Algorithmic Design: Fairness Versus Accuracy*

Annie Liang† Jay Lu‡ Xiaosheng Mu§

January 21, 2022

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 errors of these algorithms differ sharply across subgroups of the population. What are the tradeoffs between accuracy and fairness, and how do these tradeoffs depend on the inputs to the algorithm? To answer these questions, we propose a model in which a designer chooses an algorithm that maps observed inputs into decisions. We characterize the fairness-accuracy Pareto frontier, and identify how the algorithm's inputs govern the shape of this frontier, showing (for example) that access to group identity necessarily reduces the error for the worse-off group. We then apply these results to study the consequences of banning algorithmic inputs. We show that: (1) banning group identity makes all designers worse off if and only if covariates are *group-balanced*; (2) banning a biased input (such as a test score) makes all designers worse off if group identity is permitted as an input, but can be strictly optimal when group identity is not.

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, Krishna Dasaratha, Sergiu Hart, Yair Livne, and Sendhil Mullainathan for helpful comments, and National Science Foundation Grant SES-1851629 for financial support. We also thank Andrei Iakovlev for valuable research assistance on this project.

†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 important decisions, policymakers have become concerned with the possibility that algorithms are "unfair," in the sense that their errors differ sharply across subgroups of the population. These concerns are supported by a growing body of similar empirical results for algorithms predicting need for medical treatment (Obermeyer et al., 2019), risk of criminal reoffense (Arnold et al., 2021), and risk of mortgage default (Fuster et al., 2021).

Fairness, however, is not the only criterion that matters—we also care about the algorithm's accuracy. This paper seeks to understand how the tradeoff between these objectives is governed by the available inputs to the algorithm. Besides our basic theoretical interest in this problem, we are motivated by practical challenges regarding algorithm design and regulation. For example, what are the consequences for either group of permitting the algorithm access to group identity as an input? If an input is "biased" against a particular group, will sufficiently fairness-minded designers prefer to ban this input?

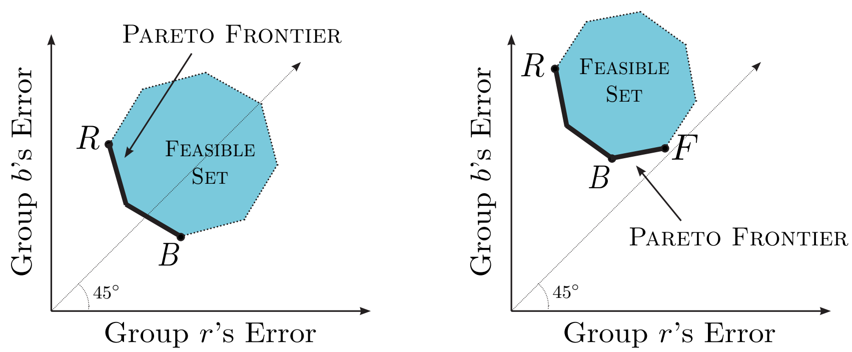
To answer these questions, we propose a framework in which a designer chooses an algorithm that takes observed covariates as inputs (e.g., past criminal background, psychological evaluations, social network data) and outputs an action (e.g., whether or not to recommend bail). The algorithm's consequences for any given individual are measured using a loss function, which can be interpreted either as a measure of the inaccuracy of the algorithm's decision (our leading interpretation), or as the dis-utility received by the individual. We then aggregate losses within members of two pre-defined groups, group r (red) and group b (blue). A *group error* is the expected loss for members of that group. We say that an algorithm is more accurate if it implies lower group errors, and more fair if it implies a smaller difference between the two groups' errors.

We do not commit to a single "right" way of trading off these goals. Instead, we assume that the designer's preferences obey the following order: 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). This weak criterion accommodates a broad range of preferences, including for example Utilitarian designers (who minimize the aggregate error in the population), Rawlsian designers (who minimize the greater of the two group errors), and Egalitarian designers (who minimize the difference between group errors). The *fairness-accuracy Pareto frontier* is the set of all feasible group error pairs (given the inputs to the algorithm) that are Pareto-undominated.

Our results identify a simple property of the algorithm's inputs that is critical to the shape of the fairness-accuracy frontier. If we choose the algorithm that is best for one of the groups, will this algorithm induce a lower error for that group compared to the other?

If so, we say that the covariates are *group-balanced*, and if not, we say the covariates are *group-skewed*. Roughly speaking, covariates can be group-skewed if they are systematically more informative about one group than another. For example, because individuals belonging to a lower socioeconomic (SES) 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 high SES individuals than for low SES individuals (Obermeyer et al., 2019). The algorithm (based on this covariate alone) that is best for the low SES group results in a *higher* error for this group than for the high SES group.

Our first main result says that depending on whether inputs are group-balanced or group-skewed, the frontier takes either of two possible forms, as depicted in Figure 1.



(a) Inputs are Group-Balanced (b) Inputs are Group-Skewed

Figure 1: The Fairness-Accuracy Pareto frontier.

In the case of group-balanced inputs, the frontier begins at the point that is best for group r (labeled R) and ends at the point that is best for group b (labeled B). Moving along the frontier increases one group's error while decreasing the other's. In the case of group-skewed inputs, the frontier again spans the lower boundary of the feasible set (i.e., the group errors achievable using some algorithm) from the best point for group r to the best point for group b. However, in this case the Pareto frontier also includes an additional segment (from B to the fairness-maximizing point F) along which both groups' errors increase.

Can a policy proposal that increases errors for both groups, but reduces the gap between group errors, be justified by fairness considerations? If the algorithm's inputs are group-balanced, then our characterization implies that the answer is *no*: Uniformly increasing both groups' errors necessarily moves off the Pareto frontier, and so cannot be optimal for any designer, regardless of the designer's preferences. Intuitively, group-balance means that inputs do not favor any one particular group, so it is possible to increase fairness by redistributing errors from one group to another. On the other hand, if inputs are group-skewed (as in the healthcare example above), it may be that the only way to decrease the gap in errors is to increase errors for both groups. A designer who places sufficient weight on fairness relative to accuracy may prefer to do this.

We next apply this characterization to derive more specific results for the important special case where covariates reveal group identity. When group identity is known to the algorithm, then the Pareto frontier turns out to be Rawlsian, in the sense that the worse-off group (with the higher error) receives its minimal feasible error at every point on the frontier. Indeed, access to group-identity must reduce the disadvantaged group's error, regardless of how the designer prefers to trade off fairness and accuracy. This is because giving the algorithm access to group-identity permits use of separate rules for each group. This means that it is possible to reduce either group's error without changing the error for the other group. All else equal, reducing the error of the disadvantaged group not only weakly improves accuracy, but also improves fairness, and thus must be preferred by all designers with preferences in our permitted class.

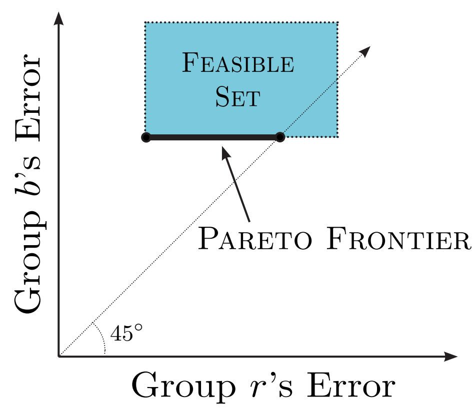


Figure 2: Depiction of the Pareto frontier in the case where X reveals G.

In the second half of the paper, we investigate what happens if the designer does not choose the algorithm, but instead controls the inputs of the algorithm. This question is motivated by settings in which a designer has fairness concerns, but the agent setting the algorithm does not. For example, a judge (agent) determining sentencing may seek to maximize the number of correct verdicts, while a policymaker (designer) may additionally prefer that the accuracy of the judge's verdicts is equitable across certain social groups. In these cases, the policymaker can pass regulation that constrains the inputs that the algorithm can use, for example by excluding the use of a specific covariate in the algorithm.

We model this interaction by supposing that the designer chooses a garbling of the available covariates, and an agent chooses an algorithm (taking the garbling as input) to maximize accuracy. We show that under weak conditions, it is without loss for the designer to have control only of the algorithm's inputs. That is, although the designer cannot choose the algorithm, he can still obtain his most preferred outcome with the use of informational constraints.

We next consider what these informational constraints look like, and whether they might involve excluding a covariate entirely. We study two leading cases. First, we consider the consequences of excluding group identity as an algorithmic input. We show that if and only if the remaining covariates are group-balanced, then every designer—including the Egalitarian designer, who seeks only to minimize the difference in group errors—would strictly prefer for the algorithm to use some information about group identity. These results show that although conditioning on group identity is unfair from the perspective of disparate *treatment* (i.e., whether the policy discriminates between individuals on the basis of a group identity), it 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).

Next, we consider the consequences of excluding a covariate different from group identity—for example, excluding test scores as an input into college admissions decisions. We show that when group identity is also permitted, completely excluding any covariate makes every designer strictly worse off, so long as that covariate satisfies a minimally informative condition we call "decision-relevance." The condition of decision-relevance does not depend on whether the covariate is biased towards either group. Our result thus suggests the following: So long as group identities are permissible inputs for college admission decisions (as is the case in most states in US), then excluding test scores is welfare-reducing for all designer preferences—regardless of how biased the score may be. On the other hand, if group identity is not permitted as an input into college admissions decisions (as is the case in the state of California[1]), we provide an example demonstrating that the optimal garbling of covariates (for some designer preference) may indeed involve completely excluding group identity.

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

[1]Proposition 209 (1996) states that "the government and public institutions cannot discriminate against or grant preferential treatment to persons on the basis of race, sex, color, ethnicity, or national origin in public employment, public education, and public contracting."

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 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 fairness constraint (e.g., requiring approximately equal group errors), 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.[2] 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[3] seeks to eliminate inequality across groups (Parfit, 2002; Knight, 2013).[4] 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

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

[3]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.

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

perspectives, some of which we make explicit in Section 2.1.

Third, in Section 5, we search over possible inputs from a large space of noisy transformations of the available covariates. Here, our (input design) approach is in the tradition of the information design literature (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019), although we consider the Pareto frontier with respect to a large class of "Sender" preferences, and our focus on fairness considerations introduces non-linearities that complicate the Sender's objective function.[5] We view commitment to the information policy in our setting as legally enforceable. (See for example Yang and Dobbie (2020), which summarizes the extant law and proposes new legal policies for mitigating algorithmic bias.)

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 questions regarding algorithmic design.

2 Framework

2.1 Setup and Notation

Consider a population of subjects, each possessing a *covariate* vector X taking values in the finite set $\mathcal{X}$, a *type* Y taking values in the finite set $\mathcal{Y}$, and a *group identity* G taking values r or b.[6] 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.

Each subject receives an action in $\mathcal{A} \equiv \{0, 1\}$ determined by an *algorithm* $f : \mathcal{X} \to \Delta(\mathcal{A})$ that maps 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 considers the special case where X reveals G.) Some motivating examples of types, group identities, covariates, and actions 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

[5] 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.

[6] 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.

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.[7]

Bail. 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 gang members, frequency of moves, or drug use as a child.[8]

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.

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 members of group g. We will subsequently say that an algorithm is more accurate if it implies lower group errors, and more fair if it implies a smaller difference between the two groups' errors.

Our leading interpretation of the loss function is a measure of inaccuracy of the algorithm's decision, and we refer to $e_g(f)$ as error throughout the paper. But 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 interpret $\ell(a, y)$ as the disutility received by a subject with type y and action a, and $e_g(f)$ as the average disutility for members of group g. These two interpretations are associated with different notions of fairness, which are contrasted below:

[7]The Apple Card was investigated for gender discrimination when users noticed in certain cases that smaller lines of credit were offered to wives than to their husbands (but subsequently cleared of these charges). See `https://www.theverge.com/2021/3/23/22347127/goldman-sachs-apple-card-no-gender-discrimination`.

[8]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`.

Example 1 (Measure of Inaccuracy). The type $Y \in \{0, 1\}$ is whether the subject is high or low ability, and the action $a \in \{0, 1\}$ is whether the subject is hired. The loss function is

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

Then $e_g(f)$ is the probability of an inaccurate assessment of the individual, and the algorithm f is unfair if members of one group are substantially more likely to be wrongly evaluated than members of the other.

Example 2 (Measure of Disutility). Fix Y and a as in Example 1. The loss function is

$$\ell(a, y) = \begin{cases} 0 & \text{if } a = 1 \\ 1 & \text{if } a = 0 \end{cases} \tag{2}$$

and reflects subjects' disutility from the outcome: Subjects prefer to be hired regardless of type. Then, $e_g(f)$ is the fraction of individuals in group g who are not hired, and the algorithm f is unfair if members of one group are substantially more likely to be hired than members of the other. If all members of group r are low ability, while all members of group b are high ability, exclusive hiring of individuals from group b is fair using the loss function in Example 1, but not fair using this loss function.

We view the choice of the right loss function as application-specific, and demonstrate results that hold for arbitrary ℓ.

2.2 Fairness-Accuracy Pareto Frontier

We suppose that the designer cares about both accuracy and fairness, preferring lower group errors and also preferring for errors to differ less across groups.[9] We do not privilege a specific way of trading off between these two objectives and view the following partial order as allowing for the largest set of plausible designer preferences.

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.[10]

[9] 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. We consider a standalone and static framework in the present paper, leaving to future work the interesting question of how these algorithimic design decisions might impact outcomes in a larger game, or in a future period.

[10] It is straightforward to see that all of our results extend if our measure of unfairness is replaced with any strictly increasing function of $|e_r - e_b|$.

Below we provide a few prominent examples of designer preferences that are consistent with this partial order; that is, 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).[11]

Example 3 (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 $\succeq_u$.

Example 4 (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 5 (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 then compares ties using the Utilitarian utility w_u. We say that a designer is *Egalitarian* if his preference over error pairs is $\succeq_e$.

The designer can flexibly choose from the set $\mathscr{F}_X$ of all mappings $f : \mathcal{X} \to \Delta(\mathcal{A})$ and the Pareto frontier corresponds to all error pairs that are optimal for some designer whose preference respects our Pareto-dominance ordering (see Appendix B.1 for more detail).

Definition 3. The *feasible set* given covariate X is

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

The *Pareto frontier* 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.

[11] Our definitions for the Rawlsian and Egalitarian designers consider fairness over groups, rather than fairness over individuals. We formulate their preferences in this way because of our motivating settings, but we note also a conceptual challenge with the latter approach: Since individuals cannot be distinguished except through their measured covariates, the "most disadvantaged person" corresponds to the most disadvantaged realization of the measured covariates, which is endogenous to which covariates are measured. Our approach of defining the group as the unit of person (with G pre-defined) avoids this complication.

3 The Fairness-Accuracy Pareto Frontier

We now characterize the fairness-accuracy Pareto frontier. In Section 3.1, we introduce a key statistical property of X, which governs the shape of $\mathcal{P}(X)$. In Section 3.2, we characterize the Pareto frontier for general X, and in Section B.2, we derive characterizations of the Pareto frontier in special cases where X possesses additional structure.

3.1 Key Property: Group-Balance

We begin by defining the property of *group-balance* that will play a key role in several of our results, including our characterization of the Pareto frontier. To define this property, we first introduce certain extreme points of the feasible set. Since the feasible set $\mathcal{E}(X)$ is closed and convex (see Lemma A.1), these points are well-defined.

Definition 4 (Group Optimal Points). For any covariate X, define

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

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

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

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

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

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]

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

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

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, 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 be the misclassification rate, as defined in (1). 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 property of group-balance precisely rules this out.

Definition 6. 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

If X is g-skewed for some group g, then we say it is *group-skewed*.

3.2 General Characterization

Depending on whether the covariates X are group-balanced or group-skewed, the Pareto frontier $\mathcal{P}(X)$ falls into either of two categories. Given two points on the boundary of a compact set, we use *lower boundary* to mean the part of the boundary of the set between the two points, and below the line segment connecting the two.

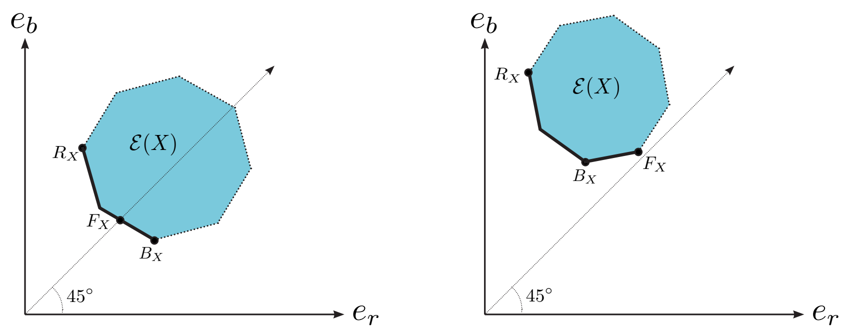
Theorem 1. *The Pareto set $\mathcal{P}(X)$ is the lower boundary of the feasible set $\mathcal{E}(X)$ between*

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

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

These two cases are depicted in Figure 3. When X is group-balanced and R_X and B_X are distinct, then R_X and B_X 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 on 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.[14]



(a) X is group-balanced

(b) X is r-skewed

Figure 3: Example feasible set and Pareto frontier for (a) a group-balanced covariate vector X and (b) an r-skewed covariate vector X.

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

Definition 7. Say that (e_r, e_b) and (e'_r, e'_b) represent a *strong fairness-accuracy conflict* if $e_r \leq e'_r$ and $e_b \leq e'_b$ with at least one inequality strict, while $|e_r - e_b| > |e'_r - e'_b|$. 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 relative to another designer. Both the Utilitarian and Rawlsian designers 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 are relevant 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.*

This corollary is evident from Figure 3. 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.

[14]In the case of g-skewed X, the fairness-optimal point F_X may not lie on the 45 degree line.

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, and so a sufficiently equity-minded designer may prefer to condition less on this information if the error is already higher for the lower socioeconomic class.

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.

4 Group Identity as an Input

We now study how use of group identity as an algorithmic input affects the two groups. We begin by characterizing the Pareto frontier when group identity G is completely revealed by the observed covariate X (either because it is itself an input, or it is perfectly correlated with other inputs), and thus known to the algorithm.

Definition 8. Say that X *reveals* G if the conditional distribution $G \mid X = x$ is degenerate for every realization x of X.

Proposition 1. *Suppose X reveals G. 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.*

[15] 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 separate 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.

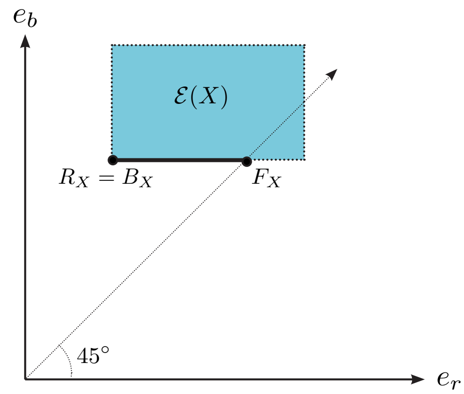


Figure 4: Example feasible set and Pareto frontier when X reveals G.

An example feasible set and Pareto frontier are depicted in Figure 4. One endpoint, the Utilitarian-optimal $R_X = B_X$, gives both groups their minimal feasible error. The other endpoint, the Egalitarian-optimal F_X, maximizes fairness. Everywhere along the Pareto frontier, the worse-off group receives its minimal feasible error, and so:

Corollary 2. *If X reveals G, then every point on the Pareto frontier $\mathcal{P}(X)$ is optimal for a Rawlsian designer.*

To understand this result, consider a simple example where x is the outcome of a lab test. Suppose that a group b individual needs treatment if and only if $x > x_b$, while a group r individual needs treatment if and only if $x > x_r$, where $x_r \neq x_b$. Without access to group identity, the algorithm must assign each realization of x to the same action for individuals in both groups. This ties together the group errors, limiting the error pairs that the designer can achieve given x alone. If in contrast the algorithm is given access to group identity, then the designer can set a separate rule for each group—for example, treating individuals in group r if $x > x_r$ and treating individuals in group b if $x > x_b$. By adjusting either group's rule, the designer can change one group's error without affecting the other. Marginally reducing the larger of the group errors not only weakly improves accuracy, but also improves fairness, and thus must be preferred by all designers with preferences in our permitted class. We generalize this insight in the next section, where we show that access to information about group-identity must improve the error for the worse-off group.

Given a covariate X, say that *group b is the disadvantaged group* if the group b error at B_X is larger than the group r error at R_X; that is, the minimal achievable error for group b (given X) is larger than that for group r. Symmetrically we can define when group r is the disadvantaged group. (In the case of a group-skewed X, the disadvantaged group receives the higher error at every point on the Pareto frontier.)

Definition 9. Say that $w : \mathbb{R}^2 \to \mathbb{R}$ is a *valid designer preference* if w respects the Pareto dominance order in Definition 2, and moreover w achieves a maximum on every compact set.

Corollary 3. *Consider a covariate X such that group b is disadvantaged. Fix any valid designer preference w. Then for any optimal point given X,*

$$(e_r^*, e_b^*) \in \underset{(e_r,e_b)\in\mathcal{E}(X)}{\arg\min}\; w(e_r, e_b),$$

there exists an optimal point given (X, G),

$$(e_r^{**}, e_b^{**}) \in \underset{(e_r,e_b)\in\mathcal{E}(X,G)}{\arg\min}\; w(e_r, e_b),$$

such that $e_b^{**} \leq e_b^*$.

This result says that for any designer preference, group b's error at the designer's optimal point given (X, G) must be weakly smaller than its error at the designer's optimal point given X. The corresponding statement for the advantaged group r is *not* true. For example, when X is r-skewed, the Egalitarian designer may choose to increase group r's error when given more information about G.[16] Thus, more information about group identity must weakly decrease the error for the disadvantaged group, but may increase the error for the advantaged group. Some caution is necessary when interpreting this result: Corollary 3 does not imply that the disadvantaged group's *welfare* necessarily increases when group identity is used, since it could be that the designer cares about inaccuracies (e.g., measuring error using the loss function (1)), while the subjects care about their outcomes (e.g., measuring welfare using loss function (2)). See further discussion in Section 6.

5 Control of Algorithmic Inputs

We have so far assumed that the designer directly chooses the best algorithm according to a preference that takes into account both fairness and accuracy. This is a good description of some settings—for example, a company may internalize fairness concerns in its hiring algorithm. In other settings, the algorithm is set by an agent who does not intrinsically care about equity across groups, but the inputs used by the algorithm are constrained by a regulator who does. For example, a judge determining sentencing may seek to maximize the

[16]See for example Panel (b) of the subsequent Figure 6.

number of correct verdicts, while a policymaker may additionally prefer that the accuracy of the judge's verdicts is equitable across certain social groups. Or, a bank may seek to maximize profit from loans, while a regulator may prefer that no subpopulation is shut out from the possibility of obtaining a loan. In these settings, the regulator can often influence the algorithm by passing regulation that constrains the algorithm's inputs, for example by excluding the use of a specific covariate in the algorithm.

In Section 5.1, we model this interaction by allowing the designer to design the inputs of the algorithm, while the algorithm itself is chosen by another agent. In Section 5.2, we ask when the designer prefers to completely exclude a given input (e.g., group identity) by making any information about this input unavailable.

5.1 Input Design for Algorithms

Suppose a designer first determines what data can be legally used as inputs into the algorithm, and then an agent chooses an algorithm given the permitted inputs. Formally, the designer chooses a *garbling* of the covariate vector X, which is represented as a stochastic map $T : \mathcal{X} \to \Delta(\mathcal{T})$ taking realizations of X into distributions over the possible realizations of T. The following are some common examples.

Example 6 (Banning an Input). $X = (X_1, X_2, X_3)$ and $T(x_1, x_2, x_3) = (x_1, x_2)$ with probability 1.

Example 7 (Adding Noise). $T(x) = x + \varepsilon$ where ε is noise independent of X, Y, G.

Example 8 (No Information). $T(x) = t_0$ with probability 1 for all $x \in \mathcal{X}$.

Given the garbling chosen by the designer, the agent chooses an algorithm $f : \mathcal{T} \to \Delta(A)$ that minimizes

$$\alpha_r \cdot e_r(f) + \alpha_b \cdot e_b(f)$$

where $\alpha_r, \alpha_b \geq 0$. That is, the agent only cares about accuracy and maximizes a utility function that is linear and decreasing in the group errors.[17] Since the agent's utility is linear in group error, we can rewrite the utility as

$$\begin{aligned}
\alpha_r e_r(f) + \alpha_b e_b(f) &= \sum_g \alpha_g \mathbb{E}\left[\ell\left(f(T), Y\right) \mid G = g\right] \\
&= \sum_{t \in \mathcal{T}} p_t \sum_{y,g} \frac{\alpha_g}{p_g} \cdot \mathbb{P}\left(Y = y, G = g \mid T = t\right) \cdot \ell\left(f(t), y\right),
\end{aligned}$$

[17] In Appendix B.3.1 we consider the case in which some coefficient α_g is negative (so that the agent's payoffs are increasing in some group's error).

where p_t is the probability of $T = t$. Thus, the agent's problem of minimizing ex-ante error is equivalent to solving the following ex-post problem[18]

$$f(t) \in \arg\min_{a \in \mathcal{A}} \sum_{y,g} \frac{\alpha_g}{p_g} \cdot \mathbb{P}(Y = y, G = g \mid T = t) \cdot \ell(a, y). \tag{3}$$

The special case when $\alpha_g = p_g$ corresponds to the Utilitarian agent, since the objective function in (3) reduces to $\mathbb{E}(\ell(a, Y) \mid T = t)$.[19]

Definition 10. The pair of group errors (e_r, e_b) is *implemented by* T if there exists an algorithm f_T satisfying (3) such that $(e_r, e_b) = (e_r(f_T), e_b(f_T))$.

Definition 11. The *input-designfeasible set* given covariate X includes all error pairs that the designer can implement by choosing different garblings of X to make available to the agent:

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

The *input-design Pareto frontier* $\mathcal{P}^*(X)$ includes those error pairs $(e_r, e_b) \in \mathcal{E}^*(X)$ that are Pareto-undominated in $\mathcal{E}^*(X)$.

We show that under relatively weak conditions, it is *without loss* to have control only of the algorithm's inputs: Any outcome that is optimal for the designer under unconstrained design can also be achieved under input design. To state the result, we define

$$e_0 = \min_{a \in \mathcal{A}} \left(\alpha_r \cdot \mathbb{E}[\ell(a, Y) \mid G = r] + \alpha_b \cdot \mathbb{E}[\ell(a, Y) \mid G = b]\right)$$

to be the best payoff that the agent can achieve given no information, and

$$H = \{(e_r, e_b) : \alpha_r e_r + \alpha_b e_b \leq e_0\}$$

to be the halfspace including all error pairs that improve the agent's payoff relative to e_0.

Theorem 2 (When Informational Constraints Are Without Loss). *The following hold:*

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

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

[18] When the agent's utility is non-linear in group errors, the ex-ante and ex-post problems are not equivalent in general.

[19] The agent's utility may involve weights different from the utilitarian weights if errors for either group are differentially costly. For example, suppose the agent is a bank manager and group b is wealthier than group r. In this case, the loans for group b may be of higher value, so that incorrectly classifying creditworthy individuals in group r is more costly. This corresponds to scaling the loss ℓ for group r by $\alpha_r/p_r > 1$.

This result follows from the subsequent lemma, which says that the input-designfeasible set is equal to the intersection of the unconstrained feasible set and H, with an analogous statement relating the Pareto frontiers.

Lemma 1. *For every covariate X, the input-design feasible set is $\mathcal{E}^*(X) = \mathcal{E}(X) \cap H$ and the input-design Pareto set is $\mathcal{P}^*(X) = \mathcal{P}(X) \cap H$.*

One direction of the lemma is straightforward: The agent's payoff cannot be made worse off than if the agent were given no information, so $\mathcal{E}^*(X) \subseteq \mathcal{E}(X) \cap H$. We demonstrate the converse: Every point in $\mathcal{E}(X) \cap H$ can be implemented by some garbling of X. Our proof is by construction and garbles X into recommendation of actions. We show that the obedience constraints reduce precisely to the condition that the agent's payoff is improved relative to no information. Thus, the lemma follows. Figure 5 provides an illustration of how Theorem 2 is implied by Lemma 1.

Lemma 1 and Theorem 2 tell us that informational constraints are always sufficient to recover part of the original Pareto frontier. Moreover, so long as certain points improve the agent's payoffs relative to no information, then although the designer does not have explicit control over the algorithm set by the agent, the designer can employ informational constraints 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.

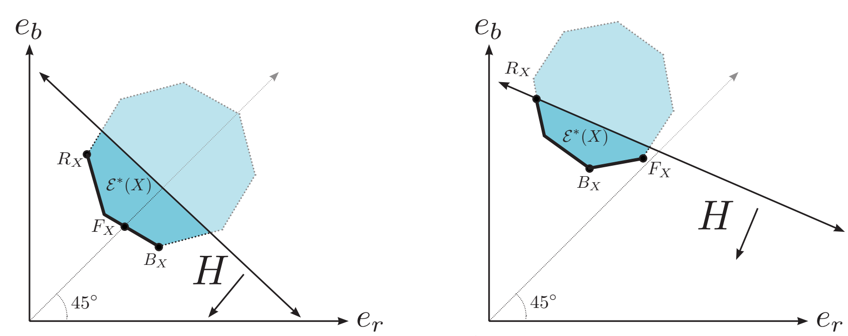


Figure 5: Depiction of an example input-design Pareto frontier for (a) a group-balanced covariate vector X and (b) an r-skewed covariate vector X.

5.2 Excluding a Covariate

In practice, constraints on algorithmic inputs often completely ban use of a given covariate. For example, protected group identities such as race and religion are illegal inputs into lending and hiring decisions,[20] and the University of California university system recently banned consideration of standardized test scores in their admissions decisions.[21]

To study what happens when a covariate is excluded from use, we compare the Pareto frontier when the designer chooses a garbling of (X, X') versus when the designer chooses a garbling of X only. We consider two leading cases: In Section 5.2.1, we suppose that X' is group identity, and in Section 5.2.2, we allow for X' to be an arbitrary covariate (such as a test score) under the assumption that X reveals group identity. Appendix B.3.1 reports an additional result that does not require (X, X') to reveal G.

Excluding a covariate can be strictly optimal for the designer, and we provide an example demonstrating this in Section 5.2.3. But in the settings of Sections 5.2.1 and 5.2.2, we show that although the designer may prefer to *limit* information about the covariate, under weak conditions all designers are made strictly worse off by excluding the covariate. Formally, we will demonstrate conditions for the following:

Definition 12. Say that excluding covariate X' given X *uniformly worsens the frontier* if every point in $\mathcal{P}^*(X)$ is Pareto-dominated by a point in $\mathcal{P}^*(X, X')$.

This property does not imply a comparison between the choice of revealing (X, X') completely versus revealing X completely. So, for example, if the designer does not have flexibility over design of the inputs, the designer may prefer to ban X' rather than to reveal it. Our results do imply however that (under the provided conditions), excluding X' can only be optimal under some exogenous constraint on what the designer can do.

5.2.1 Excluding Group Identity

We first consider the consequences of excluding group identity. The property of group balance (suitably strengthened) emerges as the critical one:

Definition 13. Say that X is *strictly group-balanced* if $e_r < e_b$ at R_X and $e_b < e_r$ at B_X.

[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` and Garg et al. (2021).

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 point on the 45-degree line is strictly group-balanced.

Proposition 2. *Suppose $R_X, B_X \in H$. Then, excluding G given X uniformly worsens the frontier if and only if X is strictly group-balanced.*

The assumption $R_X, B_X \in H$ makes the above result easier to state as an if-and-only-if condition. But it follows from our proof of Proposition 2 that X being strictly group-balanced is a sufficient condition for the frontier to uniformly worsen when excluding G.

The key observation towards this result is that the minimal (and maximal) feasible error for both groups is the same given X and 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. Moreover, we know that when X is group-balanced, then $\mathcal{P}^*(X)$ is characterized by Part (a) of Theorem 1 while $\mathcal{P}^*(X, G)$ is characterized by Proposition 1 (using the equivalence in Theorem 2 for both cases). As depicted in Panel (a) of Figure 6, the Pareto frontier given X does not intersect with the frontier given (X, G), so every point on the new frontier (after excluding G) is dominated by a point on the original frontier. On the other hand, when X is group-skewed, then the two frontiers necessarily overlap as depicted in Panel (b).

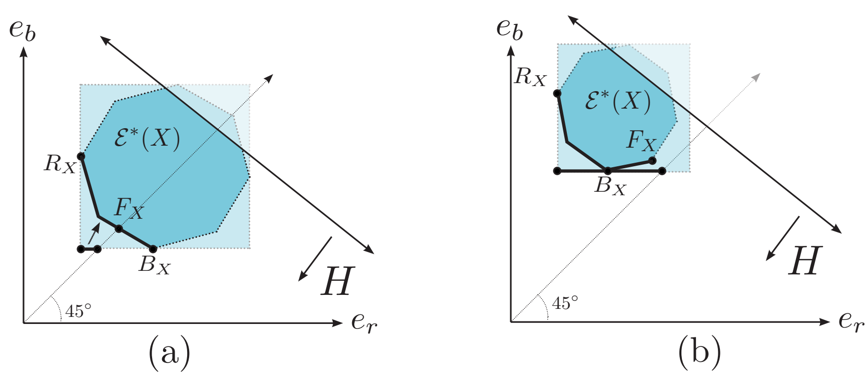


Figure 6: (a) X is strictly group-balanced and excluding G given X uniformly worsens the frontier; (b) X is r-skewed and excluding G given X does not uniformly worsen the frontier.

Proposition 2 implies that for a large class of covariate vectors (any X that is strictly group-balanced), *every* designer can improve their payoffs by choosing a garbling that includes information about G. Most notably, even the Egalitarian designer (who seeks only to minimize the difference in group errors) would strictly prefer for the algorithm to use some information about group identity, beyond what is already available in X.[22]

In our setting, conditioning on G allows the designer to use garblings of X that potentially differ across groups. Traditionally, most policies that restrict use of inputs—for example, the "ban the box" campaign, which restricted employers from asking about criminal histories (Agan and Starr, 2018)—apply symmetrically across groups. Our result shows that even equity-minded designers may strictly prefer to implement noisy transformations that are asymmetric between the two groups. Such policies are 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).[23] Our analysis helps to formalize the tension between these goals, and further demonstrates how to implement such policies in practice.

5.2.2 Excluding a Covariate When Group Identity is Known

Next we consider the case of excluding a covariate (such as a test score) when group identity G is a permitted input.

Definition 14. Say that X' is *decision-relevant over* X *for group* g if there are realizations (x, x') and $(x, \tilde{x}')$ of (X, X') that have strictly positive probability conditional on $G = g$, where 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 a test score, then X' is decision-relevant for group g so long as there is one individual in group g for whom taking the test score into consideration reverses the admission decision.

Proposition 3. *Suppose X reveals G. For any X' we have the following:*

(a) If X is g-skewed, then excluding X' given X uniformly worsens the frontier if and only if X' is decision-relevant over X for group $g' \neq g$.

(b) If X is group-balanced, then excluding X' given X uniformly worsens the frontier if and only if X' is decision-relevant over X for both groups.

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

[22] We show in Appendix B.3.1 that this result extends even if the agent is adversarial against one of the groups (i.e., preferring to increase that group's error) so long as the agent is not "too strongly" adversarial.

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

g. Applying Proposition 1, both the Pareto frontier given X and the Pareto frontier given (X, X') are single line segments. When X' is decision-relevant over X for the disadvantaged group, then the minimal feasible error for that group is strictly lower, pushing the Pareto frontier downwards (see Panel (a) of Figure 7). On the other hand, 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 (see Panel (b) of Figure 7).

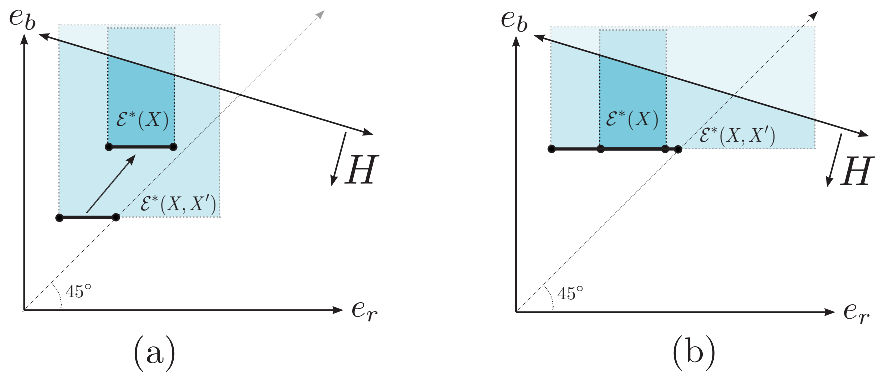


Figure 7: (a) Example in which X' is decision-relevant for group b, and excluding X' uniformly worsens the frontier; (b) Example in which X' is not decision-relevant for group b, and excluding X' does not uniformly worsen the frontier.

There is currently an active policy debate concerning whether universities should permit test scores as an input into admissions decisions. The condition of decision-relevance does not depend on whether the covariate X' is "biased"—in the sense of being systematically lower-valued or less informative for either group—and is very likely satisfied by test scores in practice.[24] Our result thus suggests the following: So long as group identities are permissible inputs for college admission decisions (as is the case in most states in US), then excluding test scores is welfare-reducing for all designer preferences—regardless of how biased the score may be. On the other hand, if group identity is not permitted as an input into college admissions decisions (as is the case in the state of California), then it may be that for the designer's fairness-accuracy preferences, the optimal garbling of covariates indeed involves completely excluding group identity, as in the next section.

[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.

5.2.3 Excluding a Covariate Can Be Strictly Optimal

We conclude with a simple example in which completely excluding a covariate is strictly optimal for the designer.[25] Suppose $\mathcal{Y} = \{0, 1\}$ and the Y, G independently and uniformly distributed, i.e., $\mathbb{P}(Y = y, G = g) = 1/4$ for any $y \in \{0, 1\}$ and $g \in \{r, b\}$. Let X be a null signal, that is, $X = x_0$ with probability one. Further let X' be a binary signal with the following conditional probabilities $\mathbb{P}(X' \mid Y, G)$:

	$X' = 1$	$X' = 0$
$Y = 1$	1	0
$Y = 0$	0	1

$G = r$

	$X' = 1$	$X' = 0$
$Y = 1$	0.6	0.4
$Y = 0$	0.4	0.6

$G = b$

Thus, X' is perfectly informative about the individuals in group r, and imperfectly informative about those in group b. Suppose the loss function ℓ is the misclassification rate, as defined in (1), and the agent is Utilitarian ($\alpha_r = p_r = 1/2$ and $\alpha_b = p_b = 1/2$).

We now study the input-design feasible set $\mathcal{E}^*(X, X') = \mathcal{E}^*(X')$. Without loss, we can restrict attention to garblings of X' that take two values, $a = 1$ and $a = 0$, which correspond to the designer's action recommendation for the agent. Any such garbling can be identified with a pair (α, β), where α is the probability with which $X' = 1$ is mapped into $a = 1$, and β is the probability with which $X' = 0$ is mapped into $a = 1$. It is easy to show that the agent's obedience constraint reduces to the simple inequality $\alpha \geq \beta$, which intuitively requires the agent to take the action $a = 1$ more often when $X' = 1$ and thus Y is more likely to be 1.

For any pair (α, β), group r's error can be calculated as

$$e_r(\alpha, \beta) = \frac{1}{2}(1 - \alpha) + \frac{1}{2}\beta = 0.5 - 0.5(\alpha - \beta),$$

while group b's error is

$$e_b(\alpha, \beta) = \frac{1}{2} \cdot 0.6(1 - \alpha) + \frac{1}{2} \cdot 0.4(1 - \beta) + \frac{1}{2} \cdot 0.4\alpha + \frac{1}{2} \cdot 0.6\beta = 0.5 - 0.1(\alpha - \beta).$$

So as $\alpha - \beta$ ranges from 0 to 1, the implementable group errors constitute the line segment connecting $(0, 0.4)$ with $(0.5, 0.5)$. This entire line segment is also the Pareto frontier $\mathcal{P}^*(X, X')$, as illustrated in Figure 8:

[25] This example falls outside of the settings considered in the previous sections.

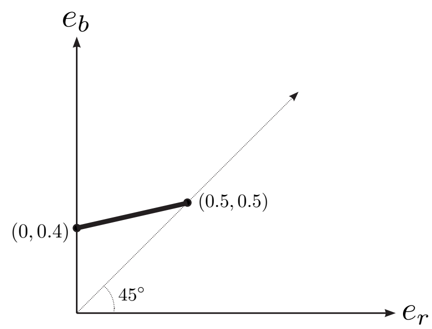


Figure 8: Illustration of the input-design Pareto frontier given (X, X')

Now suppose the designer is Egalitarian. Then, sending the null signal X leads to a payoff of 0. But if the designer chooses any nontrivial garbling of (X, X'), namely any $\alpha \neq \beta$, then the designer's payoffs become $-0.4(\alpha - \beta) < 0$. Thus, the designer's payoffs strictly reduce when any information about X' is provided to the agent, and so the designer strictly prefers to exclude X'. Intuitively, any information the designer provides will be used by the agent to maximize aggregate accuracy, but this information is inevitably more informative about group r, thus increasing the gap between the two group's errors.

6 Extensions

Characterizations of the Pareto set. We have informally described the Pareto frontier as the set of optimal points for designers whose preferences respect the Pareto dominance relation in Definition 2. In Appendix B.1, we formalize this connection with three characterizations of the Pareto set. First, we show that the Pareto set is the smallest set containing an optimal point for every (permitted) designer preference. So our Pareto set is minimal in the sense that we cannot exclude any points from it without hurting some designer. Second, we show that for every point in the Pareto set, there is some designer preference for whom that point is uniquely optimal. Third, we characterize the Pareto set as the set of optimal points for a class of "simple" designer preferences, which are linear in accuracy (group errors) and fairness (difference in group errors).

Group-dependent loss functions. We have defined our loss function to be a function of the subject's action and type. A natural extension is to consider loss functions that are group-dependent. For example, in the healthcare example, if G is ethnic background, then the same medical procedure may have different risk levels depending on group identity. In the lending example, if G is socioeconomic background, then a bank manager may value loans to the wealthier group more and attach greater costs to errors in the less wealthy group. In these cases, the loss ℓ would depend on G. All of our main results, including the characterizations of the Pareto set (Theorems 1 and 2) hold for group-dependent loss functions.

Other agent preferences. Section 5 considers misaligned incentives between a designer controlling inputs and an agent setting the algorithm. There, we assume that the agent cares about accuracy and prefers for both group errors to be lower. In Appendix B.3.1, we consider what happens when this misalignment is more extreme and the agent is adversarial (i.e. negatively biased) towards one of the two groups, preferring for that groups' errors to be higher. We generalize several results from Section 5 and show that, perhaps surprisingly, even if the agent is negatively biased (but the bias is not too extreme), it can still be optimal for the designer to provide information about group identity.

Another potential generalization would be to permit the agent and designer to have different loss functions. When the agent's loss function is different from the designer's, the set of points that the agent prefers over the prior (what we defined to be H) is no longer guaranteed to be a halfspace from the designer's perspective. This introduces interesting technical complications, and we leave the problem of different loss functions to future work.[26]

Finally, we have assumed that the agent only cares about accuracy and does not have fairness concerns. This assumption is more critical, since fairness concerns introduce non-linearities into the agent's objective function. With linearity, the agent's ex-ante and ex-post problems are the same. Without linearity, the ex-ante and ex-post problems may not be the same, so this raises the question of whether the agent commits to the algorithm or chooses the action after the realization of the garbling. We conjecture that the introduction of fairness concerns in the agent's preferences makes it harder for the designer to implement desired outcomes.

More than two actions. We have assumed that there are two actions $\mathcal{A} = \{0, 1\}$. All of our results in Section 3 about the unconstrained problem directly extend for any finite $\mathcal{A}$. However, Lemma 1 (the relationship between the input-designPareto frontier and the unconstrained Pareto frontier) relies critically on the assumption of two actions; see Appendix B.4 for a counterexample to this lemma when $|\mathcal{A}| > 2$. With more than two actions, a characterization of the Pareto set will be more complicated, and we leave its analysis for future work.

[26]Our result does include the special case when the agent's loss function $\ell_a = \alpha_g \ell_d$ is just a group-specific multiple of the designer's loss function. This is mathematically equivalent to the setup in Section 5

More than two groups. We have assumed that there are two groups $\mathcal{G} = \{r, b\}$. Some of our results, such as Theorem 2 and Lemma 1, can be shown to directly extend for any finite $\mathcal{G}$. However, in order 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 errors $(e_g)_{g\in\mathcal{G}}$ Pareto dominates another vector $(e'_g)_{g\in\mathcal{G}}$ if $e_g \le e'_g$ for every group g, and also $|e_g - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} e_g| \le |e'_g - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} e'_g|$ for every $g \in \mathcal{G}$, with at least one inequality holding strictly. That is, fairness is improved if each group's error is closer to the average group error. In this case, Proposition 1 generalizes to the feasible set being a hyperrectangle with sides parallel to the axes' hyperplanes, again with one group receiving its minimal feasible error everywhere along the frontier.

A Proofs for Results in Main Text

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 errors 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 errors 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. $\square$

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 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 of $\mathcal{E}(X)$ 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 errors 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' errors 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 group error 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 errors $\alpha e_r + \beta e_b$ 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 sum of group errors $\alpha e_r + \beta e_b$, 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 group error 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 errors 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 errors than B_X for both groups, but reduces the difference in group errors. This is a strong fairness-accuracy conflict as we desire to show.

A.4 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 $\mathcal{E}(X)$ 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.

A.5 Proof of Corollary 2

Suppose without loss of generality that $R_X = B_X$ lies above the 45-degree line. Then from Proposition 1 we know that the Pareto frontier is the horizontal line segment from $R_X = B_X$ to F_X. Thus, every point on the Pareto frontier has the same group b error as B_X, which is the minimal feasible error given the covariate X. For concreteness let us use $\underline{e}_b$ to denote this minimal group b error. Then we have $e_r \leq e_b = \underline{e}_b$ at every Pareto optimal point, where the first inequality holds because such a point lies above the 45-degree line. A Rawlsian designer whose utility function is $-\max\{e_r, e_b\}$ thus gets $-\underline{e}_b$ in payoff at any Pareto optimal point. On the other hand, any feasible point (e_r, e_b) satisfies $e_b \geq \underline{e}_b$ by definition of $\underline{e}_b$. Thus $-\max\{e_r, e_b\} \leq -e_b \leq -\underline{e}_b$, showing that a Rawlsian designer's payoff is maximized along the Pareto frontier.

A.6 Proof of Corollary 3

From the definition it is easy to see that if group b is disadvantaged given covariate X, then when given covariate (X, G) we have that $B_{X,G} = R_{X,G}$ lies above the 45-degree line (in fact, the group b error at $B_{X,G}$ is the same as the group b error at B_X, similarly for group r). Thus, every Pareto optimal point given (X, G) achieves the minimal feasible group b error given (X, G). Now, for any valid designer preference w, there must exist an optimal point that lies on the Pareto frontier given (X, G). Such an optimal point (e_r^{**}, e_b^{**}) thus achieves the minimal feasible group b error given (X, G), which is weakly lower than the minimal feasible group b error given X alone. It follows that $e_b^{**} \leq e_b^*$ for any feasible point (e_r^*, e_b^*) given X. This comparison certainly holds also for any optimal point (e_r^*, e_b^*) given X.

A.7 Proof of Lemma 1

We first characterize the input-design feasible set, and later study the input-design Pareto set. It is clear that regardless of what garbling the designer gives the agent, the agent's payoff will be weakly better than what can be achieved under no information. Thus any error pair that is implementable by input-design must belong to the halfspace H. Such an error pair must also belong to the feasible set $\mathcal{E}(X)$, so we obtain the easy direction $\mathcal{E}^*(X) \subseteq \mathcal{E}(X) \cap H$ in the lemma.

Conversely, we need to show that a feasible error pair $(e_r, e_b) \in \mathcal{E}(X)$ that satisfies $\alpha_r e_r + \alpha_b e_b \le e_0$ can be implemented by some garbling T. We will in fact prove this result for a general group-dependent loss function $\ell(a, y, g)$, which covers an extension discussed in Section 6.

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 agent. 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 $a = 1$:[27]

$$\sum_{y,g} \frac{\alpha_g}{p_g} \sum_x p_{x,y,g} \cdot f(x) \cdot \ell(1, y, g) \le \sum_{y,g} \frac{\alpha_g}{p_g} \sum_x p_{x,y,g} \cdot f(x) \cdot \ell(0, y, g).$$

The above is just a direct generalization of equation (3) to group-dependent loss functions. It is adapted to the current setting with the observation that given the recommendation $T = 1$, the conditional probability of $Y = y$ and $G = g$ is proportional to the recommendation probability $\sum_x p_{x,y,g} \cdot f(x)$, where we use $p_{x,y,g}$ as a shorthand for $\mathbb{P}(X = x, Y = y, G = g)$.

Let us rewrite the above displayed equation as

$$\sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot f(x)\ell(1, y, g) \le \sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot f(x)\ell(0, y, g).$$

If we add $p_{x,y,g} \frac{\alpha_g}{p_g}(1 - f(x))\ell(0, y, g)$ to each summand above, we obtain

$$\sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot (f(x)\ell(1, y, g) + (1 - f(x))\ell(0, y, g)) \le \sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot \ell(0, y, g). \tag{A.1}$$

Now, the LHS above can be rewritten as $\sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot \mathbb{E}[\ell(A, y, g) \mid X = x, Y = y, G = g]$, which is also equal to $\sum_g \alpha_g \cdot \mathbb{E}[\ell(A, Y, g) \mid G = g]$. This is precisely the agent's expected loss when following the designer's recommended actions.

On the other hand, the RHS in (A.1) can be seen to be the agent's expected loss when taking the action $a = 0$ regardless of the designer's recommendation. Thus, we deduce that the obedience constraint for the recommendation $a = 1$ is equivalent to (A.1), which simply says that the agent's payoff under the designer's recommendation should be weakly better than the constant action $a = 0$ ignoring the recommendation. Symmetrically, the other obedience constraint for the recommendation $a = 1$ is equivalent to the agent's payoff being better than the constant action $a = 1$. Put together, these obedience constraints thus reduce

[27] 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.

to the requirement that the designer's recommendation gives the agent a payoff that exceeds what can be achieved with no information.

For any error pair (e_r, e_b) that is feasible under unconstrained design, we can construct an action recommendation/garbling T that implements it assuming that the recommendation would be obedient for the agent. If (e_r, e_b) belongs to the halfspace H, then by the previous analysis we know that obedience is satisfied. Thus (e_r, e_b) is implementable under input-design, showing that $\mathcal{E}(X) \cap H = \mathcal{E}^*(X)$ as desired.

Finally we turn to the Pareto set and argue that $\mathcal{P}^*(X) = \mathcal{P}(X) \cap H$. In one direction, if an error pair is undominated in $\mathcal{E}(X)$ and implementable under input design, then it is also undominated in the smaller set $\mathcal{E}^*(X)$. This proves $\mathcal{P}(X) \cap H \subseteq \mathcal{P}^*(X)$. 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 $\mathcal{P}^*(X) \subseteq \mathcal{E}^*(X) \subseteq H$, 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 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 $\alpha_r \widehat{e_r} + \alpha_b \widehat{e_b} \leq \alpha_r e_r + \alpha_b e_b \leq e_0$ (the first inequality uses $\alpha_r, \alpha_b \geq 0$ and the second uses $(e_r, e_b) \in \mathcal{P}^*(X) \subseteq \mathcal{E}^*(X)$). 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 completes the proof.

A.8 Proof of Theorem 2

We will deduce Theorem 2 from Lemma 1. 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 from below. Clearly, $\mathcal{P}^*(X) = \mathcal{P}(X)$ can only hold if $R_X, B_X \in \mathcal{P}^*(X) \subseteq H$, so we focus on the "if" direction of the result. Suppose $R_X, B_X \in H$, then we claim that the entire lower boundary of $\mathcal{E}(X)$ from R_X to B_X belongs to H. Indeed, let L_X be the error 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. Because R_X, B_X have respectively minimal group errors in the feasible set $\mathcal{E}(X)$, 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 the halfspace H ($L_X \in H$ because the agent's payoff weights α_r, α_b are non-negative), we deduce that $\mathcal{P}(X) \subseteq H$. Hence whenever $R_X, B_X \in H$, we have by Lemma 1 that $\mathcal{P}^*(X) = \mathcal{P}(X) \cap H = \mathcal{P}(X)$. This argument proves Theorem 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.[28] So across all points on this part of the lower boundary, F_X maximizes $\alpha_r e_r + \alpha_b e_b$. 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 (by the same argument as in the group-balanced case before). Hence the entire lower boundary from R_X to F_X belongs to H, as we desire to show.

A.9 Proof of Proposition 2

We first present a simple lemma which conveniently restates the property of "uniform worsening of frontier":

Lemma A.2. *Excluding covaraite X' given X uniformly worsens the frontier if and only if $\mathcal{P}^*(X)$ does not intersect with $\mathcal{P}^*(X, X')$.*

The proof of this lemma is straightforward: If there exists a point in $\mathcal{P}^*(X)$ that also belongs to $\mathcal{P}^*(X, X')$, then this point is not Pareto-dominated by any point in $\mathcal{P}^*(X, X')$, so that the frontier does not uniformly worsen when excluding X'. On the other hand, suppose no point in $\mathcal{P}^*(X)$ belongs to $\mathcal{P}^*(X, X')$. Note that any point in $\mathcal{P}^*(X)$ is implementable via a garbling of X and thus implementable via a garbling of X, X'. Thus any such point belongs to $\mathcal{E}^*(X, X')$, and since it is not Pareto-optimal in this set, it must be Pareto-dominated by some Pareto optimal point in this (compact) set. In this case we do have uniform worsening of the frontier, as we desire to show.

Below we use Lemma A.2 to deduce Proposition 2. The key observation is that whether or not G is excluded does not affect the minimal (or maximal) feasible error for either group. This is because if we want to minimize the error of a particular group g using an algorithm that depends on X, then we essentially condition on $G = g$ anyways.

With this observation, suppose X is strictly group-balanced. Then R_X lies strictly above the 45-degree line and B_X lies strictly below. Since we assume $R_X, B_X \in H$, Theorem 2 tells us that the input-design Pareto frontier $\mathcal{P}^*(X)$ is the same as the unconstrained Pareto frontier $\mathcal{P}(X)$, and by Theorem 1 this frontier is the lower boundary of the feasible set $\mathcal{E}(X)$ connecting R_X to B_X. By Lemma A.2, we just need to show that in this case the lower boundary of $\mathcal{E}(X)$ from R_X to B_X does not intersect with the input-design Pareto frontier $\mathcal{P}^*(X, G)$ given (X, G). To characterize the latter frontier, let $L_X = R_{X,G} = B_{X,G}$ denote

[28] 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.

the error pair that has the same e_r as R_X and the same e_b as B_X. Without loss of generality assume L_X lies weakly above the 45-degree line. Then from Proposition 1 we know that the unconstrained Pareto frontier $\mathcal{P}(X,G)$ is the horizontal line segment from L_X to $F_{X,G}$. This point $F_{X,G}$ is the intersection between the line segment $L_X B_X$ and the 45-degree line (here we use the fact that L_X lies above the 45-degree line and B_X lies below). As $B_X \in H$, the points L_X and $F_{X,G}$ also belong to H because they have equal e_b and smaller e_r compared to B_X. Hence the input-design Pareto frontier $\mathcal{P}^*(X,G)$ is also the line segment from L_X to $F_{X,G}$. To see that this horizontal line segment does not intersect the boundary of $\mathcal{E}(X)$ from R_X to B_X, just note that B_X is the only point on that boundary with the same (minimal) e_b as any point on the horizontal line segment. But B_X does not belong to that line segment because it is strictly below the 45-degree line. This proves the result when X is strictly group-balanced.

Now suppose X is not strictly group-balanced. Then R_X and B_X lie weakly on the same side of the 45-degree line, and without loss of generality let us assume they lie weakly above. It is still the case that the unconstrained Pareto frontier $\mathcal{P}(X,G)$ is the horizontal line segment from L_X to $F_{X,G}$. But in the current setting $F_{X,G}$ must be weakly closer to the 45-degree line than B_X, which means that B_X now lies in between L_X and $F_{X,G}$. In other words, $B_X \in \mathcal{P}(X)$ and $B_X \in \mathcal{P}(X,G)$. But by assumption, B_X also belongs to H. So Lemma 1 tells us that B_X belongs to the input-design Pareto frontiers $\mathcal{P}^*(X)$ and $\mathcal{P}^*(X,G)$. This shows that the two frontiers $\mathcal{P}^*(X)$ and $\mathcal{P}^*(X,G)$ intersect, which completes the proof by Lemma A.2.

A.10 Proof of Proposition 3

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 A.3. *$\underline{e}^*_g < \underline{e}_g$ 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.,

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

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^*,

$$\begin{aligned}\mathbb{E}&\left[\ell(a_g^*(x,x'),Y) \mid G=g, X=x, X'=x'\right] \\ &\le \mathbb{E}\left[\ell(a_g(x),Y) \mid G=g, X=x, X=x'\right] \quad \forall x \in \mathcal{X}, \forall x' \in \mathcal{X}'. \end{aligned} \tag{A.2}$$

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 (A.2) must hold strictly at either (x,x') or $(x,\tilde{x}')$. Thus, by taking the expectation of (A.2) conditional on $G=g$, we obtain

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

If X' is not decision-relevant over X for group g, then (A.2) holds with equality at every x, x', and the equivalence $\underline{e}_g^* = \underline{e}_g$ follows. □

We now use Lemma A.2 and A.3 to prove Proposition 3. First suppose X is r-skewed. Together with the assumption that X reveals G, we know that $R_X = B_X$ lies strictly above the 45-degree line. In this case the unconstrained Pareto frontier $\mathcal{P}(X)$ is the horizontal line segment from $R_X = B_X$ to F_X, by Proposition 1.

Now if X' is not decision-relevant over X for group b, then from Lemma A.3 we know that the minimal feasible error for group b is the same given (X,X') as given X. Note that the group b minimal error given X exceeds the group r minimal error given X. The former remains the same given (X,X'), while the latter becomes weakly smaller. Thus the group b minimal error given (X,X') also exceeds the group r minimal error given (X,X'). In other words, $R_{X,X'} = B_{X,X'}$ also lies strictly above the 45-degree line, and the Pareto frontier $\mathcal{P}(X,X')$ is the horizontal line segment from $R_{X,X'} = B_{X,X'}$ to $F_{X,X'}$. Crucially, this line segment shares the same e_b as the line segment from $R_X = B_X$ to F_X. In addition, as $R_{X,X'}$ must have weakly smaller e_r than R_X, and $F_{X,X'}$ must be weakly closer to the 45-degree line than F_X, we deduce that the unconstrained Pareto frontier $\mathcal{P}(X,X')$ is a horizontal line segment that is a superset of the line segment $\mathcal{P}(X)$. Thus, in particular, $R_X = B_X$ belongs to both of these frontiers. Lemma 1 thus imply that $R_X = B_X$ also belongs to the input-design Pareto frontiers $\mathcal{P}^*(X)$ and $\mathcal{P}^*(X,X')$ ($R_X = B_X$ belongs to H because this point can be implemented by giving X to the agent, who will then minimize both groups' errors given this information). By Lemma A.2, uniform worsening of the frontier does not occur when excluding X', as we desire to show.

If X' is decision-relevant over X for group b, then Lemma A.3 tells us that $\underline{e}_b^* < \underline{e}_b$ with strict inequality. There are two cases to consider here. One case involves $\underline{e}_b^* > \underline{e}_r^*$, so that (X, X') is r-skewed just as X is. Then the unconstrained Pareto frontier $\mathcal{P}(X, X')$ is again a horizontal line segment, but with e_b equal to $\underline{e}_b^*$. Since $\underline{e}_b^* < \underline{e}_b$, this frontier is parallel but lower than the Pareto frontier $\mathcal{P}(X)$. Thus $\mathcal{P}(X)$ does not intersect $\mathcal{P}(X, X')$. As their subsets, the input-design Pareto frontiers $\mathcal{P}^*(X)$ and $\mathcal{P}^*(X, X')$ also do not intersect. Thus by Lemma A.2, there is uniform worsening of the frontier. In the remaining case we have $\underline{e}_b^* \leq \underline{e}_r^*$, so that (X, X') is b-skewed. Then the unconstrained Pareto frontier $\mathcal{P}(X, X')$ is now a *vertical* line segment with $e_r = \underline{e}_r^*$. The points on this frontier have varying e_b, but any of the e_b does not exceed $\underline{e}_r^*$ because these points are below the 45-degree line. Because $\underline{e}_r^* \leq \underline{e}_r < \underline{e}_b$, we thus know that any point on the frontier $\mathcal{P}(X, X')$ has strictly smaller e_b compared to any point on $\mathcal{P}(X)$. Once again these two unconstrained frontiers do not intersect, and nor do the input-design frontiers. This proves Proposition 3 when X is r-skewed.

A symmetric argument applies when X is b-skewed, so below we focus on the case where X is group-balanced. That is, $R_X = B_X$ lies on the 45-degree line. In this case the Pareto frontiers $\mathcal{P}(X)$ and $\mathcal{P}^*(X)$ are both this singleton point. If X' is not decision-relevant over X for group b, then Lemma A.3 tells us that $\underline{e}_b^* = \underline{e}_b = \underline{e}_r \geq \underline{e}_r^*$. When equality holds the Pareto frontiers $\mathcal{P}(X, X')$ and $\mathcal{P}^*(X, X')$ are also the singleton point $R_X = B_X$, and uniform worsening does not occur. If we instead have strict inequality $\underline{e}_b^* = \underline{e}_b > \underline{e}_r^*$, then (X, X') is r-skewed and the unconstrained Pareto frontier $\mathcal{P}(X, X')$ is a horizontal line segment with one of the endpoints being $F_{X,X'} = R_X = B_X$. Thus $R_X = B_X$ belongs also to the input-design Pareto frontier $\mathcal{P}^*(X, X')$, showing that $\mathcal{P}^*(X)$ and $\mathcal{P}^*(X, X')$ intersect. Uniform worsening of the frontier does not occur either way.

Conversely, suppose X' is decision-relevant over X for both groups. Then by Proposition 1, the unconstrained frontier $\mathcal{P}(X, X')$ is either a horizontal line segment with $e_b = \underline{e}_b^* < \underline{e}_b = \underline{e}_b$, or a vertical line segment with $e_r = \underline{e}_r^* < \underline{e}_r = \underline{e}_b$. Either way the point $R_X = B_X$ does not belong to this frontier, showing that $\mathcal{P}(X)$ does not intersect with $\mathcal{P}(X, X')$. Hence $\mathcal{P}^*(X)$ and $\mathcal{P}^*(X, X')$ also do not intersect, and by Lemma A.2 we know that there is uniform worsening of the frontier. This completes the entire proof of Proposition 3.

B Additional Material

B.1 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 B.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 B.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 B.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. $\square$

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 B.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. □

B.2 Supplementary Material for Section 3

We now apply the general characterization demonstrated in Section 3 to derive more specific results for cases where X satisfies additional structure. We first consider a conditional independence condition that nests the case in which X reveals G:

Assumption 1 (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 once the algorithm has conditioned on X, there is no additional predictive value to knowing the group's identity.

We first characterize the Pareto set under conditional independence. We say the Pareto set $\mathcal{P}(X)$ is increasing if for any $e, e' \in \mathcal{P}(X)$, $e_r > e'_r$ implies $e_b \geq e'_b$.

Proposition B.3. *Suppose Assumption 1 holds. Then $\mathcal{P}(X)$ is increasing from $B_X = R_X$ to F_X. Moreover, it is convex (concave) in e_r if and only if $e_r \leq e_b$ ($e_r \geq e_b$) at $B_X = R_X$.*

Proof. Recall that

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

where

$$E(x) = \left\{ \alpha \left(\sum_y \frac{x_{y,r}}{p_r} \ell(1, y) \sum_y \frac{x_{y,b}}{p_b} \ell(1, y) \right.\right.$$
$$\left.\left. + (1 - \alpha) \left(\sum_y \frac{x_{y,r}}{p_r} \ell(0, y), \sum_y \frac{x_{y,b}}{p_b} \ell(0, y) \right) \ : \ \alpha \in [0, 1] \right\}$$

Under Assumption 3, $x_{y,g} = x_y x_g$ so we have

$$E(x) = \left\{ \left(\alpha \sum_y x_y\, \ell(1, y) + (1 - \alpha) \sum_y x_y \ell(0, y) \right) \left(\frac{x_r}{p_r}, \frac{x_b}{p_b} \right) \ : \ \alpha \in [0, 1] \right\}$$

Since $\frac{x_g}{p_g} \geq 0$, every $E(x)$ has a positive slope so $\mathcal{P}(X)$ must be increasing. Note that since,

$$\arg \min_{e \in \mathcal{E}(X)} e_g = \sum_{x \in \mathcal{X}} \left(\arg \min_{e \in E(x)} e_g \right) p_x$$

we have $B_X = R_X$. Convexity (concavity) follows from the fact that $\mathcal{E}(X)$ is a convex set. $\square$

An immediate corollary is the following:

Corollary 4. *Suppose Assumption 1 holds. Then every pair of distinct points $e, e' \in \mathcal{P}(X)$ represents a strong fairness-accuracy conflict.*

Proof. 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 by the definition of B_X, 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. $\square$

An example Pareto frontier for a covariate satisfying Conditional Independence is depicted in Figure 9. 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. 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.[29]

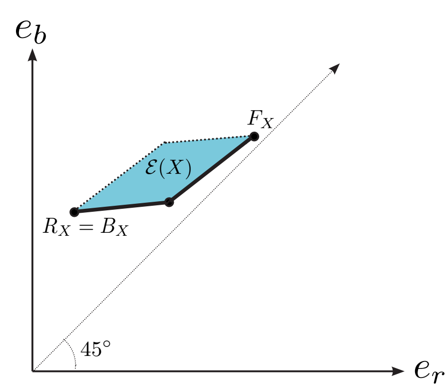


Figure 9: Depiction of the Pareto frontier under assumption of conditional independence of G and Y.

Since $R_X = B_X$ under conditional independence, that means that the only case of group-balance is when both groups have equal errors at $R_X = B_X = F_X$. Thus, under Assumption 3, X is group-balanced if and only if the Pareto set is a singleton on the 45 degree line. Barring this special case, X is going to be group-skewed and there will be a strong fariness-accurarcy conflict as per the corollary above.

Conditional independence allows for a clear relationship between statistical properties of X and the shape of the Pareto frontier. Suppose the designer has more information about G. Note that if X' is only informative about G, then (X, X') is also conditionally independent.

Finally, we can further strengthen conditional independence to the following strong independence condition:

[29]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. (Corollary 4 is vacuous in this case, since there are no two distinct points on the Pareto frontier.)

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

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

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 10.

Proposition B.4. *Suppose Assumption 2 holds. Then the Pareto frontier is a single point on the 45-degree line.*

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

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

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 $\left(\sum_y \frac{x_{y,r}}{p_r} \ell(1, y), \sum_y \frac{x_{y,b}}{p_b} \ell(1, y)\right)$ and $\left(\sum_y \frac{x_{y,r}}{p_r} \ell(0, y), \sum_y \frac{x_{y,b}}{p_b} \ell(0, y)\right)$, lies on the 45-degree line. Therefore $\mathcal{E}(X) = \sum_x E(x) \cdot p_x$ is also 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.

B.3 Supplementary Material for Section 5

B.3.1 Adversarial Agents

We now consider the problem outlined in Section 5, when one of the weights α_r, α_b is negative.[30] Without loss, let $\alpha_r > 0 > \alpha_b$, reflecting an adversarial agent who prefers for group b's error to be higher. The first half of Lemma 1 extends fully.

[30]It is straightforward also to consider the case where both weights are negative, but we do not consider this setting to be practically relevant.

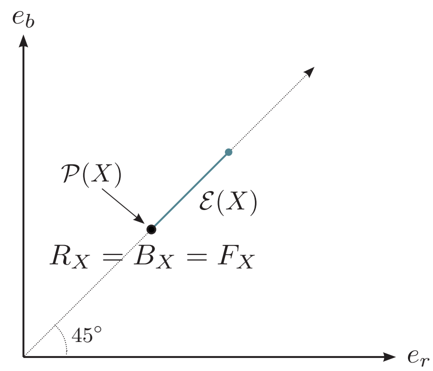


Figure 10: Depiction of the Pareto frontier under assumption of strong independence

Lemma B.1. *For every covariate X, $\mathcal{E}^*(X) = \mathcal{E}(X) \cap H$.*

But the analogous equivalence for the Pareto frontier does not extend. Instead, similar to the development of R_X, B_X, and F_X, define

$$G^*_X \equiv \underset{(e_r,e_b)\in\mathcal{E}^*(X)}{\arg\min}\; e_g$$

to be the feasible point in $\mathcal{E}^*(X)$ that minimizes group g's error (breaking ties by minimizing the other group's error), and 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 (breaking ties by minimizing either group's error).

Definition B.3. Covariate X is:

- *input-design-r-skewed* if $e_r < e_b$ at R^*_X and $e_r \leq e_b$ at B^*_X
- *input-design-b-skewed* if $e_b < e_r$ at B^*_X and $e_b \leq e_r$ at R^*_X
- *input-design-group-balanced* otherwise

The proof for Theorem 1 applies for any compact and convex feasible set, and so directly implies:

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

(a) *R^*_X and B^*_X if X is input-design-group-balanced*

(b) *G^*_X and F^*_X if X is input-design-g-skewed*

We can use this characterization to extend our result from Section 5.2.1.

Definition B.4. X is *strictly input-design-group-balanced* if $e_r < e_b$ at R^*_X and $e_b < e_r$ at B^*_X.

Proposition B.5. *(a) Suppose $|\alpha_r| > |\alpha_b|$. Then excluding G given X always uniformly worsens the frontier.*

(b) Suppose $|\alpha_r| \leq |\alpha_b|$. Then, excluding G given X uniformly worsens the frontier if and only if X is strictly input-design-group-balanced.

This result says that, potentially surprisingly, even if the agent choosing the algorithm has adversarial motives against one of the groups, the designer may still prefer to send information about group identity. Indeed, if the agent is only slightly adversarial—caring more about reducing group r's error than about increasing group b's error—then excluding G given X always uniformly worsens the frontier. When the agent is strongly adversarial—caring less about reducing group r's error than about increasing group b's error—then the condition for a uniform worsening of the frontier is the same as the one outlined in the main text. From the point of view of policy, this result suggests that even in the presence of malicious actors, naively excluding group identity may not be the optimal approach.

We outline the argument here. Consider a strictly input-design-group-balanced X (see Figure 11). Theorem B.1 tells us that the Pareto frontier is that part of the lower boundary of $\mathcal{E}^*(X)$ connecting R^*_X and B^*_X. By Lemma B.1 and the characterization of $\mathcal{E}(X,G)$ in the proof of Lemma 1, the feasible set $\mathcal{E}^*(X,G)$ is the intersection of H and the smallest box containing $\mathcal{E}^*(X)$. Again applying Theorem B.1, the Pareto frontier $\mathcal{P}^*(X,G)$ connects $R^*_{X,G} = B^*_{X,G}$ to $F^*_{X,G}$. This frontier is nowhere overlapping with the original frontier, so it follows that excluding G given X uniformly worsens the frontier.

Now suppose X is input-design-r-skewed X (see Figure 12). There are two possibilities. If $|\alpha_r| > |\alpha_b|$ (Panel (a)), then B^*_X and F^*_X are identical, so the Pareto frontier is that part of the lower boundary of $\mathcal{E}^*(X)$ connecting R^*_X to $B^*_X = F^*_X$. The Pareto frontier $\mathcal{P}^*(X,G)$ consists of the line segment connecting $R^*_{X,G} = B^*_{X,G}$ to $F^*_{X,G}$ (which may be degenerate). This frontier is nowhere overlapping with the previous, so again, excluding G given X uniformly worsens the frontier. On the other hand, if $|\alpha_r| \leq |\alpha_b|$ (Panel (b)), then B^*_X and F^*_X are different, and the line segment from B^*_X to F^*_X belongs to both $\mathcal{P}^*(X)$ and also $\mathcal{P}^*(X,G)$. Thus, excluding G given X does not uniformly worsen the frontier.

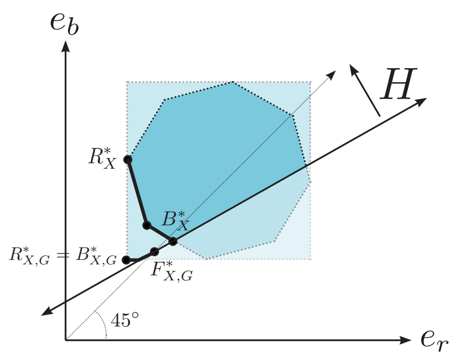


Figure 11: X is strictly input-design-group-balanced.

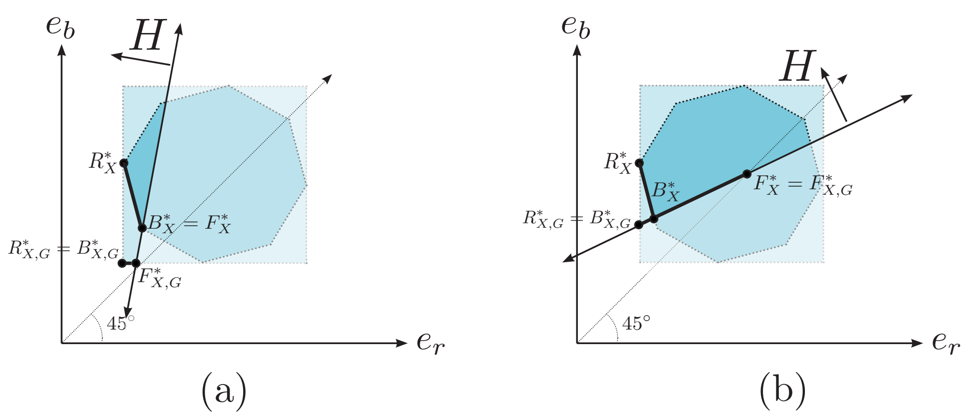


Figure 12: X is r-skewed. Panel (a): $|\alpha_r| > |\alpha_b|$; Panel (b): $|\alpha_r| \leq |\alpha_b|$.

B.3.2 Result for Excluding X' Given Group-Balanced X

Our results in the main text assume that group identity is revealed by (X, X'), allowing us to exploit the special structure of the Pareto frontier when G is revealed. We now provide a sufficient condition for when excluding X' given X uniformly worsens the frontier, without assuming that G is revealed.

Definition B.5. 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 B.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 excluding X' given X uniformly worsens the frontier.*

Proof. 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 pair 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 group errors 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. □

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.

B.4 Counterexample to Lemma 1 When $|A| > 2$

This example demonstrates that Lemma 1 can fail if there are more than two actions, i.e., there may exist a point (e_r, e_b) satisfying $p_r e_r + p_b e_b \leq e_0$, which cannot be implemented under input design.

Let $X = \{x_1, x_2, x_3\}$, $Y = \{y_1, y_2, y_3\}$, $G = \{r, b\}$, and $A = \{1, 2, 3\}$. Suppose that $p_r = 0.2$ and $p_b = 0.8$, and the conditional distribution of (X, Y) given G is:

	$X = x_1$	$X = x_2$	$X = x_3$
$Y = y_1$	0	0.07	0.006
$Y = y_2$	0	0.06	0
$Y = y_3$	0.1	0	0.764

$G = r$

	$X = x_1$	$X = x_2$	$X = x_3$
$Y = y_1$	0	0.08	0.1
$Y = y_2$	0.05	0.06	0.04
$Y = y_3$	0.1	0.1	0.47

$G = b$

The loss function is described by the following matrix L, whose (i, j)-th entry is the loss to choosing action i when the true type is y_j:

$$L = \begin{pmatrix} 5.2 & 5.88 & 5.02 \\ 5.12 & 6.42 & 4.91 \\ 4.16 & 6.84 & 5.06 \end{pmatrix}$$

The best action under the prior is $a = 2$ and yields an aggregate error of approximately 5.15. Define f and $\tilde{f}$ to be deterministic algorithms satisfying

$$f(x_1) = 1 \quad f(x_2) = 2 \quad f(x_3) = 3$$

$$\tilde{f}(x_1) = 1 \quad \tilde{f}(x_2) = 3 \quad \tilde{f}(x_3) = 2$$

Then $e \equiv e(f) \approx (5.14, 5.15)$ while $\tilde{e} \equiv e(\tilde{f}) \approx (4.99, 5.12)$. Both points yield aggregate errors that improve upon the prior, and hence satisfy the condition of Lemma 1. Nevertheless, the latter point cannot be implemented by input design. The reason is as follows: e and $\tilde{e}$ are extreme points of the feasible set $\mathcal{E}(X)$. Thus, if they can be implemented, then they can be implemented by a direct mechanism that suggests play of the action. Suppose towards contradiction that the designer can implement e. By measurability of the actions of the decision-maker, the decision-maker must be able to distinguish between x_1, x_2, x_3. This means that the decision-maker has enough information to choose the algorithm $\tilde{f}$, achieving the error pair $\tilde{e}$ which is strictly better. Observing that e and $\tilde{e}$ belong to $\mathcal{P}(X)$ completes the argument.

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," .

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*.

Garg, N., H. Li, and F. Monachou (2021): "Dropping Standardized Testing for Admissions Trades Off Information and Access," Working Paper.

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., 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.

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.