning a covariate typically cannot be justified by fairness or accuracy concerns (as defined in our framework), although it is generally the case that the designer prefers to condition on a noisy transformation of the available information, rather than use the information directly as given. Moreover they demonstrate that policies that are unfair from the perspective of disparate *treatment*—such as including group identity as a covariate—may improve fairness in the sense of disparate *impact* (i.e., whether the adverse effects of the algorithm are disproportionately borne by members in a specific group).

1.1 Related Literature

Our work builds on a recent literature in computer science on algorithmic fairness (see Kleinberg et al. (2018) and Roth and Kearns (2019) for overviews). Kleinberg et al. (2017) and Chouldechova (2017) demonstrated that certain notions of fairness (equal false positive rates, equal false negative rates, calibration) cannot be simultaneously satisfied. This important early work pointed not only to the necessity of tradeoffs between fairness and traditional goals such as accuracy, but also to potentially different definitions of fairness. A large subsequent literature has explored alternative notions of fairness—for example, fairness defined over individuals rather than groups (Dwork et al., 2012; Kearns et al., 2019), fairness that takes into account the endogenous decisions of agents (Jung et al., 2020), and fairness for when the algorithm does not directly output a decision, but instead guides a human decision-maker (Rambachan et al., 2021; Gillis et al., 2021). Concurrently, a separate branch of the literature has focused on developing novel algorithms that optimize for a more traditional goal (e..g, efficiency or profit) subject to a constraint on fairness (Hardt et al., 2016; Diana et al., 2021).

Our work differs from the previous literature in the following important ways. First,

rather than developing an optimal algorithm subject to a hard fairness constraint (e.g., requiring approximately equal group error rates), we solve for the Pareto frontier between fairness and accuracy. Several authors have pointed to such a frontier as a useful conceptual tool (Roth and Kearns, 2019), and others have estimated this frontier for specific data sets (Wei and Niethammer, 2020). Our work provides theoretical results for how this frontier will look depending on statistical properties of the algorithm's inputs.

Second, we use a general definition of group error, which nests several of the popular fairness metrics in the literature, but can also be interpreted more broadly as (negative) group utility.[1] This more general formulation facilitates comparison between our framework and the literature in philosophy and economics, which considers the question of how to choose between different distributions of outcomes (broadly construed) across individuals within a society. Several classical perspectives have natural analogues in our problem. The familiar utilitarian perspective (Harsanyi, 1953, 1955) translates in our framework to a preference that minimizes the algorithm's average error across all individuals, without regard for how the algorithm's errors may differ across groups. At the other extreme, a pure egalitarian or luck egalitarian[2] seeks to eliminate inequality across groups (Parfit, 2002; Knight, 2013).[3] Still other approaches are intermediate: For example, the Rawlsian approach maximizes the payoff for the most disadvantaged individuals, and Grant et al. (2010) characterize a generalization of utilitarianism that allows for non-linear aggregation of individual payoffs in order to capture fairness considerations. Our Pareto frontier accommodates these various perspectives, some of which we make explicit in Section 2.1.

Third, in our Bayes design problem, we do not take the input to the algorithm as given but instead search over possible covariates from a large space of noisy transformations of the available covariates. Here, our (Bayes-design) approach follows the information design approach (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019) with the following differences: (1) we consider the Pareto frontier with respect to a large class of "Sender" preferences; (2) our Sender only has access to a restricted set of information structures that are garblings of the available covariates; (3) our focus on fairness considerations introduces non-linearities that complicate the Sender's objective function.[4] We note that certain cri-

[1] See Corbett-Davies and Goel (2018) for a critical review of several of the popular error rate metrics.

[2] Luck egalitarians ask that people are made equal "in the benefits and burdens that accrue to them via brute luck" (namely, luck that falls on a person in ways beyond their control), but allows for inequities that result from intentional choices. Most of the group identities that are relevant in our motivating applications (see Section 2.1) are not chosen by individuals.

[3] Derek Parfit's "Principle of Equality" asserts that "it is bad in of itself if some people are worse off than others."

[4] In particular, the Sender's objective function is not posterior-separable and cannot be expressed as a straightforward expectation of payoffs conditional on realized posteriors.

tiques of the information design approach—especially commitment to a flexible information policy—are less relevant in our setting. We view the noisy transformations in our model as policy instruments that the policymaker can potentially commit to by law. (See for example Yang and Dobbie (2020), which summarizes the extant law and proposes new legal policies for mitigating algorithmic bias.)

The theoretical literature on algorithmic fairness is informed by a growing body of empirical work, which points to the existence of algorithmic bias in multiple settings. For example, Obermeyer et al. (2019) find that a widely used algorithm for predicting need of medical treatment has the property that at every risk score, non-white patients are considerably sicker than white patients. Arnold et al. (2021) find that a risk assessment tool used to guide pretrial bail decisions in NYC recommends that white defendants be released before trial at a higher rate than non-white defendants with equal risk of pretrial misconduct. Fuster et al. (2021) argue that use of sophisticated machine learning techniques to predict mortgage default leads to greater improvements for white borrowers than for non-white borrowers.

Finally, our framework shares certain features with the literature on statistical discrimination (see Fang and Moro (2011) for a survey). For example, the point that observable characteristics may be correlated with unobserved characteristics of interest (such as ability) is one that has been well-noted in this literature. Models of statistical discrimination have primarily focused on explaining why inequality emerges and persists in equilibrium, while our paper focuses instead on practical questions regarding algorithmic design.

2 Framework

2.1 Setup and Notation

Consider a population of subjects, each described by a *type* Y taking values in the finite set $\mathcal{Y}$, a *group identity* G taking values r or b, and a *covariate* vector X taking values in the finite set $\mathcal{X}$. Throughout we think of G, X, Y as random variables with joint distribution $\mathbb{P}$. For each group $g \in \{r, b\}$, we let $p_g \equiv \mathbb{P}(G = g)$ denote the fraction of the overall population that belongs to group g and we suppose that $p_g > 0$ for each group.

A designer chooses an action in $\mathcal{A} = \{0, 1\}$ for each individual, using an *algorithm* $f : \mathcal{X} \to \Delta(\mathcal{A})$ that maps observed covariates into distributions over actions. The variables Y and G are not directly observed by the designer and so cannot be used as inputs into the algorithm, but may be correlated with X. (Section 4.1 considers the special case where X reveals G.)

The error of choosing action a for a subject whose true type is y is measured using a loss

function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$. We further aggregate these losses across individuals within each group:

Definition 1. For any algorithm f and group $g \in \{r, b\}$, the *group g error* is

$$e_g(f) := \mathbb{E}\left[\ell(f(X), Y) \mid G = g\right].$$

That is, the group g error is the average loss for a member of group g. Throughout, we assume the definition of the relevant groups to be a primitive of the setting, determined by sociopolitical precedent and outside the scope of our model.[5]

Since we impose no restrictions on the loss function ℓ (in particular, we do not require that its range is positive), it is possible to alternatively set $\ell(a, y) = -u(a, y)$ where u is the payoff received by a subject with type y and action a. Under this interpretation, lower values of $e_g(f)$ correspond to higher average payoffs for subjects in group g. All of our subsequent results hold regardless of whether we interpret $e_g(f)$ as a group error or as (the negative of) a group payoff.

Some motivating examples of types Y, group identities G, covariates X, and actions a are given below:

Healthcare. Y is need of treatment, G is socioeconomic class (low SES or high SES), and the action is whether the individual receives treatment. The covariate vector X includes possible attributes such as image scans, number of past hospital visits, family history of illness, and blood tests.

Credit scoring. Y is creditworthiness, G is gender, and the action is whether the borrower's loan request is approved. The covariate vector X includes possible attributes such as purchase histories, social network data, income level, and past defaults.[6]

Criminal sentencing. Y is whether an individual is high-risk or low-risk of crime reoffense, G is race (white or non-white), and the action is whether the individual is released on bail. The covariate vector X includes possible attributes such as the individual's past criminal record, psychological evaluations, family criminal background, number of friends who are

[5] In general, the group fairness achieved by an algorithm will very much depend on how those groups are defined. Kearns et al. (2018) extend statistical notions of fairness to apply to various subgroups identified by an adversarial evaluator.

[6] The Apple Card was accused of gender discrimination when users noticed in certain cases that smaller lines of credit were offered to wives than to ther husbands. Apple was subsequently cleared of these charges on the basis that the algorithms used to set these credit limits "did not consider prohibited characteristics of applicants and would not produce disparate impacts." See `https://www.theverge.com/2021/3/23/22347127/goldman-sachs-apple-card-no-gender-discrimination`.

gang members, frequency of moves, or drug use as a child.[7]

Job hiring. Y is whether a job applicant is high or low quality, G is citizenship (immigrant or domestic applicants), and the action is whether the applicant is hired. The covariate X includes possible attributes such as past work history, resume, and references.

A natural choice for the loss function in the above applications is

$$\ell(a, y) = \begin{cases} 0 & \text{if } a = y \\ \lambda^+ & \text{if } (a, y) = (1, 0) \\ \lambda^- & \text{if } (a, y) = (0, 1) \end{cases} \tag{1}$$

where the parameters $\lambda^+, \lambda^- > 0$ respectively denote the error to a false positive (e.g., denying bail to a low-risk individual) and a false negative (e.g., granting bail to a high-risk individual).[8]

2.2 Fairness-Accuracy Pareto Frontier

We suppose that the designer cares about both accuracy and fairness, in the sense that he prefers lower group errors but also prefers errors to differ less across groups. We do not privilege a specific way of trading off between these two objectives and view this partial order as allowing for the largest set of possible designer preferences (see Appendix D for more detail).

Definition 2. Say that a pair of group errors (e_r, e_b) *Pareto-dominates* another pair (e'_r, e'_b) if $e_r \leq e'_r$, $e_b \leq e'_b$, and $|e_r - e_b| \leq |e'_r - e'_b|$, with at least one of these inequalities strict.

Below we provide a few prominent examples of different designer preferences that are consistent with this partial order, i.e. whenever (e_r, e_b) Pareto-dominates (e'_r, e'_b), then the designer strictly prefers (e_r, e_b) to (e'_r, e'_b).

Example 1 (Utilitarian). The designer evaluates errors $e = (e_r, e_b)$ according to the weighted sum in the population. That is, let

$$w_u(e) = p_r e_r + p_b e_b$$

and let $\succeq_u$ be the ordering represented by w_u, i.e. $e \succeq_u e'$ if and only if $w_u(e) \geq w^u(e')$. (Note that the minority population, which has a lower weight by definition, will be naturally discounted as a group in this evaluation.) We say that a designer is *Utilitarian* if his preference over error pairs is $\succeq_u$.

[7]These example covariates are based on the survey used by the Northpointe COMPAS risk tool. See for reference: `https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html`.

[8]If we instead interpret ℓ as negative utility, then loss functions different from the one in (1) may be more natural. For example, a suspect may prefer to be released on bail regardless of risk of crime reoffense.

Example 2 (Rawlsian). The designer evaluates errors $e = (e_r, e_b)$ according to the greater error. That is, let

$$w_r(e) = \max\{e_r, e_b\}.$$

and let $\succeq_r$ be the corresponding ordering represented by w_r. We say that a designer is *Rawlsian* if his preference over error pairs is $\succeq_r$.

Example 3 (Egalitarian). The designer evaluates errors $e = (e_r, e_b)$ according to their difference. That is, let

$$w_e(e) = |e_r - e_b|$$

and let $\succeq_e$ be the lexicographic order that first evaluates errors according to w_e and the compares ties using the Utilitarian utility w_u. We say a designer is *Egalitarian* if his preference over error pairs is $\succeq_e$.

Utilitarian, Rawlsian, and Egalitarian are often defined with respect to *individual utilities* rather than *group errors*. The distinction between utilities and errors is purely interpretational, since what we call group error can equally be interpreted as a negative group utility. The difference between "individual" and "group" is more substantive,[9] and our choice of defining fairness concerns over social groups follows the motivation of the paper.[10]

We consider two design problems, which reflect different constraints on the designer. In the first problem, which we call *full design*, the designer can flexibly choose from the set $\mathscr{F}$ of all mappings $f : X \to \Delta(A)$, so the set of feasible group error pairs given covariate X is

$$\mathcal{E}(X) \equiv \{(e_r(f), e_b(f)) : f \in \mathscr{F}\}.$$

We propose also a second *Bayes design* problem in which the designer is constrained to algorithms that choose the optimal action given some garbling of the available information. Recall that any garbling T of X can be associated with a stochastic map $T : \mathcal{X} \to \Delta(\mathcal{T})$ that takes realizations of X into distributions over the possible realizations of T. Fixing a garbling T, let $f_T : \mathcal{T} \to \Delta(A)$ denote any mapping that assigns each realization of T to a optimal action. That is, each $f_T(t)$ is supported on the set of actions a minimizing $\mathbb{E}_P[\ell(a, Y) \mid T = t]$.

[9] The exception is the case of the Utilitarian designer, where it is not relevant.

[10] Note that the alternative route, in which a person is the unit over which fairness is imposed, requires resolution of what distinguishes a unique person. In our framework, individuals cannot be distinguished except through their measured covariates, so the "most disadvantaged person" corresponds to the most disadvantaged realization of the measured covariates. But this most disadvantaged entity is dependent, then, on which set of covariates we measure, a strange and undesirable endogeneity. In contrast, our approach of defining the group as the unit of person (with G pre-defined) avoids this issue.

Definition 3. Say that a pair of group errors (e_r, e_b) is *implemented by* T if $(e_r, e_b) = (e_r(f_T), e_b(f_T))$.

The feasible set of group errors under Bayes-design given covariate X is then

$$\mathcal{E}^*(X) \equiv \{(e_r, e_b) \,:\, (e_r, e_b) \text{ is implemented by a garbling } T \text{ of } X\}.$$

One interpretation of the Bayes-design problem is that it describes the interaction between a designer who controls information and a decision-maker who sets the algorithm. The designer first determines what data about subjects can be legally used as inputs into the algorithm, and the decision-maker then uses the permitted inputs to choose actions. The decision-maker cares only to maximize accuracy, while the designer may have other fairness concerns. For example, a judge determining sentencing may seek to maximize the number of correct verdicts, while a policymaker may additionally prefer that the accuracy of the judge's verdicts is equitable across certain social groups. The set of feasible outcomes under Bayes-design are exactly those that the designer can implement by choosing different garblings of X to make available to the decision-maker.

A second interpretation of Bayes design is that it models the training of a machine learning algorithm, where the garbling does not reflect an intentional policy strategy but rather noise in the data. Under this interpretation, the covariate vector X represents the variables that we are trying to measure (e.g., family criminal background), while its garbling T describes the measurements that are available as inputs into the algorithm (e.g., number of family members who have been arrested). The feasible set under Bayes design reflects the different outcomes that are implied by different noisy measurements of X.[11]

Definition 4. The full-design Pareto set given X, denoted $\mathcal{P}(X)$, is the set of all pairs $(e_r, e_b) \in \mathcal{E}(X)$ that are *Pareto-undominated*, i.e. no other error pair $(e'_r, e'_b) \in \mathcal{E}(X)$ Pareto-dominates it. The Bayes-design Pareto set given X, denoted $\mathcal{P}^*(X)$, is the set of all pairs $(e_r, e_b) \in \mathcal{E}^*(X)$ that are Pareto-undominated in $\mathcal{E}^*(X)$.

3 Characterization of the Pareto Frontier

Our results in this section describe the fairness-accuracy Pareto frontier and how it depends on properties of the covariate X. We characterize the full design Pareto frontier in Section 3.1 and the Bayes design Pareto frontier in Section 3.2. In the subsequent Section 4, we use

[11] Noise in these measurements may very well be correlated with G. For example, suppose that members of group r are arrested at random (independent of criminality), while only individuals in group b that have committed a crime are arrested. Then, the number of prior family arrests is a noisier measurement of family criminal background for group r than group b.

these general characterizations to derive characterizations of the Pareto frontier in special cases where X possesses additional structure.

3.1 Full Design

The full design Pareto frontier $\mathcal{P}(X)$ corresponds to a part of the boundary of the feasible set $\mathcal{E}(X)$, where a special role is played by those feasible error pairs that minimize each group's error (individually), and the error pair that minimizes the difference between group errors. Since the feasible set $\mathcal{E}(X)$ is closed and convex (see Lemma A.1), the following optima are well-defined.

Definition 5 (Group Optimal Points). For any covariate X, define

$$R_X \equiv \underset{(e_r,e_b)\in\mathcal{E}(X)}{\arg\min}\ e_r$$

to be the feasible point that minimizes group r's error, and define

$$B_X \equiv \underset{(e_r,e_b)\in\mathcal{E}(X)}{\arg\min}\ e_b$$

to be the feasible point that minimizes group b's error. In both cases, if the minimizer is not unique, we break ties by choosing the point that minimizes the other group's error. We let G_X denote the group optimal point for group g.

Group optimal points can be easily derived from data. For instance, to calculate R_X, set the algorithm to choose the optimal action for group r for each realization of X (breaking ties in favor of group b).[12] R_X is then the error pair resulting from this algorithm.

Definition 6 (Fairness Optimal Point). For any covariate X, define

$$F_X \equiv \underset{(e_r,e_b)\in\mathcal{E}(X)}{\arg\min}\ |e_r - e_b|$$

to be the point that minimizes the absolute difference between group errors. If the minimizer is not unique, we choose the point that further minimizes either group's error.[13]

While R_X and B_X respectively denote the points that minimize group r and b's errors, the group whose error is minimized need not be the group with the lower error. For example,

[12]Throughout, when we say "optimal action for group g at realization x," we mean any $a \in \arg\min_{a'\in\mathcal{A}} \mathbb{E}[\ell(a', Y) \mid X = x, G = g]$.

[13]It can be shown that for every X, this point is the same regardless of which group is used to break the tie.

suppose $\mathbb{P}(Y = 1 \mid G = r) = \mathbb{P}(Y = 1 \mid G = b) = 1/2$, and X is a binary score with the following conditional probabilities:

	$X = 0$	$X = 1$
$Y = 0$	3/4	1/4
$Y = 1$	1/4	3/4

$G = r$

	$X = 0$	$X = 1$
$Y = 0$	2/3	1/3
$Y = 1$	1/3	2/3

$G = b$

Let the loss function ℓ satisfy $\ell(a, y) = 1$ if $a \neq y$, and otherwise $\ell(a, y) = 0$. Then the b-optimal point B_X is achieved by the algorithm that maps $X = 1$ to $a = 1$ and $X = 0$ to $a = 0$, which leads to a *higher* error of 1/3 for group b, compared to the error of 1/4 for group r. Thus, using X to maximally reduce errors for group b results in an even greater reduction in error for group r. The subsequent definition formalizes this property.

Definition 7. Covariate X is:

- *r-skewed* if $e_r < e_b$ at R_X and $e_r \leq e_b$ at B_X
- *b-skewed* if $e_b < e_r$ at B_X and $e_b \leq e_r$ at R_X
- *group-balanced* otherwise

Group balance is easily evaluated on data by testing whether the algorithm that chooses the optimal action for group g indeed leads to a lower error for that group. X is group-balanced when this is true for both groups. On the other hand, if X is g-skewed for some group g, then we say it is *group-skewed*.

We now characterize the Pareto frontier in each of these cases. Given two points on the boundary of a set, the *lower boundary* is the part of the boundary of the set between the two points, and below the line segment connecting the two.

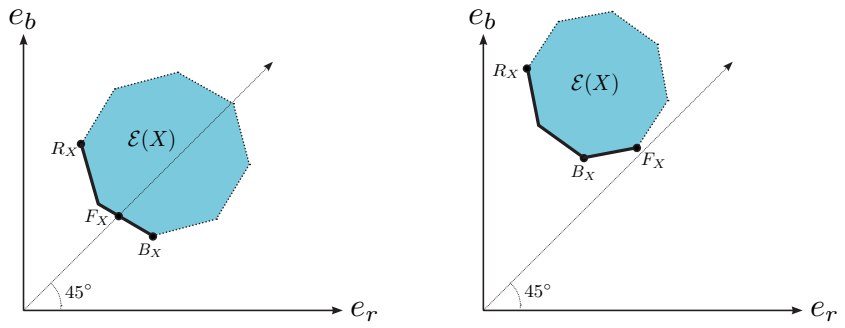
Theorem 1. *The full-design Pareto set $\mathcal{P}(X)$ is the lower boundary of the full-design feasible set $\mathcal{E}(X)$ between*

(a) R_X and B_X if X is group-balanced

(b) G_X and F_X if X is g-skewed

The cases described in Theorem 1 are depicted in Figure 1. When X is group-balanced and R_X and B_X are distinct, then they fall on opposite sides of the 45-degree line, and the Pareto frontier is that part of the lower boundary of the feasible set connecting these two points. When X is g-skewed, then both R_X and B_X fall to the same side of the 45-degree

line, and the Pareto frontier is that part of the lower boundary of the feasible set connecting G_X to F_X. In general, this Pareto frontier need not include a point where the group errors are identical, although F_X is guaranteed to belong to the 45-degree line when the loss function takes the form given in (1).[14]



(a) X is group-balanced

(b) X is r-skewed

Figure 1: Depiction of an example full-design Pareto frontier for (a) a group-balanced covariate vector X and (b) an r-skewed covariate vector X. The case of a b skewed covariate vector X is the mirror image of (b) reflected across the 45° line.

Theorem 1 immediately implies the following corollary, which establishes an equivalence between group skewness and the existence of a particularly strong kind of fairness-accuracy conflict along the Pareto frontier.

Definition 8. Say that (e_r, e_b) and (e'_r, e'_b) represent a *strong fairness-accuracy conflict* if $e_r \leq e'_r$ and $e_b \leq e'_b$ but $|e_r - e_b| > |e'_r - e'_b|$. Say that X *implies a strong fairness-accuracy conflict* if there are two points in $\mathcal{P}(X)$ with this property.

A strong fairness-accuracy conflict means that the tradeoff between fairness and accuracy is especially stark: a designer's optimal point may involve higher errors for *both* groups. Both the Utilitarian and Rawlsian designer consider uniform increases across group errors to be welfare-reducing, but a designer who places sufficient weight on fairness (e.g., the Egalitarian designer) might prefer to increase both groups' errors if it reduces the difference between those errors. Our next corollary says that disagreements of this kind emerge only when X fails group balance.

Corollary 1. *Suppose F_X is distinct from R_X and B_X. Then X implies a strong fairness-accuracy conflict if and only if it is group-skewed.*

[14]Consider the algorithm that assigns action 1 with probability $\frac{\lambda^+}{\lambda^+ + \lambda^-}$ regardless of the covariate. This yields an error of $\frac{\lambda^+\lambda^-}{\lambda^+ + \lambda^-}$ for both groups.

This corollary is evident from Figure 1. When X is group-balanced (Panel (a)), the Pareto frontier consists exclusively of negatively-sloped line segments, so moving along the frontier necessarily lowers one group's error while raising another's. In contrast, when X is r-skewed (Panel (b)), then that part of the frontier connecting B_X to F_X has a positive slope. Moving along this part of the frontier thus increases errors for both groups, but decreases the difference between these errors. A similar observation holds in the case where X is b-skewed.

In practice, the kind of covariates that are likely to fail group balance (and hence, create strong fairness-accuracy conflicts) are those that are systematically more informative about one group than another. For example, because individuals belonging to a lower socioeconomic class are less likely to go to the hospital in case of sickness, the number of past hospital visits is more informative about need-of-care for wealthier individuals than for less wealthy individuals (Obermeyer et al., 2019). Conditioning on this covariate reduces errors for both groups but reduces errors for wealthy individuals by more. The only way to increase fairness is to condition less on this information, resulting in higher errors for both groups.

If we interpret $\mathbb{P}$ as a prior informed by historical data, then a similar asymmetry can emerge when there is less historical data on the relationship between observed covariates X and type Y for minority groups. For example, if medical data is drawn from experiments that predominantly involved male subjects, then beliefs about need-for-treatment for women may be less accurate than for men at every symptom profile. Again, this would mean that a way to increase fairness is to condition less on the available information, which reduces accuracy for both groups but decreases the gap in errors.[15] Whether this change is an improvement depends on the designer's fairness-accuracy preference.

3.2 Bayes Design

We turn now to Bayes design, where the designer is constrained to algorithms that map the available information (a garbling of X) to Bayes optimal decisions.

In this case, one constraint on the feasible group errors is that the aggregate error (averaging across groups) must improve upon the prior. Formally, let a_0 denote the optimal action given no information, and define $e_0 = \mathbb{E}(\ell(a_0, Y))$ to be the expected error when assigning a_0 to all individuals. Any errors achieved under Bayes design must then belong to

[15] Finally, we note that it is important in these examples that the meaning of X is the same across groups—e.g., the same symptoms indicate need for treatment—which has the consequence that the designer cannot disentangle the needs of the two groups. When different realizations of X imply different optimal actions for the two groups, then X can be group-balanced even if it is substantially more informative about one group than the other.

the halfspace

$$H = \{(e_r, e_b) \,:\, p_r e_r + p_b e_b \leq e_0\}.$$

Since the Bayes-design feasible set $\mathcal{E}^*(X)$ is also a subset of the full-design feasible set $\mathcal{E}(X)$, it follows that $\mathcal{E}^*(X) \subseteq \mathcal{E}(X) \cap H$.

Theorem 2 says that this is the only constraint on the Bayes-design feasible set: Every point in $\mathcal{E}(X) \cap H$ can be implemented by some garbling of X. Thus, the Bayes-design feasible set is exactly the intersection of the full-design feasible set and the halfspace H, and the Bayes-design Pareto set must also be that part of the full-design Pareto set that belongs to H.

Theorem 2. *For every covariate X, the Bayes-design feasible set is $\mathcal{E}^*(X) = \mathcal{E}(X) \cap H$ and the Bayes-design Pareto set is $\mathcal{P}^*(X) = \mathcal{P}(X) \cap H$.*

Figure 2 depicts these relationships. An important implication of our characterization of the Bayes frontier is the following. Recall that we can interpret Bayes design as a description of the conflict between a regulator who can control information (e.g. a regulator setting "ban-the-box" style policies) and a decision-maker who sets the algorithm to maximize accuracy (e.g. a firm setting the hiring policy). Theorems 1 and 2 imply the following corollary, which characterizes when such a conflict is irrelevant from the perspective of the regulator.

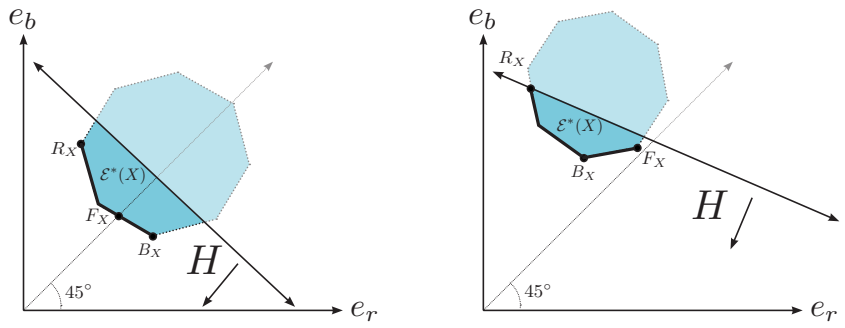


Figure 2: Depiction of an example Bayes-design Pareto frontier for (a) a group-balanced covariate vector X and (b) an r-skewed covariate vector X.

Corollary 2. *The following hold:*

(a) If X is g-skewed, then $\mathcal{P}(X) = \mathcal{P}^*(X)$ if and only if $G_X, F_X \in H$.

(b) If X is group-balanced, then $\mathcal{P}(X) = \mathcal{P}^*(X)$ if and only if $R_X, B_X \in H$.

When these conditions hold, whatever outcome that is optimal for the designer under full design can also be achieved under Bayes design. In other words, although the designer does not have explicit control over the algorithm set by the decision-maker, he can employ informational constraints (i.e. garblings) to generate his most preferred outcome. Conversely, when these conditions do not hold, then misaligned incentives between the designer and the decision-maker do matter; the designer may be unable to achieve his most preferred outcome even with full use of informational constraints.

4 Special Cases

We now use the general characterization demonstrated in the preceding section to derive more specific results for cases where X satisfies additional structure. We consider three settings: (a) when group identity G is completely revealed by the observed covariate X, (b) when group identity G is completely independent of (X, Y), and (c) when group identity G is independent of Y conditional on X. In each of these cases, we comment in particular on the implications for Utilitarian, Rawlsian, and Egalitarian designers (as defined in Section 2.1). For brevity, we state the results below for the case of the full design frontier, but it is easy to show (by applying Theorem 2) that every result stated in this section holds also for the Bayes design frontier.

4.1 X Reveals G: The Frontier is Rawlsian

First suppose that X reveals G, so that each subject's group identity is known to the algorithm. In practice, this would arise either if the covariate vector X includes the group identity as one of the covariates, or if certain covariates in X perfectly proxy for the unreported group identity.

Assumption 1 (X Reveals G)**.** *The conditional distribution $G \mid X = x$ is degenerate for every realization x of X.*

Proposition 1. *Suppose Assumption 1 holds. Then the feasible set $\mathcal{E}(X)$ is a rectangle whose sides are parallel to the axes, and $\mathcal{P}(X)$ is the line segment from $R_X = B_X$ to F_X.*

An example feasible set and Pareto frontier are depicted in Figure 3. One endpoint, $R_X = B_X$, gives both groups their minimal feasible error. The other endpoint, F_X, maximizes fairness. The special property that the Pareto frontier is parallel to the axis holds because it is possible to change one group's error without affecting the other group's error.

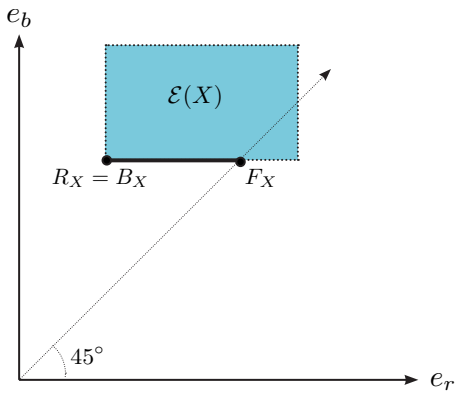


Figure 3: Depiction of the Pareto frontier in the case where X reveals G.

Different points along the Pareto frontier can be achieved in the following way. Under the assumption that X reveals G, one can partition all the realizations of X into those that reveal that the group is b (call these group-b realizations) and those that reveal that the group is r (call these group-r realizations). Group-b realizations should be sent directly to the optimal action (which is also the optimal action for group b). Group-r realizations should be sent to a random action, where the closer the desired point is to the fairness-optimal F_X, the greater the probability of choosing the suboptimal action for group r. The implementation procedure is similar under Bayes design: the designer can increases fairness along the frontier by only garbling the advantaged group's signals while leaving the disadvantaged group's signals intact.

It immediately follows from this result that every point on the Pareto frontier gives the disadvantaged group its minimal feasible error, and so:

Corollary 3. *Under Assumption 1, every point on the Pareto frontier is optimal for a Rawlsian designer.*

Across this continuum of Rawlsian-optimal points, the error for the advantaged group ranges from its minimal feasible error (the Utilitarian-optimal point R_X) to the error closest to that of the disadvantaged group (the Egalitarian-optimal point F_X). The content of fairness-accuracy preferences in the setting where X reveals G, therefore, lies only in its implications for the advantaged group.

In practice, our approach is related to restrictions on disclosure to improve fairness in different contexts. For instance, many jurisdictions have passed "ban the box" policies that restrict employers from asking about past criminal histories in order to reduce racial dispar-

ities in employment (Agan and Starr, 2018). In these cases, the noisy transformations are usually symmetric between the two groups, i.e. if disclosure of criminal history is banned at all, then it is banned equally for all racial groups. In contrast, our Bayes-design Pareto frontier uses garblings of X that implement noisy transformations that are asymmetric between the two groups. Such policies may be unfair from the perspective of disparate *treatment* (i.e., whether the policy discriminates between individuals on the basis of a group identity), but may be necessary to impose fairness in the sense of disparate *impact* (i.e., whether the adverse effects of the policy are disproportionately borne by members in a specific group).[16] Our analysis helps formalize the tension between these goals, and further demonstrates how to implement such policies in practice.

4.2 Strong Independence: Rawlsian = Utilitarian = Egalitarian

Next we suppose to the contrary that (X, Y) is completely uninformative about group identity. Formally, the relative proportion of subjects belonging to either group is independent of the realizations of X and Y.

Assumption 2 (Strong Independence)**.** *For both groups g,*

$$\mathbb{P}(G = g \mid Y = y, X = x) = p_g \quad \forall x, y.$$

This is a strong property. One pair of sufficient conditions for the assumption to hold is if $G \perp\!\!\!\perp Y$ (group identity and type are independent) and $G \perp\!\!\!\perp X \mid Y$ (group identity is conditionally independent of the measured covariate given type). Independence of group identity and type will fail in many of the applications described in Section 2.1, but perhaps not in all—for example, gender and risk of certain medical outcomes are thought to be independent.[17] Conditional independence of group identity and covariates is in the spirit of certain legal restrictions that forbid use of correlates of group identity for prediction (Yang and Dobbie, 2020), but our condition is formally different.[18] In contrast to the previous two settings, we view Strong Independence primarily as a useful theoretical benchmark and not as a description of covariate vectors that are likely to appear in practice.

The feasible set in this case turns out to be a line segment on the 45-degree line, and the Pareto set is a single point, as depicted in Figure 4.

[16] See `https://www.justice.gov/crt/book/file/1364106/download` for definitions of disparate treatment and impact.

[17] For example, Crabtree et al. (1999) find no significant difference in mortality based on gender in a large study of hospital infections.

[18] That is, neither $G \perp\!\!\!\perp X$ nor $G \perp\!\!\!\perp X \mid Y$ imply one another.

Proposition 2. *Suppose Assumption 2 holds. Then the Pareto frontier is a single point on the 45-degree line.*

The Pareto frontier consists of the single point that is achieved by conditioning on all of the available information in X. Since this point is on the 45-degree line, both groups have the same error. Thus, this point is simultaneously optimal for Rawlsian, Utilitarian, and Egalitarian designers—indeed, fairness-accuracy preferences are completely irrelevant here: All designers who agree on the basic Pareto dominance principle outlined in Definition 2 prefer the same policy.

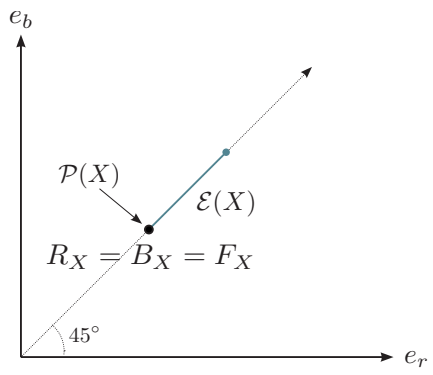


Figure 4: Depiction of the Pareto frontier under assumption of strong independence

4.3 Conditional Independence: Rawlsians and Utilitarians Agree

Finally we consider a condition that generalizes the previous conditions: We suppose that group identity G and type Y are independent conditional on the observed covariate X.

Assumption 3 (Conditional Independence)**.** $G \perp\!\!\!\perp Y \mid X$.

Under this assumption, X contains all of the information in the group identity that is relevant to predicting Y, so that once the algorithm has conditioned on X, there is no additional predictive value to knowing the group's identity. This kind of conditional independence appears for example when the coefficient on group identity is zero in a regression of Y on observables, e.g. Ludwig and Mullainathan (2021) find that race (G) is not predictive of a criminal's risk (Y) conditional on arrest (X) in their data.

Proposition 3. *Suppose Assumption 3 holds. Then every pair of distinct points* $(e_r, e_b), (e'_r, e'_b) \in \mathcal{P}(X)$ *represents a strong fairness-accuracy conflict.*

An example Pareto frontier for a covariate satisfying Conditional Independence is depicted in Figure 5. The left point is the (shared) group optimal point $R_X = B_X$, which is the preferred point for both a Rawlsian and Utilitarian designer. To show existence of such a point, we prove that under Assumption 3, the optimal action at each realization of X is the same for both groups. Thus, the algorithm that minimizes error for the disadvantaged group is the same as the algorithm that minimizes error for the advantaged group, and hence it is also the algorithm that minimizes aggregate error.

The right endpoint is the fairness optimal point F_X, and this is the preferred point for an Egalitarian designer. From $R_X = B_X$ to F_X, the Pareto frontier consists entirely of positively sloped line segments. Thus, everywhere along the frontier, the two groups' errors move in the same direction, implying that the only way to improve fairness is to decrease accuracy uniformly across groups, and that the only difference across designers that matters is how they choose to resolve strong fairness-accuracy conflicts.[19] Strengthening Conditional Independence to the assumption that X reveals G (Section 4.1) yields that the line from $R_X = B_X$ to F_X has slope 0, while Strong Independence (Section 4.2) yields that $R_X = B_X$ and F_X are the same point.

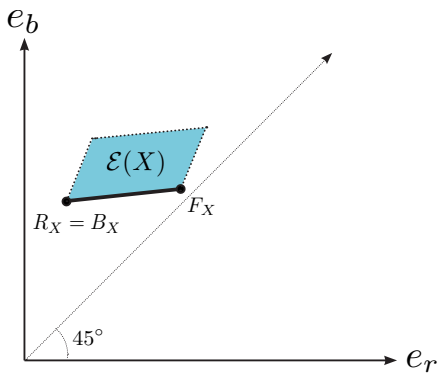


Figure 5: Depiction of the Pareto frontier under assumption of conditional independence of G and Y.

[19] In the special case when $R_X = B_X = F_X$, the Pareto set is just a singleton, and there is no strong fairness-accuracy conflict. (Proposition 3 is vacuous in this case, since there are no two distinct points on the Pareto frontier.)

5 When Should Covariates be Banned?

An important question facing lawmakers and regulators of AI is whether to prohibit use of certain covariates by predictive algorithms. We conclude by applying our main results to study how the Pareto frontier changes when a covariate is banned.

The typical justification for prohibiting a covariate is that use of the covariate increases inequity across two groups. Historically, protected group identities such as race and religion have been illegal inputs into consequential decisions such as lending and hiring.[20] But increasingly, covariates different from group identity are also prohibited, on the grounds that these covariates are biased against certain groups. For example, as of May, 2021, the University of California university system no longer considers test scores in their admissions decisions.[21]

To study the effect of banning a new covariate, we now rewrite the previous covariate vector as a pair (X, X'), where we interpret X as those covariates that have been approved for use, and X' as the covariate in question. We use $\mathcal{X}$ and $\mathcal{X}'$ to denote the sets of possible realizations for X and X'. For a given fairness-accuracy preference, the optimal use of a new covariate X' may be to discard it. In this section, we derive conditions under which access to a new covariate X' leads to a *uniform* welfare improvement beyond what is achievable with a background covariate vector X.[22] When X' uniformly Pareto improves upon X, then it is never optimal for a designer to discard X'.

Definition 9. Say that X' *uniformly Pareto-improves* upon X if every point in $\mathcal{P}(X)$ is Pareto-dominated by a point in $\mathcal{P}(X, X')$.

Sections 5.1 and Section 5.2 provide sufficient and necessary conditions for X' to uniformly Pareto improve upon X in three special cases. Section 5.1 considers the setting where the approved covariates X reveal the group identity. Section 5.2 considers the setting where the covariate in question X' *is* the group identity. Section 5.3 gives a sufficient condition for any X that is group-balanced. As in Section 4, we state our results only for the full design

[20]For example, the Equal Opportunity Act forbids any creditor to discriminate on the basis of "race, color, religion, national origin, sex or marital status, or age" (see `https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf`), and Title VII of the Civil Rights Act prohibits discrimination by employers on the basis of "race, color, religion, sex, or national origin" except in cases where the protected trait is an occupational qualification.

[21]See for reference: `https://www.nytimes.com/2021/05/15/us/SAT-scores-uc-university-of-california.html`.

[22]If Definition 9 is satisfied, then every point in the Bayes-design Pareto frontier $\mathcal{P}^*(X)$ is also Pareto-dominated by a point in $\mathcal{P}^*(X, X')$, so the uniform improvement holds whether we consider full design or Bayes design. In the Bayes-design setting, the implication is that the regulator can always achieve a better outcome by giving the decision-maker some garbling of X' as an input.

Pareto frontier. The conditions in Propositions 4 and 5 remain sufficient for a uniform Pareto improvement of the Bayes design Pareto frontier, but may not be necessary.[23] Proposition 6 (which only provides a sufficient condition) holds for Bayes design without modification.

All of the conditions we derive are quite weak, suggesting that access to many covariates that are relevant in practice represent a uniform Pareto improvement. None of these conditions rely on whether the covariate X' is "biased" in the sense of being systematically lower-valued or less informative for one group than another, although interestingly it can matter whether the baseline covariate vector X is systematically more informative about one group (via the property of g-skew).[24] Our result depends crucially on our assumption that the designer can either (1) choose a flexible (not necessarily optimal) algorithm based on the available information, or (2) choose a flexible garbling of the available information to condition on.

5.1 When X Reveals G

We begin by considering the case where the approved covariates reveal the group identity G. Whether a uniform Pareto improvement obtains in this setting turns out to reduce to the following condition:

Definition 10. Say that X' is *decision-relevant over X for group g* if there exist $x \in X$ and $x', \tilde{x}' \in \mathcal{X}'$ (where (x, x') and $(x, \tilde{x}')$ have strictly positive probability conditional on $G = g$) such that the optimal assignment for group g is uniquely equal to 1 at (x, x') and 0 at $(x, \tilde{x}')$.

This is a weak condition that says only that the additional information in X' sometimes matters for the optimal assignment for individuals in group g. For example, if X' is the outcome of a medical scan, then X' fails to be decision-relevant group for group g if and only if the (perceived) best treatment for *every* individual in group g is unchanged by the outcome of the scan.

Proposition 4. *Suppose X reveals G, and without loss let X be group-balanced or r-skewed. Choose any additional covariate X'.*

(a) If X is r-skewed, then X' uniformly Pareto-improves upon X if and only if X' is decision-relevant over X for group b.

(b) If X is group-balanced, then X' uniformly Pareto-improves upon X if and only if X' is decision-relevant over X for both groups.

[23] When the conditions in Corollary 2 hold for X, in which case $\mathcal{P}(X) = \mathcal{P}^*(X)$, then Propositions 4 and 5 extend fully for Bayes design.

[24] This finding is similar to a result in Rambachan et al. (2021), which shows (in a different model) that any new covariate, however biased, will be optimally used by a social planner, as long as it is informative.

We prove this result by demonstrating a lemma that says that access to X' reduces the minimal feasible error for group g if and only if X' is decision-relevant over X for group g. Applying Proposition 1, both the Pareto frontier given X and the Pareto frontier given (X, X') are single line segments (either horizontal or vertical). When X' fails to be decision-relevant over X for the disadvantaged group, then the new Pareto frontier must remain a line that overlaps with the previous frontier, so a uniform Pareto-improvement is not obtained. On the other hand, when X' is decision-relevant over X for the disadvantaged group, then the minimal feasible error for that group strictly reduces, pushing the Pareto frontier downwards. The different possible uniform Pareto improvements are depicted in Figure 6.

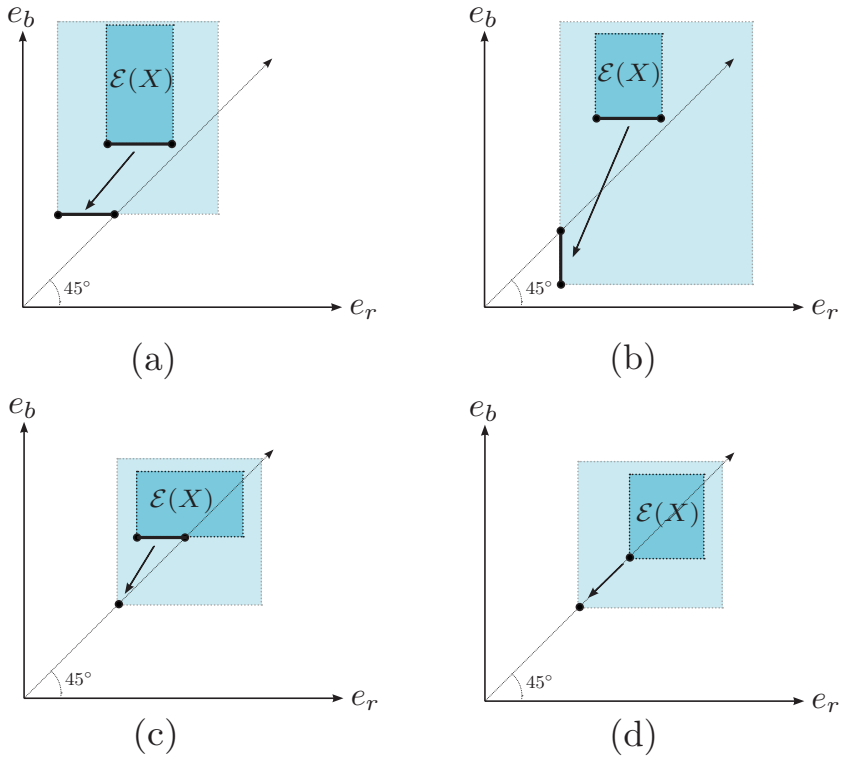


Figure 6: Depiction of the four kinds of uniform Pareto improvement when X reveals G: (a) Both X and (X, X') are r-skewed, (b) X is r-skewed while (X, X') is b-skewed, (c) X is r-skewed while (X, X') is group-balanced, (d) both X and (X, X') are group-balanced

5.2 When X' is G

Next we consider the setting where the covariate in question *is* the group identity G. Again the property of group balance emerges as the critical one.

Definition 11. Say that X is *strictly* group-balanced if $e_r < e_b$ at R_X and $e_b < e_r$ at B_X.

Relative to group-balance, strict group-balance rules out covariate vectors X for which $R_X = B_X = F_X$. Any group-balanced covariate vector X for which $\mathcal{P}(X)$ is not a singleton is strictly group-balanced.

Proposition 5. *G uniformly Pareto-improves upon X if and only if X is strictly group-balanced.*

The key observation towards this result is that G cannot be decision-relevant over X for either group, so the minimal feasible error for both groups is the same given X or given (X, G). Geometrically, this means that the Pareto frontier given (X, G) is contained within the smallest rectangle enclosing the Pareto frontier given X. When X is group-balanced, then $\mathcal{P}(X)$ is characterized by Part (c) of Theorem 1 while $\mathcal{P}(X, G)$ is characterized by Proposition 1. We show that the new Pareto frontier does not intersect with the original frontier, and that this implies that G uniformly Pareto-improves upon X. See Panel (a) of Figure 7 for an illustration. On the other hand, when X is r-skewed, then the Pareto frontier $\mathcal{P}(X, G)$ includes the point B_X, so a uniform Pareto improvement is not obtained. See Panel (b) of Figure 7 for an illustration.

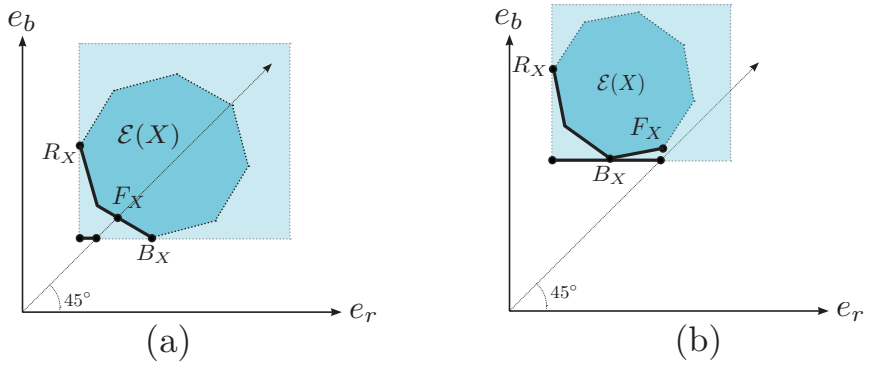


Figure 7: (a) X is group-balanced and (X, G) uniformly Pareto-improves upon X; (b) X is r-skewed and (X, G) does not uniformly Pareto-improve upon X.

This result has the following implications. First, for a large class of covariate vectors (any X that is group-balanced), use of the group identity G leads to a strict welfare improvement *regardless* of the designer's specific fairness-accuracy preferences. Most notably, the Egalitarian designer (who seeks only to minimize the difference in group errors) would strictly prefer for the algorithm to condition on G. This observation is reminiscent of Section 5.1, where we observed that disparate treatment may be necessary to minimize disparate impact.

Second, the *only* case in which precluding use of G can be optimal is the following: If X is r-skewed, and if the designer's optimal point is B_X, then the designer weakly prefers

to discard G.[25,26] Otherwise, prohibiting covariates cannot be justified in our framework, regardless of the designer's concern for fairness or accuracy (as we have defined these concepts). Recall here that our framework can accommodate the case where what the designer cares about is equalizing outcomes rather than equalizing error rates, so a fairness-based justification for prohibiting covariates would have to go beyond these two most common notions.[27,28]

5.3 When X is Group-Balanced

Finally we provide a sufficient condition for a covariate X' to uniformly Pareto improve upon X when X is strictly group-balanced.

Definition 12. Say that X' is *uniformly decision-relevant at* $x \in \mathcal{X}$ if there exist $x', \tilde{x}' \in \mathcal{X}'$ such that:

(i) the optimal action for both groups at (x, x') is uniquely equal to 1

(ii) the optimal action for both groups at $(x, \tilde{x}')$ is uniquely equal to 0

(iii) $\mathbb{P}(X = x, X' = x' \mid G = g), \mathbb{P}(X = x, X' = \tilde{x}' \mid G = g) > 0$ for both groups

This definition says that the realization x is "split" into (x, x') and $(x, \tilde{x}')$, where the optimal action is the same for both groups at each of these realizations, but different across (x, x') and $(x, \tilde{x}')$.

Proposition 6. *Let X and X' be any two covariates, where X is strictly group-balanced. Suppose X' is uniformly decision-relevant at any $x \in \mathcal{X}$. Then X' uniformly Pareto-improves upon X.*

[25]The designer might prefer B_X because he only seeks to minimize the error for group b (because of properties of group b that are outside of this model), or because the designer is Rawlsian and seeks to minimize the error for the more disadvatanged group.

[26]Even in this case, there is another algorithm which *does* condition on G that achieves the same optimal outcome B_X, so conditioning on G need not imply a worse outcome for this designer.

[27]For example, let $\ell(a, y)$ take value 1 if $a = 1$, and 0 otherwise. Then e_g is the fraction of group g individuals who receive action $a = 1$. What our result says is that—unless the designer cares only to maximize the fraction of individuals from group b that receive action 1—then the designer will always strictly benefit from using group identity.

[28]One possible reason to ban protected group identities, which goes beyond our framework, is if the decision-maker does not maximize for accuracy—and in fact, does not respect our Pareto dominance relation at all. For example, the decision-maker may prefer larger differences in group errors, or prefer for the error for a specific group to be large. Additionally, as Kasy and Abebe (2021) point out, an algorithm that is fair in the narrow context of one decision may perpetuate or exacerbate inequalities within a larger context.

This proposition provides a weak sufficient condition for a uniform Pareto improvement: the additional information in X' only needs to allow for a more accurate decision for both groups at *some* realization of the covariate vector X.

To put it more concretely, consider the question of whether access to test scores (X') uniformly Pareto improves upon other application materials (X). The proposition says the following. Suppose first that X is strictly group-balanced.[29] Then, test scores (X') uniformly improve upon the existing covariates whenever there is *some* application profile, such that additionally taking the test score into consideration reverses the decision for at least one individual of each group (with that profile). Our point here is not a recommendation for this (or any) particular domain, since the conditions stated in Proposition 6 would need to be empirically verified on data, and with other domain-specific considerations taken into account. But the results in this section clarify primitive conditions that should be investigated in decisions about whether or not to ban a covariate such as this.

6 Extensions

Other definitions of fairness. In the main text, the designer's fairness criterion is the difference in group errors $|e_r - e_b|$. It is straightforward to see that all of our results extend for any strictly increasing function of this difference.[30]

More than two actions. In the main text we fix the action set to be $\mathcal{A} = \{0, 1\}$. All of our proofs and results about the full design problem directly extend for any $\mathcal{A}$ satisfying $|\mathcal{A}| < \infty$. However, our proof for Theorem 2 (regarding the relationship between the Bayes-design Pareto frontier and the full-design Pareto frontier) relies critically on the assumption of two actions. We leave to future work the question of whether this result extends more generally.

More than two groups. Another interesting extension is characterization of the Pareto frontier for more than two groups. Some of our results, such as Theorem 2, can be shown to directly extend (as stated) to arbitrary finite numbers of groups. But to extend our other results, we would first have to specify a definition of fairness for multiple groups. One possible generalization of the Pareto dominance relationship is to say that a vector of group

[29] It may not be in practice if in aggregate, the existing application materials necessarily lead to a worse average error for one group than another.

[30] Similarly, our results hold for strictly increasing transformations of either group error, although these generalizations are less interpretable.

errors $(e_g)_{g\in\mathcal{G}}$ Pareto dominates another vector $(e'_g)_{g\in\mathcal{G}}$ if $e_g \leq e'_g$ for every group g, and also $|e_g - e_{g'}| \leq |e'_g - e'_{g'}|$ for every pair of groups $g, g' \in \mathcal{G}$, with at least one inequality holding strictly. We conjecture that all of our main results have analogues in this case.

A Proofs for Section 3

A.1 Characterization of Feasible Set

Lemma A.1. *The full-design feasible set $\mathcal{E}(X)$ is a closed and convex polygon.*

Proof. Given algorithm f, we slightly abuse notation to let $f(x)$ denote the probability of choosing action $a = 1$ at covariate x. We further let $x_{y,g}$ denote the conditional probability that $Y = y$ and $G = g$ given $X = x$. Finally, let p_x denote the probability of $X = x$. Then the group error rates can be written as follows:

$$\begin{aligned} e_g(f) &= \mathbb{E}\left[f(X)\,\ell(1,Y) + (1 - f(X))\,\ell(0,Y) \mid G = g\right] \\ &= \sum_x \left(f(x) \sum_y \frac{x_{y,g}}{p_g} \ell(1,y) + (1 - f(x)) \sum_y \frac{x_{y,g}}{p_g} \ell(0,y) \right) \cdot p_x, \end{aligned}$$

where p_g is the prior probability that $G = g$. The set of all feasible error rates is given by

$$\mathcal{E}(X) = \left\{ (e_r(f), e_b(f)) \,:\, f(x) \in [0,1]\ \forall x \right\}.$$

If we let

$$\begin{aligned} E(x) := \Bigg\{ &\lambda \left(\sum_y \frac{x_{y,r}}{p_r} \ell(1,y), \sum_y \frac{x_{y,b}}{p_b} \ell(1,y) \right) \\ &+ (1-\lambda) \left(\sum_y \frac{x_{y,r}}{p_r} \ell(0,y), \sum_y \frac{x_{y,b}}{p_b} \ell(0,y) \right) \,:\, \lambda \in [0,1] \Bigg\} \end{aligned}$$

represent a line segment in $\mathbb{R}^2$, then we see that

$$\mathcal{E}(X) = \sum_{x \in \mathcal{X}} E(x) \cdot p_x.$$

This is a (weighted) Minkowski sum of line segments, which must be a closed and convex polygon. □

A.2 Proof of Theorem 1

First observe that the Pareto frontier must be part of the boundary of the feasible set $\mathcal{E}(X)$, because any interior point (e_r, e_b) is Pareto dominated by $(e_r - \epsilon, e_b - \epsilon)$ which is feasible when ϵ is small.

Consider the group-balanced case, where R_X lies weakly above the 45-degree line and B_X lies weakly below. If $R_X = B_X$, then this point simultaneously achieves minimal error rates for both groups, as well as minimal unfairness. In this case it is clear that the Pareto frontier consists of that single point, which dominates every other feasible point. Another degenerate case is when the entire feasible set $\mathcal{E}(X)$ consists of the line segment $R_X B_X$. Here again it is easy to see that the entire line segment is Pareto undominated, and the result also holds.

Next we show that the upper boundary connecting R_X to B_X (excluding R_X and B_X) is Pareto dominated. One possibility is that the upper boundary consists entirely of the line segment $R_X B_X$. Take any point Q on this line segment, and through it draw a line parallel to the 45-degree line. Then this line intersects the boundary of $\mathcal{E}(X)$ at another point Q' (otherwise we return to the degenerate case above). By our current assumption about the upper boundary, this point Q' must be strictly below the line segment $R_X B_X$. It follows that Q' reduces both group error rates compared to Q, by the same amount. Thus Q' Pareto dominates Q. If instead the upper boundary is strictly above the line segment $R_X B_X$, then through any such boundary point Q we can still draw a line parallel to the 45-degree line. But now let Q^* be the intersection of this line with the extended line $R_X B_X$. If Q^* lies between R_X and B_X, then it is feasible and Pareto dominates Q because both groups' error rates are reduced by the same amount. Suppose instead that Q^* lies on the extension of the ray $B_X R_X$ (the other case being symmetric), then we claim that R_X itself Pareto dominates Q. Indeed, by definition Q must have weakly larger e_r than R_X. And because in this case Q^* is farther away from the 45-degree line than R_X (this is where we use the assumption that R_X is already above that line), Q^* and thus Q also induce strictly larger error rate difference $e_b - e_r$ than R_X. Hence Q has larger e_r, $e_b - e_r$ as well as e_b when compared to R_X, as we desire to show.

To complete the proof for the group-balanced case, we need to show that the lower boundary connecting R_X to B_X is *not* Pareto dominated. R_X (and symmetrically B_X) cannot be Pareto dominated, because it minimizes e_r and conditional on that further minimizes e_b uniquely. Take any other point Q on the lower boundary. If Q lies on the line segment $R_X B_X$, then the lower boundary consists entirely of this line segment. In this case Q minimizes a certain weighted average of group error rates $\alpha e_r + \beta e_b$ across all feasible points, where $\alpha, \beta > 0$ are such that the vector (α, β) is orthogonal to the line segment $R_X B_X$ (which necessarily has a negative slope). Any such point Q cannot be Pareto dominated,

since a dominant point would have smaller $\alpha e_r + \beta e_b$. Finally suppose Q is a boundary point strictly below the line segment $R_X B_X$. Then it minimizes some weighted error rate $\alpha e_r + \beta e_b$, and it will suffice to show that the weights α, β must be positive. Indeed, $\alpha, \beta \leq 0$ cannot happen because Q induces smaller e_r, e_b than Q^* (Q^* defined in the same way as before but now to the top-right of Q) and thus larger $\alpha e_r + \beta e_b$. $\alpha > 0 \geq \beta$ cannot happen because Q induces larger e_r and smaller e_b than R_X, and thus also larger $\alpha e_r + \beta e_b$. Symmetrically $\beta > 0 \geq \alpha$ cannot happen either. So we indeed have $\alpha, \beta > 0$, which implies that Q is Pareto undominated. This proves the result for the group-balanced case.

This argument can be adapted to the group-skewed case as follows. Suppose X is r-skewed, so that R_X and B_X are both above the 45-degree line. To show that the upper boundary connecting R_X to F_X is Pareto dominated, we choose any boundary point Q and (similar to the above) let Q^* be on the extended line $R_X F_X$ such that QQ^* is parallel to the 45-degree line. If Q^* is on the line segment $R_X F_X$ then it is a feasible point that dominates Q. If Q^* lies on the extension of the ray $F_X R_X$, then as before it can be shown that R_X dominates Q. Finally if Q^* lies on the extension of the ray $R_X F_X$, then it must be the case that F_X lies on the 45-degree line (otherwise it will not minimize $|e_r - e_b|$ as defined). In this case Q is a point that is below the 45-degree line, but also above the extended line $B_X F_X$ by convexity of the feasible set. Since F_X already has larger e_b than B_X, we see that Q must in turn have larger e_b than F_X. But then it follows that Q is dominated by F_X because it has larger e_b, larger $e_r - e_b$ (being below the 45-degree line where F_X belongs to), and thus also larger e_r.

It remains to show that the lower boundary connecting R_X to F_X is Pareto undominated. By essentially the same argument, we know that the lower boundary from R_X to B_X is Pareto undominated. As for the lower boundary from B_X to F_X, note that if some point Q here is dominated by another boundary point $\widehat{Q}$, then $\widehat{Q}$ must induce smaller $|e_b - e_r|$. Since $e_b - e_r$ is positive at Q, this means that $\widehat{Q}$ induces smaller $e_b - e_r$ than Q, without the absolute value applied to the difference. So either $\widehat{Q}$ lies on the lower boundary from Q to F_X, or $\widehat{Q}$ belongs to the other side of the 45-degree line (i.e., below it). Either way the alternative point $\widehat{Q}$ must be farther away from B_X than Q on the lower boundary, so that by convexity $\widehat{Q}$ lies above the extended line $B_X Q$. Given that Q already has larger e_b than B_X, this implies that $\widehat{Q}$ has even larger e_b than Q. Hence $\widehat{Q}$ cannot in fact Pareto dominate Q, completing the proof.

A.3 Proof of Corollary 1

Suppose X is group-balanced, then by Theorem 1 the Pareto frontier is the lower boundary from R_X to B_X. Let L_X be the error rate pair that consists of the e_r in R_X and the e_b

in B_X (geometrically, L_X is such that the line segments $R_X L_X$ and $B_X L_X$ are parallel to the axes). Then because R_X, B_X have respectively minimal group error rates in the feasible set, and because we are considering the lower boundary, any point on this lower boundary $\mathcal{P}(X)$ must belong to the triangle with vertices R_X, B_X and L_X. This implies by convexity that each edge of this lower boundary has a negative slope (just note that the first and final edges must have negative slopes). Because of this, if we start from R_X and traverse along this lower boundary, it must be the case that e_r continuously increases while e_b continuously decreases. Thus in the group-balanced case there does not exist any strong fairness-accuracy conflict along the Pareto frontier.

On the other hand, suppose X is r-skewed. Then we claim that B_X and F_X (which are assumed to be distinct) present a strong fairness-accuracy conflict. Indeed, by assumption of r-skewness, B_X is weakly above the 45-degree line. F_X must also be weakly above the 45-degree line because otherwise it would be less fair compared to the point on the line segment $B_X F_X$ that also belongs to the 45-degree line. Thus, the fact that F_X is weakly more fair than B_X implies that F_X entails smaller $e_b - e_r$ than B_X. By definition of B_X, F_X entails larger e_b than B_X. Combining the above two observations, we know that F_X also entails larger e_r than B_X. Hence F_X induces larger group error rates than B_X for both groups, but reduces the difference in error rates. This is a strong fairness-accuracy conflict as we desire to show.

A.4 Proof of Theorem 2

We first characterize the Bayes design feasible set, and later study the Bayes design Pareto set. As argued in the main text, any feasible error rate pair under Bayes design must improve upon the optimal constant action based on the prior, in terms of the society's overall average error rate. Thus the direction $\mathcal{E}^*(X) \subset \mathcal{E}(X) \cap H$ is straightforward.

Conversely, we need to show that a feasible error rate pair (e_r, e_b) that satisfies $p_r e_r + p_b e_b \leq e_0$ can be implemented by some garbling T. Consider a garbling T that maps X to $\Delta(A)$, with the interpretation that the realization of $T(x)$ is the recommended action for the decision maker. If we abuse notation to let $f(x)$ denote the probability that the recommendation is $a = 1$ at covariate x, then this algorithm f needs to satisfy the following obedience constraint for $a = 1$:[31]

$$\sum_x f(x) \left(\sum_y x_y \ell(1, y) \right) \cdot p_x \leq \sum_x f(x) \left(\sum_y x_y \ell(0, y) \right) \cdot p_x,$$

[31] By a version of the revelation principle, such garblings together with the following obedience constraints are without loss for studying the feasible outcomes, in a general setting.

where we recall from the proof of Theorem 1 that x_y is a shorthand for the conditional probability of $Y = y$ given $X = x$.

Symmetrically, the obedience constraint for $a = 0$ is

$$\sum_x (1 - f(x)) \left(\sum_y x_y \ell(0, y) \right) \cdot p_x \leq \sum_x (1 - f(x)) \left(\sum_y x_y \ell(1, y) \right) \cdot p_x.$$

Thus, an error rate pair $e(f) \in \mathcal{E}(X)$ can be Bayes implemented if and only if the above two obedience constraints are satisfied. Rearranging the two, we get

$$\sum_x f(x) \left(\sum_y x_y (\ell(1, y) - \ell(0, y)) \right) \cdot p_x \leq 0$$
$$\sum_x (1 - f(x)) \left(\sum_y x_y (\ell(0, y) - \ell(1, y)) \right) \cdot p_x \leq 0.$$

Rearranging the former gives

$$\sum_x \left(f(x) \sum_y x_y \ell(1, y) + (1 - f(x)) \sum_y x_y \ell(0, y) \right) \cdot p_x \leq \sum_x \sum_y x_y \ell(0, y) \cdot p_x = \sum_y p_y \ell(0, y),$$

where $p_y = \sum_x x_y \cdot p_x$ denotes the prior probability that $Y = y$. Similarly, rearranging the latter obedience constraint gives

$$\sum_x \left(f(x) \sum_y x_y \ell(1, y) + (1 - f(x)) \sum_y x_y \ell(0, y) \right) \cdot p_x \leq \sum_x \sum_y x_y \ell(1, y) \cdot p_x = \sum_y p_y \ell(1, y),$$

Combining the above inequalities, we can rewrite the obedience constraints as

$$\sum_x \left(f(x) \sum_y x_y \ell(1, y) + (1 - f(x)) \sum_y x_y \ell(0, y) \right) \cdot p_x \leq \min_{a \in \{0,1\}} \sum_y p_y \ell(a, y).$$

This precisely says that the population average error rate under the algorithm f should not exceed the minimal error rate achievable in the prior, as we desire to show.

Next we turn to the Pareto set and argue that $\mathcal{P}^*(X) = \mathcal{P}(X) \cap H$. In one direction, if an error rate pair is undominated in $\mathcal{E}(X)$ and Bayes feasible, then it is also undominated in the smaller set $\mathcal{E}^*(X)$. This proves $\mathcal{P}(X) \cap H \subset \mathcal{P}^*(X)$. In the opposite direction, suppose

for contradiction that a certain point $(e_r, e_b) \in \mathcal{P}^*(X)$ does not belong to $\mathcal{P}(X) \cap H$. Since $\mathcal{P}^*(X) \subset \mathcal{E}^*(X) \subset H$, we know that (e_r, e_b) must not belong to $\mathcal{P}(X)$. Thus by definition of $\mathcal{P}(X)$, (e_r, e_b) is Pareto dominated by some other error rate pair $(\widehat{e_r}, \widehat{e_b}) \in \mathcal{E}(X)$. In particular, we must have $\widehat{e_r} \leq e_r$ and $\widehat{e_b} \leq e_b$, which implies $p_r\widehat{e_r} + p_b\widehat{e_b} \leq p_r e_r + p_b e_b \leq e_0$, where the latter inequality uses the assumption that $(e_r, e_b) \in \mathcal{P}^*(X) \subset \mathcal{E}^*(X)$. It follows that the dominant point $(\widehat{e_r}, \widehat{e_b})$ also belongs to H and thus $\mathcal{E}^*(X)$. But this contradicts the assumption that (e_r, e_b) is undominated in $\mathcal{E}^*(X)$. Such a contradiction proves the result.

A.5 Proof of Corollary 2

If X is group-balanced, then by Theorem 1 we know that $\mathcal{P}(X)$ is the part of the boundary of $\mathcal{E}(X)$ that connects R_X to B_X. If $R_X, B_X \in H$, then we claim that this entire boundary belongs to H. Indeed, let L_X be the error rate pair that consists of the e_r in R_X and the e_b in B_X (geometrically, L_X is such that the line segments $R_X L_X$ and $B_X L_X$ are parallel to the axes). Then because R_X, B_X have respectively minimal group error rates in the feasible set, and because we are considering the lower boundary, any point on this lower boundary $\mathcal{P}(X)$ must belong to the triangle with vertices R_X, B_X and L_X. Since R_X, B_X, L_X all belong to H, we deduce that $\mathcal{P}(X) \subset H$. Hence whenever $R_X, B_X \in H$, we have by Theorem 2 that $\mathcal{P}^*(X) = \mathcal{P}(X) \cap H = \mathcal{P}(X)$. In the opposite direction, a clearly necessary condition for $\mathcal{P}^*(X) = \mathcal{P}(X)$ is that R_X and B_X, which belong to $\mathcal{P}(X)$, also belong to $\mathcal{P}^*(X) \subset H$.

This argument proves Corollary 2 in the group-balanced case. Suppose instead that X is r-skewed (a symmetric argument applies to the b-skewed case). To generalize the above argument, we need to show that whenever R_X, F_X belong to H, then so does the entire lower boundary connecting these points. To see this, note that by the definition of B_X and F_X, the lower boundary connecting these two points consists of positively sloped edges.[32] So across all points on this part of the lower boundary, F_X maximizes the population average error rate $p_r e_r + p_b e_b$. Thus the assumption $F_X \in H$ implies that the lower boundary from B_X to F_X belongs to H. In particular $B_X \in H$, which together with $R_X \in H$ implies that the lower boundary from R_X to B_X also belongs to H (same argument as in the above group-balanced case). Hence the entire lower boundary from R_X to F_X belongs to H, as we desire to show.

[32] If we start from B_X and traverse the lower boundary to the right until F_X, then the first edge of this boundary must be weakly positive because B_X has minimum e_b. The final edge of this boundary must also be positive, since otherwise the starting vertex of this edge would be closer to the 45-degree line than F_X. It follows by convexity that the entire boundary from B_X to F_X has positive slopes.

B Proofs for Section 4

B.1 Proof of Proposition 1

We recall the proof of Lemma A.1, where we showed that the feasible set $\mathcal{E}(X)$ can be written as $\sum_x E(x) \cdot p_x$, with $E(x)$ representing the line segment connecting the two points $\left(\sum_y \frac{x_{y,r}}{p_r}\ell(1,y), \sum_y \frac{x_{y,b}}{p_b}\ell(1,y)\right)$ and $\left(\sum_y \frac{x_{y,r}}{p_r}\ell(0,y), \sum_y \frac{x_{y,b}}{p_b}\ell(0,y)\right)$. If X reveals G, then for each realization x, either $x_{y,r} = 0$ for all y or $x_{y,b} = 0$ for all y. Thus each $E(x)$ is a horizontal or vertical line segment, implying that $\mathcal{E}(X)$ must be a rectangle with $R_X = B_X$ being its bottom-left vertex.

Suppose without loss of generality that $R_X = B_X$ lies above the 45-degree line. If the rectangle $\mathcal{E}(X)$ does not intersect the 45-degree line, then it is easy to see that F_X must be the bottom-right vertex of $\mathcal{E}(X)$. In this case the Pareto frontier is the entire bottom edge of the rectangle, which is a horizontal line segment. If instead the rectangle $\mathcal{E}(X)$ intersects the 45-degree line, then F_X is the intersection between the bottom edge of the $\mathcal{E}(X)$ and the 45-degree line. Again the Pareto frontier is the horizontal line segment from $R_X = B_X$ to F_X. This proves the result.

B.2 Proof of Proposition 3

We first show that $R_X = B_X$ under conditional independence. Indeed, in order to minimize a group g's error rate, the algorithm should choose for each realization x the action $a(x)$ that minimizes

$$\sum_y x_{y,g}\ell(a,y),$$

where $x_{y,g}$ continues to denote the conditional probability of $Y = y, G = g$ given $X = x$. Under conditional independence, $x_{y,g}$ can be rewritten as the product $x_y \times x_g$. Since x_g is a constant that can be pulled out of the above sum, the optimal action $a(x)$ equivalently minimizes $\sum_y x_y \ell(a,y)$. From this we see that the optimal action $a(x)$ is the same regardless of whether we seek to minimize e_r or e_b. Hence $R_X = B_X$ as we desire to show.

Now, if $R_X = B_X$ lies on the 45-degree line, then this is the only point in the Pareto frontier, and the result holds vacuously. Otherwise suppose without loss of generality that $R_X = B_X$ lies above the 45-degree line. Then we are in the r-skewed case, and by Theorem 1 the Pareto frontier is the lower boundary of $\mathcal{E}(X)$ from R_X to F_X. Since $R_X = B_X$, the Pareto frontier in this case is also the lower boundary from B_X to F_X. But we know that this part of the lower boundary consists of positively sloped edges. So there is a strong fairness-accuracy conflict everywhere along the frontier.

B.3 Proof of Proposition 2

We continue to follow the notation laid out in the proof of Lemma A.1. Note that under strong independence,

$$\begin{aligned}\frac{x_{y,r}}{x_{y,b}} &= \frac{\mathbb{P}(Y=y, G=r \mid X=x)}{\mathbb{P}(Y=y, G=b \mid X=x)} \\ &\quad \frac{\mathbb{P}(Y=y, G=r, X=x)}{\mathbb{P}(Y=y, G=b, X=x)} \\ &= \frac{\mathbb{P}(G=r \mid Y=y, X=x)}{\mathbb{P}(G=b \mid Y=y, X=x)} = \frac{p_r}{p_b}.\end{aligned}$$

Thus $\frac{x_{y,r}}{p_r} = \frac{x_{y,b}}{p_b}$ for all x, y. It follows that the line segment $E(x)$, which connects the two points $\left(\sum_y \frac{x_{y,r}}{p_r}\ell\left(1,y\right), \sum_y \frac{x_{y,b}}{p_b}\ell\left(1,y\right)\right)$ and $\left(\sum_y \frac{x_{y,r}}{p_r}\ell\left(0,y\right), \sum_y \frac{x_{y,b}}{p_b}\ell\left(0,y\right)\right)$, lies on the 45-degree line. Therefore $\mathcal{E}\left(X\right) = \sum_x E\left(x\right) \cdot p_x$ is also on the 45-degree line.

C Proofs for Section 5

C.1 Proof of Proposition 4

Let $\underline{e}_g = \min\{e_g \mid (e_r, e_b) \in \mathcal{E}(X)\}$ and $\overline{e}_g = \max\{e_g \mid (e_r, e_b) \in \mathcal{E}(X)\}$ be the minimal and maximal feasible errors for group g given X, and define $\underline{e}^*_g = \min\{e_g \mid (e_r, e_b) \in \mathcal{E}(X, X')\}$ and $\overline{e}^*_g = \max\{e_g \mid (e_r, e_b) \in \mathcal{E}(X, X')\}$ to be the corresponding quantities given X and X'. The following lemma says that access to X' reduces the minimal feasible error for group g if and only if X' is decision-relevant over X for group g.

Lemma C.1. *$\underline{e}^*_g < \underline{e}_g$ if X' is decision-relevant over X for group g, and $\underline{e}^*_g = \underline{e}_g$ if it is not.*

Proof. Let $a_g : \mathcal{X} \to \{0,1\}$ be any strategy mapping each realization of X into an optimal action for group g, i.e.,

$$a_g(x) \in \operatorname*{arg\,min}_{a' \in \{0,1\}} \mathbb{E}\left[\ell(a', Y) \mid G = g, X = x)\right] \quad \forall x \in \mathcal{X}.$$

Likewise let $a^*_g : \mathcal{X} \times \mathcal{X}' \to \{0,1\}$ satisfy

$$a^*_g(x, x') \in \operatorname*{arg\,min}_{a' \in \{0,1\}} \mathbb{E}\left[\ell(a', Y) \mid G = g, X = x, X' = x')\right] \quad \forall x \in \mathcal{X},\ \forall x' \in \mathcal{X}'.$$

By optimality of a^*_g,

$$\mathbb{E}\left[-\ell(a^*_g(x, x'), Y) \mid G = g, X = x, X' = x'\right]$$

$$\leq \mathbb{E}\left[-\ell(a_g(x), Y) \mid G = g, X = x, X = x'\right] \quad \forall x \in \mathcal{X}, \forall x' \in \mathcal{X}'. \quad \text{(C.1)}$$

Suppose X' is decision-relevant over X for group g. Then there exist $x \in \mathcal{X}$ and $x', \tilde{x}' \in \mathcal{X}'$ such that the optimal assignment for group g is uniquely equal to 1 at (x, x') and 0 at $(x, \tilde{x}')$, where both (x, x') and $(x, \tilde{x}')$ have positive probability conditional on $G = g$. But then (C.1) must hold strictly at either (x, x') or $(x, \tilde{x}')$. Thus

$$\underline{e}_g^* = \mathbb{E}\left[-\ell(a_g^*(X, X'), Y) \mid G = g\right] < \mathbb{E}\left[-\ell(a_g(X), Y) \mid G = g\right] = \underline{e}_g.$$

If X' is not decision-relevant over X for group g, then (C.1) holds with equality at every x, x', and the equivalence $\underline{e}_g^* = \underline{e}_g$ is immediate. □

We now proceed to prove Part (a) of Proposition 4. Suppose X is r-skewed and X' is decision-relevant over X for group b. Then by Lemma C.1, $\underline{e}_b^* < \underline{e}_b$ while $\underline{e}_r^* \leq \underline{e}_r$. By Proposition 1, the Pareto frontier given X is a horizontal line segment

$$\mathcal{P}(X) = \{(e_r, \underline{e}_b) \,:\, e_r \in [\underline{e}_r, \min\{\overline{e}_r, \underline{e}_b\}]\}.$$

and there are two possibilities for the Pareto frontier $\mathcal{P}(X, X')$. Suppose first that (X, X') is also r-skewed, in which case

$$\mathcal{P}(X, X') = \{(e_r, \underline{e}_b^*) \,:\, e_r \in [\underline{e}_r^*, \min\{\overline{e}_r^*, \underline{e}_b^*\}]\}.$$

Geometrically, $\mathcal{P}(X, X')$ is a horizontal line below $\mathcal{P}(X)$ that lies everywhere above the 45-degree line (Panel (a) in Figure 6). We will show that every point $(e_r, \underline{e}_b) \in \mathcal{P}(X)$ is Pareto dominated by a point on $\mathcal{P}(X, X')$.

First observe that every $(e_r, \underline{e}_b)$ with $e_r \leq \tilde{e}_r^* \equiv \min\{\overline{e}_r^*, \underline{e}_b^*\}$ is Pareto dominated by $(e_r, \underline{e}_b^*)$, which is directly below $(e_r, \underline{e}_b)$ and belongs to $\mathcal{P}(X, X')$. If instead $e_r > \tilde{e}_r^*$, then the point $(e_r, \underline{e}_b)$ is Pareto dominated by $F_{X,X'} = (\tilde{e}_r^*, \underline{e}_b^*)$. To see this, note that both $\tilde{e}_r^* \leq e_r$ and also $\underline{e}_b^* < \underline{e}_b$ (geometrically, the point $(\tilde{e}_r^*, \underline{e}_b^*)$ falls strictly below and weakly to the left of $(e_r, \underline{e}_b)$). Thus it remains to show that $|\tilde{e}_r^* - \underline{e}_b^*| \leq |e_r - \underline{e}_b| = \underline{e}_b - e_r$. In fact, this inequality holds even when e_r is at its maximum value $e_r = \min\{\overline{e}_r^*, \underline{e}_b^*\}$, because the difference between group errors must be weakly smaller at $F_{X,X'}$ than at F_X (as F_X is also a feasible point in $\mathcal{E}(X, X')$). So we do have uniform Pareto improvement in this case.

Now suppose that (X, X') is b-skewed, in which case

$$\mathcal{P}(X, X') = \{(\underline{e}_r^*, e_b) \,:\, e_b \in [\underline{e}_b^*, \min\{\overline{e}_r^*, \underline{e}_b^*\}]\}.$$

Geometrically, $\mathcal{P}(X, X')$ is a vertical line everywhere below the 45-degree line. (See Panel (b) in Figure 6.) We will show that the point $F_{X,X'} = (\underline{e}_r^*, \min\{\overline{e}_b^*, \underline{e}_r^*\})$ Pareto dominates

every point on the original frontier $\mathcal{P}(X)$. To see this, choose an arbitrary $(e_r, \underline{e}_b) \in \mathcal{P}(X)$. Note that $\underline{e}^*_r \leq \underline{e}_r$ and thus

$$\min\{\overline{e}^*_b, \underline{e}^*_r\} \leq \underline{e}^*_r \leq \underline{e}_r < \underline{e}_b,$$

where the final strict inequality holds because X is r-skewed ($\underline{e}_r = \underline{e}_b$ would imply that X is group-balanced). Thus $F_{X,X'}$ induces smaller group error rates than $(e_r, \underline{e}_b)$. By the same argument as before, the difference in group errors must be weakly smaller at $F_{X,X'}$ than at every point on $\mathcal{P}(X)$. So we again have uniform Pareto improvement.

Finally suppose (X, X') is group-balanced, in which case it is a single point (e^*, e^*) where $e^* = \underline{e}^*_r = \underline{e}^*_b$. (See Panel (c) of Figure 6.) Compared to any point $(e_r, e_b) \in \mathcal{P}(X)$, the point (e^*, e^*) involves a strictly smaller group error for group b, a weakly smaller group error for group r, and a weakly smaller difference in group errors. So (e^*, e^*) Pareto dominates every point on $\mathcal{P}(X)$.

Now we consider the other direction of Part (a) and suppose X' is not decision-relevant over X for group b. This implies by Lemma C.1 that $\underline{e}^*_b = \underline{e}_b > \underline{e}_r \geq \underline{e}^*_r$. So (X, X') is also r-skewed, and the Pareto frontier $\mathcal{P}(X, X')$ must lie on the same horizontal line as the original frontier $\mathcal{P}(X)$. It is straightforward to see that for any two points both above the 45-degree line and having the same e_b, neither of them Pareto dominates the other (since a smaller e_r necessarily implies a larger $e_b - e_r$). Thus there is no uniform Pareto improvement in this case, as we desire to show.

We proceed to prove Part (b) of Proposition 4 and suppose X is group-balanced. In this case, Proposition 1 implies that the Pareto frontier $\mathcal{P}(X)$ is a single point on the 45-degree line, which we can denote by (e, e) where $e = \underline{e}_r = \underline{e}_b$. (See Panel (d) of Figure 6.)

Suppose X' is decision-relevant over X for both groups. Then $\underline{e}^*_b < \underline{e}_b = e$ and $\underline{e}^*_r < \underline{e}_r = e$ by Lemma C.1. This means that the Pareto frontier $\mathcal{P}(X, X')$ is either a horizontal line strictly below $\mathcal{P}(X)$, or a vertical line strictly to the left of $\mathcal{P}(X)$. Neither segment intersects with $\mathcal{P}(X)$, implying $(e, e) \notin \mathcal{P}(X, X')$. But since (e, e) eliminates all difference in group errors, it must be that $F_{X,X'}$ does as well. So the point $F_{X,X'}$ involves a weakly smaller difference in group errors relative to (e, e), and a strictly smaller group error for both groups. Thus $\mathcal{P}(X, X')$ uniformly improves upon $\mathcal{P}(X)$.

In the other direction, suppose X' is not decision-relevant over X for group b. Then $\underline{e}^*_b = \underline{e}_b = \underline{e}_r \geq \underline{e}^*_r$ by Lemma C.1, and by Proposition 1, $\mathcal{P}(X, X')$ must be a horizontal line segment with one endpoint at $F_{X,X'} = (e, e)$. Similar to Part (a), such a situation does not present a uniform Pareto improvement.

C.2 Proof of Proposition 5

As in the previous proof, define $\underline{e}_g = \min\{e_g \mid (e_r, e_b) \in \mathcal{E}(X)\}$ and $\overline{e}_g = \max\{e_g \mid (e_r, e_b) \in \mathcal{E}(X)\}$ to be the minimal and maximal feasible errors for group g given X, and define $\underline{e}^*_g = \min\{e_g \mid (e_r, e_b) \in \mathcal{E}(X, G)\}$ and $\overline{e}^*_g = \max\{e_g \mid (e_r, e_b) \in \mathcal{E}(X, G)\}$ to be the corresponding quantities given X and G. We will use the following lemma.

Lemma C.2. *G is not decision-relevant over X for either group g.*

Proof. A necessary condition for G to be decision-relevant over X for group r is for there to exist an $x \in \mathcal{X}$ such that (x, r) and (x, b) both have positive probability conditional on $G = r$. But clearly $\mathbb{P}(X = x, G = b \mid G = r) = 0$. So this cannot be. The same argument shows that G cannot be decision-relevant over X for group b. □

Corollary 4. *The minimal feasible error for group g is the same given X and given (X, G); that is, $\underline{e}^*_g = \underline{e}_g$ for both groups g.*

Proof. Immediate from Lemmata C.1 and C.2. □

Suppose X is not strictly group-balanced. There are two possibilities. First suppose $R_X = B_X = F_X$. Then by Corollary 4 and Proposition 1, the Pareto frontier $\mathcal{P}(X, G)$ is also this singleton point. So a uniform Pareto improvement is not obtained.

Now suppose X is g-skewed for some group; without loss, let it be r-skewed. Then by Corollary 4, $\underline{e}^*_b = \underline{e}_b$ is unchanged, which implies that the point $B_{X,G}$ must remain above the 45-degree line, so that (X, G) is also r-skewed. Moreover, by Proposition 1 the new Pareto frontier $\mathcal{P}(X, G)$ is a horizontal line segment with group b error rate fixed at $\underline{e}^*_b = \underline{e}_b$. But note that B_X on the original frontier also belongs to this horizontal line. So as shown before there is no point on the new frontier that Pareto dominates B_X, which implies that there is no uniform Pareto improvement.

In the other direction, suppose X is strictly group-balanced. Then by Theorem 1, the Pareto frontier $\mathcal{P}(X)$ is a convex piecewise linear curve extending from R_X to B_X, where R_X falls above the 45-degree line while B_X falls below. Without loss, suppose (X, G) is either group-balanced or r-skewed. Then by Proposition 1 and Corollary 4, the Pareto frontier given (X, G) is the horizontal line connecting $R_{X,G} = B_{X,G}$ to $F_{X,G}$, where the group b error is fixed at $\underline{e}_b$.

We will show that every point $(e_r, e_b) \in \mathcal{P}(X)$ is Pareto-dominated by a point on $\mathcal{P}(X, G)$. First observe that every point $(e_r, e_b) \in \mathcal{P}(X)$ with $e_r \leq \min\{\overline{e}^*_r, \underline{e}_b\}$ is Pareto-dominated by the point $(e_r, \underline{e}_b) \in \mathcal{P}(X, G)$ directly below it (note that it must be strictly below, because $e_r \leq \underline{e}_b$ rules out the possibility that $(e_r, e_b) = B_X$ in the group-balanced case). Now

consider any point $(e_r, e_b) \in \mathcal{P}(X)$ for which $e_r > \min\{\overline{e}^*_r, \underline{e}_b\}$. Any such point must be Pareto-dominated by $F_{X,G}$, since $F_{X,G}$ falls strictly to the left and weakly below (e_r, e_b) and involves a weakly smaller difference in group errors. So every point on $\mathcal{P}(X)$ is Pareto dominated by a point on $\mathcal{P}(X, G)$, as desired. This concludes the proof.

C.3 Proof of Proposition 6

Suppose the conditions of the proposition are met at $x_* \in \mathcal{X}$ and $x'_*, \tilde{x}'_* \in \mathcal{X}'$. That is, the optimal action at (x_*, x'_*) is uniquely equal to 1 for both groups, the optimal action at $(x_*, \tilde{x}'_*)$ uniquely equal to 0 for both groups, and both pairs (x_*, x'_*) and $(x_*, \tilde{x}'_*)$ have strictly positive probability conditional on both groups. We will first show that access to X' shifts the Pareto frontier $\mathcal{P}(X)$ down, and subsequently argue that this means that a uniform Pareto-improvement is obtained.

Consider any $(e_r, e_b) \in \mathcal{P}(X)$. Since this error rate is feasible, there exists an algorithm f such that $(e_r, e_b) = (e_r(f), e_b(f))$. Now define $f^* : \mathcal{X} \times \mathcal{X}' \to \Delta(\mathcal{A})$ to satisfy $f^*(x_*, x'_*) = 1$, $f^*(x_*, \tilde{x}'_*) = 0$, and $f^*(x, x') = f(x)$ at every other $x \in \mathcal{X}$, $x' \in \mathcal{X}'$. At least one of $f^*(x_*, x'_*)$ and $f^*(x_*, \tilde{x}'_*)$ must be different from $f(x_*)$. Then

$$\begin{aligned}\mathbb{E}[\ell(f^*(x, x'), Y \mid G = g, X = x, X' = x] \\ < \mathbb{E}[\ell(f(x), Y \mid G = g, X = x, X' = x] \quad \forall x \in \mathcal{X}, \forall x' \in \mathcal{X}\end{aligned}$$

So $e_r(f^*) < e_r(f^*)$ and also $e_b(f^*) < e_b(f^*)$. Thus every point on the Pareto frontier $\mathcal{P}(X)$ has a paired point strictly to the left and below it, which belongs to the feasible set $\mathcal{E}(X, X')$.

We now argue that every point on $\mathcal{P}(X)$ is Pareto-dominated. Let $(e^*, e^*) \equiv F_{X,X'}$. (This point must lie on the 45-degree line, since F_X belongs to the 45-degree line for any group-balanced X, and $F_{X,X'}$ must involve a weakly lower difference in group errors compared to F_X.) Consider any (e_r, e_b) with $e_r < e^* \leq e_b$. Then by the argument above, there exists a paired point (e'_r, e'_b) strictly below it and to the left. By convexity of $\mathcal{E}(X, X')$, we can choose (e'_r, e'_b) to be above the 45-degree line (otherwise replace it by a point on the line connecting it to (e_r, e_b)). By convexity again, we can find another feasible point $(e_r, e''_b) \in \mathcal{E}(X, X')$ on the line connecting (e^*, e^*) and (e'_r, e'_b), which is *directly below* (e_r, e_b). This point (e_r, e''_b) remains above the 45-degree line, so it is clear that it Pareto dominates (e_r, e_b).

An essentially symmetric argument applies to the case where $e_b < e^* \leq e_r$. To complete the proof, note first that e_r, e_b cannot both be strictly smaller than e^*, as that would imply that the error rates under F_X are strictly better than those under $F_{X,X'}$. Thus the remaining possibility is when $e_r, e_b \geq e^*$. If one of these inequalities holds strictly, then (e_r, e_b) is Pareto-dominated by (e^*, e^*). So the final step of the argument is to show that (e^*, e^*) cannot be

on the original Pareto frontier $\mathcal{P}(X)$; in other words, under the assumptions $F_{X,X'}$ must be strictly better than F_X.

Suppose for contradiction that $(e^*, e^*) \in \mathcal{P}(X)$. Then we can find a paired point $(e'_r, e'_b) \in \mathcal{E}(X, X')$ with $e'_r, e'_b < e^*$. Without loss suppose (e'_r, e'_b) is weakly above the 45-degree line. Then we can connect this point to B_X (which falls strictly below the 45-degree line by assumption of strict group balance) and find the intersection of this line segment with the 45-degree line, which we label as (e^{**}, e^{**}). Since (e'_r, e'_b) lies to the bottom left of (e^*, e^*) and B_X lies to its bottom right, we deduce that $e^{**} < e^*$. But then (e^{**}, e^{**}) would be a feasible point in $\mathcal{E}(X, X')$ that Pareto-dominates (e^*, e^*), contradicting the definition that $(e^*, e^*) = F_{X,X'}$. This contradiction proves the result.

D Microfoundation for the Pareto Frontier

We now provide a foundation for our Pareto frontier as the designer-optimal points across a large class of designer preferences. First, we define a *designer preference* to be any preference over error pairs that is weakly in favor of accuracy and fairness.

Definition D.1. A *designer preference* $\succeq$ is any total order such that $e \succeq e'$ whenever $e_r \leq e'_b$, $e_b \leq e'_b$ and $|e_r - e_b| \leq |e'_r - e'_b|$.

The Utilitarian, Rawlsian, and Egalitarian orderings defined in Section 2.1 are all examples of designer preferences.

Given any designer preference $\succeq$, let

$$\mathcal{P}_{\succeq}(X) = \{e \in \mathcal{E}(X) \,:\, e \succeq e' \text{ for all } e' \in \mathcal{E}(X)\}$$

denote the optimal error pairs in $\mathcal{E}(X)$ under $\succeq$. One possible definition of the Pareto frontier is the union of $\mathcal{P}_{\succeq}(X)$ over all $\succeq$, i.e., the set of all optimal points across all designer preferences. But this Pareto frontier is simply the entire feasible set $\mathcal{E}(X)$, since the preference that is completely indifferent over all error pairs is a designer preference. To obtain a more meaningful Pareto frontier, we instead consider sets that include *at least one* optimal point for every designer preference.

Definition D.2. $\mathcal{P} \subset \mathcal{E}(X)$ is *admissible* if for any designer preference $\succeq$, $\mathcal{P}_{\succeq}(X) \neq \emptyset$ implies $\mathcal{P} \cap \mathcal{P}_{\succeq}(X) \neq \emptyset$.

A set is admissible if every designer preference that achieves an optimal point also achieves an optimal point in that set. Clearly, the entire feasible set $\mathcal{E}(X)$ is admissible. Our Pareto set corresponds to the smallest admissible set.

Proposition D.1. *$\mathcal{P}(X)$ is the smallest admissible set in $\mathcal{E}(X)$.*

Proof. We first show that $\mathcal{P}(X)$ is admissible. Fix some designer preference $\succeq$ and let $e^* \in \mathcal{P}_{\succeq}(X)$ be an optimal point. If $e \in \mathcal{P}(X)$ then we already have a nonempty intersection between $\mathcal{P} = \mathcal{P}(X)$ and $\mathcal{P}_{\succeq}(X)$. Suppose $e^* \notin \mathcal{P}(X)$, then there exists some $e^{**} \in \mathcal{E}(X)$ that Pareto-dominates e. In fact, because $\mathcal{E}(X)$ is compact, we can choose e^{**} to belong to the Pareto frontier $\mathcal{P}(X)$ (just choose e^{**} to lexicographically minimize e_r and e_b among those points that Pareto dominate e^*). Now since e^{**} Pareto-dominates e^*, and the designer preference is defined to respect the Pareto ranking, we have that $e^{**} \succeq e^*$. Thus e^{**} must also be an optimal point for the preference $\succeq$, just as e^* is. This shows that $e^{**} \in \mathcal{P} \cap \mathcal{P}_{\succeq}(X)$, which must again be a nonempty set. Thus $\mathcal{P}(X)$ is admissible.

We now show that $\mathcal{P}(X)$ is the smallest admissible set. For any $e^* \in \mathcal{P}(X)$, we can define a designer preference $\succeq$ represented by the utility function w such that $w(e) = 1$ if $e = e^*$ or e Pareto-dominates e^*, and that $w(e) = 0$ otherwise. This preference clearly respects the Pareto ranking, so it is a legitimate designer preference. Moreover, the unique optimal point in $\mathcal{E}(X)$ under this preference is e itself, because by definition of Pareto optimality there cannot exist another point in $\mathcal{E}(X)$ that achieves the utility of 1 as e^* does. Thus any admissible set must include e^*. But since $e^* \in \mathcal{P}(X)$ is arbitrary, we conclude that any admissible set must contain $\mathcal{P}(X)$. This completes the proof. □

The above result shows that our Pareto set $\mathcal{P}(X)$ is minimal in the sense that we cannot exclude any points from $\mathcal{P}(X)$ without hurting some designer. In fact, our proof demonstrates that for every point $e \in \mathcal{P}(X)$, there exists some designer preference $\succeq$ such that e is the *unique* optimal error pair given $\succeq$ within the feasible set $\mathcal{E}(X)$.

Below we provide another characterization of the Pareto set via a simple class of designer preferences. Consider a designer with the following utility over errors

$$w(e_r, e_b) = \alpha_r e_r + \alpha_b e_b + \alpha_f |e_r - e_b|$$

where $\alpha_r, \alpha_b < 0$ and $\alpha_f \leq 0$. Call such designer utilities *simple*. Simple utilities are consistent with Pareto dominance. For example, both the Utilitarian and Rawlsian designers have utilities that are simple. To see this for the Utilitarian designer, set $\alpha_r = -p_r$, $\alpha_b = -p_b$ and $\alpha_f = 0$. To see this for the Rawlsian designer, set $\alpha_r = \alpha_b = \alpha_f = -1$. Our Pareto set corresponds exactly to the set of optimal points for all simple designer utilities.

Proposition D.2. *$e^* \in \mathcal{P}(X)$ if and only if there exists a simple designer utility w such that e^* maximizes w within $\mathcal{E}(X)$.*

Proof. In one direction, we want to show that if e^* maximizes some simple designer utility, then it must be Pareto optimal. Indeed, suppose for contradiction that e^{**} Pareto dominates e^*, then by definition $e_r^{**} \leq e_r^*$, $e_b^{**} \leq e_b^*$ and $|e_r^{**} - e_b^{**}| \leq |e_r^* - e_b^*|$ with at least one strict

inequality. Thus in fact there must be a strict inequality between $e_r^{**} \leq e_r^*$ and $e_b^{**} \leq e_b^*$. It follows that for weights $\alpha_r, \alpha_b < 0$, we must have $\alpha_r e_r^{**} + \alpha_b e_b^{**} > \alpha_r e_r^* + \alpha_b e_b^*$ with strict inequality. Note also that $\alpha_f |e_r^{**} - e_b^{**}| \geq \alpha_f |e_r^* - e_b^*|$ since $\alpha_f \leq 0$. Putting it together, we deduce $w(e_r^{**}, e_b^{**}) > w(e_r^*, e_b^*)$ for every simple designer utility w, contradicting the assumption about e.

In the opposite direction, we want to show that every Pareto optimal point e^* maximizes some simple designer utility. By Theorem 1, e^* must either belong to the lower boundary from R_X to B_X or the lower boundary from B_X to F_X, where the latter case only happens when X is r-skewed (we omit the symmetric situation when X is b-skewed). If e^* belongs to the boundary from R_X to B_X, then from the proof of Theorem 1 we know that e^* belongs to an edge of this boundary that has negative slope. Thus there exists a vector (α_r, α_b) that is normal to this edge, such that e^* maximizes $\alpha_r e_r + \alpha_b e_b$ among all feasible points. Since this edge has negative slope, it is straightforward to see that $\alpha_r, \alpha_b < 0$. So e maximizes the simple utility $\alpha_r e_r + \alpha_b e_b$ as desired.

If instead X is r-skewed and e^* belongs to the boundary from B_X to F_X, then again e^* belongs to an edge of this boundary. But now this edge must have weakly positive slope (since the edge starting from B_X has weakly positive slope by the definition of B_X, and since the boundary is convex). In addition, this slope must be strictly smaller than 1 because otherwise F_X would be farther away from the 45-degree line compared to its adjacent vertex on this boundary. It follows that the outward normal vector (β_r, β_b) to the edge that e^* belongs to satisfies $\beta_r \geq 0 \geq -\beta_r > \beta_b$. The point e^* of interest maximizes $\beta_r e_r + \beta_b e_b$ among all feasible points. Now let us choose any α_f to belong to the interval $(\beta_b, -\beta_r)$, which is in particular negative. Further define $\alpha_r = \beta_r + \alpha_f < 0$ and $\alpha_b = \beta_b - \alpha_f < 0$. Then $\beta_r e_r + \beta_b e_b$ can be rewritten as $\alpha_r e_r + \alpha_b e_b + \alpha_f (e_b - e_r)$. If we consider the simple utility $\alpha_r e_r + \alpha_b e_b + \alpha_f |e_b - e_r|$, then for any other feasible point e^{**} it holds that

$$\begin{aligned}
\alpha_r e_r^{**} + \alpha_b e_b^{**} + \alpha_f |e_b^{**} - e_r^{**}| &\leq \alpha_r e_r^{**} + \alpha_b e_b^{**} + \alpha_f (e_b^{**} - e_r^{**}) \\
&= \beta_r e_r^{**} + \beta_b e_b^{**} \\
&\leq \beta_r e_r^* + \beta_b e_b^* \\
&= \alpha_r e_r^* + \alpha_b e_b^* + \alpha_f (e_b^* - e_r^*) \\
&= \alpha_r e_r^* + \alpha_b e_b^* + \alpha_f |e_b^* - e_r^*|,
\end{aligned}$$

where the first inequality holds since $\alpha_f \leq 0$ and the last equality holds because $e^* \in \mathcal{P}(X)$ must be weakly above the 45-degree line. Hence the above inequality shows that e^* maximizes the simple utility we have constructed, completing the proof. □

References

Agan, A. and S. Starr (2018): "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *The Quarterly Journal of Economics*, 133, 191–235.

Angwin, J. and J. Larson (2016): "Machine bias," ProPublica.

Arnold, D., W. Dobbie, and P. Hull (2021): "Measuring Racial Discrimination in Algorithms," *AEA Papers and Proceedings*, 111, 49—54.

Bergemann, D. and S. Morris (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57, 44–95.

Chouldechova, A. (2017): "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data*, 5, 153–163.

Corbett-Davies, S. and S. Goel (2018): "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," .

Crabtree, T. D., S. J. Pelletier, T. G. Gleason, T. L. Pruett, and R. G. Sawyer (1999): "Gender-Dependent Differences in Outcome After the Treatment of Infection in Hospitalized Patients," *The Journal of the American Medical Association*, 282, 2143–2148.

Diana, E., T. Dick, H. Elzayn, M. Kearns, A. Roth, Z. Schutzman, S. Sharifi-Malvajerdi, and J. Ziani (2021): "Algorithms and Learning for Fair Portfolio Design," in *Proceedings of the 22nd ACM Conference on Economics and Computation*.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012): "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Fang, H. and A. Moro (2011): "Theories of statistical discrimination and affirmative action: A survey," in *Handbook of social economics*, vol. 1, 133–200.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2021): "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *Journal of Finance*.

Gillis, T., B. McLaughlin, and J. Spiess (2021): "On the Fairness of Machine-Assisted Human Decisions," Working Paper.

Grant, S., A. Kajii, B. Polak, and Z. Safra (2010): "Generalized Utilitarianism and Harsanyi's Impartial Observer Theorem," *Econometrica*, 79, 1939–1971.

Hardt, M., E. Price, and N. Srebro (2016): "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems*, 3315–3323.

Harsanyi, J. (1953): "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking," *Journal of Political Economy*, 61, 434–435.

——— (1955): "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment," *Journal of Political Economy*, 63, 309–321.

Jung, C., S. Kannan, C. Lee, M. M. Pai, A. Roth, , and R. Vohra (2020): "Fair Prediction with Endogenous Behavior," Working Paper.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

Kasy, M. and R. Abebe (2021): "Fairness, Equality, and Power in Algorithmic Decision-Making," in *ACM Conference on Fairness, Accountability, and Transparency*.

Kearns, M., S. Neel, A. Roth, and Z. S. Wu (2018): "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*, 2569–2577.

Kearns, M., A. Roth, and S. Sharifi-Malvajerdi (2019): "Average Individual Fairness: Algorithms, Generalization and Experiments," in *Advances in Neural Information Processing Systems*.

Kleinberg, J., J. Ludwig, S. Mullainathan, and A. Rambachan (2018): "Algorithmic Fairness," *AEA Papers and Proceedings*, 108, 22–27.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2017): "Inherent Trade-Offs in the Fair Determination of Risk Scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, vol. 67, 43:1–43:23.

Knight, C. (2013): "Luck Egalitarianism," *Philosophy Compass*, 8, 924—934.

Ludwig, J. and S. Mullainathan (2021): "Algorithmic Behavioral Science: Machine Learning as a Tool for Scientific Discovery," Working Paper.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019): "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 366, 447–453.

Parfit, D. (2002): "Equality or Priority?" in *The Ideal of Equality*, ed. by M. Clayton and A. Williams, New York: Palgrave Macmillan, 81–125.

Rambachan, A., J. Kleinberg, S. Mullainathan, and J. Ludwig (2021): "An Economic Approach to Regulating Algorithms," Working Paper.

Roth, A. and M. Kearns (2019): *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*, Oxford University Press.

Wei, S. and M. Niethammer (2020): "The Fairness-Accuracy Pareto Front," .

Yang, C. S. and W. Dobbie (2020): "Equal Protection Under Algorithms: A New Statistical and Legal Framework," *Michigan Law Review*, 119.