Algorithm Design: Fairness and Accuracy*

Annie Liang† Jay Lu‡ Xiaosheng Mu§

August 16, 2022

Abstract

Algorithms are widely used to guide high-stakes decisions, from medical recommendations to loan approvals. Designers are increasingly optimizing not only for accuracy but also "fairness" i.e. how much accuracy varies across different subgroups. We define and characterize a *fairness-accuracy frontier*, consisting of the optimal points across a broad range of criteria for trading off fairness and accuracy. Our results identify how the algorithm's inputs govern the shape of this frontier, showing (for example) that fairness considerations matter for the frontier precisely when the inputs fail a condition we call group-balance. We next study an information design problem where the designer controls inputs (e.g., by legally banning an input) but the algorithm itself is chosen by another agent. We show that all designers strictly prefer to allow access to group identity if and only if the other inputs satisfy group-balance.

1 Introduction

Decisions such as who should be targeted for high-risk healthcare management or who should receive a mortgage are increasingly guided by the predictions of algorithms (Roth and Kearns, 2019). A recent literature suggests that the error rates of commercially-deployed algorithms often differ substantially across racial or gender groups (Arnold et al., 2021; Fuster et al., 2021); for example, patients assigned to the same risk score by a widely-used healthcare

*We thank Nageeb Ali, Simon Board, Krishna Dasaratha, Will Dobbie, Sergiu Hart, Peter Hull, Navin Kartik, Yair Livne, Sendhil Mullainathan, and Derek Neal 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

algorithm were shown to have substantially different actual health risks depending on their race (Obermeyer et al., 2019), and the false positive rate of an algorithm used to predict criminal reoffense was shown to be twice as high for Black defendants as for White defendants (Angwin and Larson, 2016). These findings have led to widespread revision of algorithms' objective functions to depend not only on the aggregate accuracy of the algorithm's predictions, but also on its "fairness", defined as how much the algorithm's accuracy differs across groups. These two criteria are not always in conflict (for example, if the two groups are statistically similar) but in many situations of interest they may be.

Our goal in this paper is to understand the nature of the tradeoff between fairness and accuracy, and how this tradeoff depends on the inputs to the algorithm. The tradeoff is formalized as a *fairness-accuracy frontier* that consists of the optimal points (given a fixed set of inputs) across a broad class of objective functions, ranging from the traditional criterion of maximizing accuracy, to a criterion that enforces equal error rates across groups. Besides our basic theoretical interest in the shape of this frontier, we are additionally motivated by what it can tell us about the regulatory practice of banning certain algorithmic inputs (e.g., group identity). Conceptualizing the fairness-accuracy frontier allows us to demonstrate results that hold uniformly across the wide class of objective functions, and thus does not depend on the particular details of the designer's preferences over fairness and accuracy.

In our model, a designer chooses an algorithm that takes observed covariates as inputs (e.g., image scans, lab tests, history of hospital visits) and outputs a decision (e.g., whether to recommend a risky medical procedure). The algorithm's consequences for any given individual are measured using a loss function, which can be interpreted as the inaccuracy or the harm of the decision. We aggregate losses within two pre-defined groups, group r (red) and group b (blue). Each group's *error* is the expected loss for individuals of that group. An algorithm is more accurate if it implies lower errors for both groups, and more fair if it implies a smaller difference between the two groups' errors.

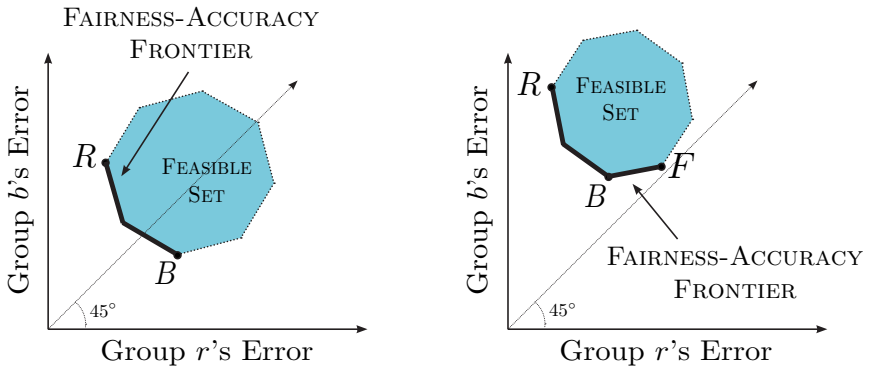
To understand the tradeoff between fairness and accuracy, we define the class of *fairness-accuracy (FA) preferences* to be all preferences over group error pairs that are consistent with the following order: one pair of group errors *FA-dominates* another if the former involves smaller errors for both groups (greater accuracy) and also a smaller difference between group errors (greater fairness).[1] This partial order is consistent with a broad range of designer preferences, including 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). Some of these

[1]We do not take a stance on the normative desirability of these preferences, instead interpreting our class as encompassing the broad range of designer preferences that could be relevant in practice.

preferences correspond directly to optimization problems that have been proposed for use in practice. For example, a Rawlsian designer is equivalent to someone who uses group distributionally robust optimization (Sagawa et al., 2020), and an Egalitarian designer is equivalent (on a restricted domain) to someone who maximizes accuracy subject to equality of error rates (as considered in Hardt et al. (2016) among others). We define the *fairness-accuracy frontier* to be the set of all feasible group error pairs that are FA-undominated within the feasible set, i.e., there is no feasible error pair that improves simultaneously on accuracy and fairness.

A simple property of the algorithm's inputs turns out to be critical for determining the shape of the fairness-accuracy frontier. Say that a covariate vector is *group-balanced* if given these inputs, group r's optimal algorithm (i.e. the one that gives r the smallest error over all feasible algorithms) yields a lower error for group r than for group b, and if the reverse is true for group b's optimal algorithm. Otherwise, we say the covariate vector is *group-skewed*. While we expect that this condition is simple to check on data, it is difficult to anticipate whether group-balance or group-skew is more typical in practice. One force that pushes towards group-balance is if the covariate has opposite implications for the two groups: for example, if frequent address changes signal higher creditworthiness for high-income borrowers but lower creditworthiness for low-income borrowers, then the algorithm (based on this covariate) that maximizes accuracy for the high-income group will lead to a lower error for the high-income group, and vice versa. On the other hand, if the covariate has the same implications for both groups but is measured more accurately for one group than the other—say, if medical data is recorded more accurately for high-income patients than low-income patients—then even the best algorithm for one group may still result in a lower error for the other.

Our first result says that depending on whether the covariate vector is group-balanced or group-skewed, the fairness-accuracy frontier takes either of two possible forms, as depicted in Figure 1. In both cases, the frontier is a part of the lower boundary of the *feasible set*, namely the error pairs that are implementable using some algorithm that takes the covariate vector as input. But in the case of group-balanced inputs, the frontier is the part of the lower boundary that 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). This is precisely the usual Pareto frontier, i.e., all error pairs that cannot be simultaneously reduced in both coordinates. In the case of group-skewed inputs, the frontier again includes the usual Pareto frontier, but now additionally includes a positively-sloped part (in Figure 1, the segment from B to the fairness-maximizing point F) along which both groups' errors increase but the gap between their errors decreases. This characterization of the frontier tells us that a policy proposal



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

Figure 1: The Fairness-Accuracy Frontier.

that increases errors for both groups, but reduces the gap between group errors, can only be justified by fairness considerations if the covariate vector is group-skewed. If instead the covariate vector is group-balanced, then regardless of how much the designer values fairness, such a policy proposal would not be justified as the optimal point must be on the usual Pareto frontier.

We next consider the important special case where group identity is an input to the algorithm. We show that the feasible set and frontier simplify as depicted in Figure 2: The feasible set is a rectangle, and the fairness-accuracy frontier is a single line segment along which the disadvantaged group (i.e., the group with the higher error) receives its minimal feasible error. A corollary of this characterization is that access to group identity must reduce the disadvantaged group's error regardless of the designer's fairness-accuracy preferences.

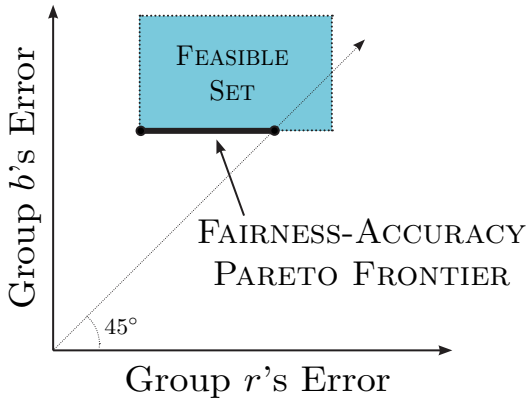


Figure 2: Depiction of the fairness-accuracy frontier in the case where X reveals G.

In the second half of the paper, we investigate what happens if the designer does not choose the algorithm, but instead regulates the inputs of the algorithm. This question is motivated by settings where 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 restricts the inputs available to the algorithm, for example, by legally banning the use of a specific input.

We model this as an information design problem (Kamenica and Gentzkow, 2011) where the designer chooses a garbling of the available inputs, and an agent chooses an algorithm (based on the garbling) to maximize accuracy. Under weak conditions, it turns out to be without loss for the designer to only control the algorithm's inputs. That is, any error pair that a designer would choose to implement given full control of the algorithm can also be achieved by appropriately garbling the inputs.

We next consider whether the optimal garbling might involve excluding a covariate entirely from use by the agent in the algorithm. We demonstrate two results: First, excluding group identity as an algorithmic input is strictly welfare-reducing for all designers (with FA preferences) if and only if the permitted covariates are group-balanced. Thus, although conditioning on group identity is unfair in terms of disparate *treatment* (i.e., whether the policy discriminates based on group identity), it may be necessary to ensure fairness in terms of disparate *impact* (i.e., whether the adverse effects of the policy are disproportionately borne by a specific group).

Second, when group identity is permitted as an input, then completely excluding any other covariate makes every designer strictly worse off, so long as that covariate satisfies a minimally informative condition we call "decision-relevance." Decision-relevance does not depend on whether the covariate is biased towards either group. When applied to the policy question of whether to permit standardized test scores in admissions decisions, our result suggests that so long as group identities are permissible inputs for admission decisions (as is the case in most states in the US), then excluding test scores is welfare-reducing for all designers—regardless of how biased the score may be against either group. 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[2]), then the optimal garbling of covariates for some designer preference may indeed involve completely excluding that covariate, and we provide an example to this effect.

[2]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."

1.1 Related Literature

A large literature in computer science has emerged recently around the topic of algorithmic fairness (see Kleinberg et al. (2018) and Roth and Kearns (2019) for overviews). We follow this literature in focusing on group-based statistical measures of fairness (Hardt et al., 2016; Kleinberg et al., 2017; Chouldechova, 2017), which compare the error rates of the algorithm across pre-defined groups (see Appendix A for examples).[3] Such fairness metrics are also focal in recent papers in empirical economics that investigate the disparate impact of algorithms (Obermeyer et al., 2019; Arnold et al., 2021; Fuster et al., 2021), and in the accompanying discussions on algorithmic discrimination in policy (Angwin and Larson, 2016).

A central goal in the computer science literature is developing algorithms that optimize for accuracy subject to fairness constraints (Hardt et al., 2016; Diana et al., 2021), and characterizing their computational and robustness properties. Our work differs from this literature in two primary ways: First, rather than focusing on a specific optimization problem,[4] we formulate a framework with minimal assumptions on how the designer prefers to trade off fairness and accuracy.[5] This allows us to identify novel statistical properties of covariate vectors (such as group-balance) that are important not because of their implications for any particular optimization problem, but because they identify properties shared across the solutions of different optimization problems.

Second, our analysis in Section 4 introduces an information design approach (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019), in which the designer can flexibly regulate the inputs to the algorithm but cannot choose the algorithm itself. We view selection of garblings of inputs as an effective policy tool,[6] which can be committed to legally,[7] and deserves further attention within the context of algorithmic fairness. Relative to the information design literature, our focus is on a frontier of solutions with respect to a class of different Sender preferences, rather than on the optimal solution for a specific Sender preference. Additionally, the Sender in our framework cannot choose a completely flexible

[3]Notable exceptions include individual fairness (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).

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

[5]Several authors have pointed to a fairness-accuracy frontier as a useful conceptual tool (Roth and Kearns, 2019), and others have estimated this frontier for specific data sets (Wei and Niethammer, 2020) or provided computationally efficient approaches for deriving this frontier Chohlas-Wood et al. (2021). Our work provides general theoretical results for how this frontier depends on statistical properties of the algorithm's inputs.

[6]In contrast, for example, Rambachan et al. (2021) studies a screening model where a designer chooses from the set of information policies that directly reveal a subset of inputs.

[7]See for example Yang and Dobbie (2020), which summarizes the extant law and proposes new legal policies for mitigating algorithmic bias.

information structure, but is constrained to garblings of a primitive covariate vector.[8]

Finally, our work connects to earlier literatures regarding fairness in economics. For example, our class of FA preferences relates to the literature on social preferences, including Fehr and Schmidt (1999)'s model of inequity aversion in games—in which players place negative weight on the absolute difference between their payoffs—and Grant et al. (2010)'s generalization of utilitarianism that allows for non-linear aggregation of individual payoffs (to capture fairness considerations). Our framework also relates to the literature on statistical discrimination (see Fang and Moro (2011) for a survey), although we focus on practical questions regarding algorithmic regulation, rather than on questions regarding why inequality emerges and persists in equilibrium.

2 Framework

2.1 Setup and Notation

There is a population of individuals, where each individual is described by a *covariate vector* X taking values in the finite set $\mathcal{X}$, a *type* Y taking values in the finite set $\mathcal{Y}$,[9] and a *group identity* G taking values r or b.[10] Throughout we think of G, X, Y as random variables with joint distribution $\mathbb{P}$, and use $p_g \equiv \mathbb{P}(G = g) > 0$ to denote the fraction of the population that belongs to group $g \in \{r, b\}$. We impose no assumptions on the joint distribution,[11] permitting for example each of the following:

Example 1 (X reveals or closely proxies for G). The group identity may be an input in the covariate vector X, or predictable from inputs in the covariate vector X. For example, Bertrand and Kamenica (2020) show that data on consumption patterns permits near perfect classification of gender and a fairly accurate prediction of other group identities such as income bracket, race, and political ideology.

[8] Fairness considerations additionally introduce non-linearities that complicate the Sender's objective function. 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.

[9] We make the finite assumption to simplify various notations in the exposition. Most of our results generalize to infinite covariate values and/or infinite types.

[10] 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.

[11] We view $\mathbb{P}$ as the population distribution on which the algorithm is both trained and tested. An interesting direction for future work would be to permit the data that the algorithm is trained on to differ in distribution from the data on which the algorithm's errors are evaluated. For example, the data on which the algorithm is trained may reflect historical biases that are no longer descriptive of the current environment. Another interesting direction would be to study optimal sampling of data on which to train the algorithm (in which case $\mathbb{P}$ is endogenous).

Example 2 (Biased Covariates). The value of an input in X may be systematically biased depending on group identity. For example, if G is income bracket, Y is ability, and X is a test score that can be improved through better access to test prep, the distribution $\mathbb{P}$ may have the property that at every ability level, the conditional distribution of test scores is shifted higher for students in the high-income bracket (i.e., the distribution of $X \mid Y = y, G = r$ first-order stochastically dominates $X \mid Y = y, G = b$ at every $y \in \mathcal{Y}$).

Example 3 (Asymmetrically Informative Covariates). The inputs in X may be more informative about Y for one group than the other. This could be because the covariate is intrinsically more informative for one of the groups—for example, if G is biological sex, X is whether the individual is pregnant, and Y is whether the individual would benefit from treatment using nonsteroidal anti-inflammatory drugs (NSAIDs) such as ibuprofen[12], then X is informative for females and not for males. But asymmetric informativeness can also arise through selective reporting of the covariate or group-dependent measurement error. For example, suppose X consists of test results that are affordable only if covered by health insurance. In this case, we may expect that the test result is recorded more frequently for high-income patients than for low-income patients.

A designer chooses an *algorithm* $a : \mathcal{X} \to \Delta(\mathcal{D})$ that maps covariate vectors into distributions over decisions in $\mathcal{D} = \{0, 1\}$. Let $\mathscr{A}_X$ denote the set of all algorithms. Some motivating examples of types, group identities, covariate vectors, and decisions are given below:

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

Credit scoring. Y is creditworthiness, G is gender, and the decision 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.

Bail. Y is whether an individual is high-risk or low-risk of criminal reoffense, G is race, and the decision 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, frequency of moves, or drug use as a child.[13]

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

[12]On October 15, 2020, the FDA warned that the use of NSAIDs for those who are pregnant around 20 weeks or later can cause serious kidney problems in the unborn baby.

[13]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`.

The consequence of choosing decision d for an individual whose true type is y is evaluated using a (potentially group-dependent) loss function $\ell : \mathcal{D} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$.[14] We further aggregate these losses across individuals within each group:

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

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

That is, group g's error is the average loss for members of group g. For example, if the type Y belongs to $\{0, 1\}$ and $\ell(d, y, g) = \mathbb{1}(d \neq y)$, then $e_g(a)$ is the probability of a type I or type II error. We view the choice of the right loss function as application-specific, and demonstrate results that hold for arbitrary ℓ.

2.2 Fairness-Accuracy Preferences

Different algorithms imply different group errors. Besides an interest in improving accuracy, i.e., reducing errors for both groups, there is also increasing interest in the "fairness" of the algorithm, as measured by the difference between the two group errors. Of the various fairness criteria that have been proposed in the literature (see Mehrabi et al. (2022) for a recent survey), many can be accommodated in our framework under a particular choice of a loss function. For example, if the type Y belongs to $\{0, 1\}$ and $\ell(d, y, g) = \mathbb{1}(d \neq y)$, then $e_r(g) = e_b(g)$ corresponds to equality of misclassification rates. If $\ell(d, y, g) = \mathbb{1}(d = 1, y = 0)$ then $e_r(g) = e_b(g)$ corresponds to equality of false positive rates (Kleinberg et al., 2017; Chouldechova, 2017). And if

$$\ell(d, y, g) = \begin{cases} \frac{P(Y=y)}{P(Y=y|G=g)} & \text{if } d = 1 \\ 0 & \text{otherwise} \end{cases}$$

then $e_r(g) = e_b(g)$ corresponds to equality of equalized odds (Hardt et al., 2016). See Appendix A for further details.

We consider the broad class of possible designer preferences over these pairs of group errors $e = (e_r, e_g) \in \mathbb{R}^2$, for which simultaneously increasing accuracy (reducing errors for both groups) and also increasing fairness (reducing the gap between these errors) is necessarily an improvement.[15]

[14] For example, if G is socioeconomic background, Y is creditworthiness, the decision is whether to grant a loan, and the loss function corresponds to financial cost, then a bank manager may experience greater losses from predicting creditworthiness incorrectly for the wealthy group.

[15] We follow the large and growing literature on algorithmic fairness (see Section 1.1) by defining fairness as a difference in group errors. This formulation does not take into account all important fairness considerations.

Definition 2. The *fairness-accuracy (FA) dominance* relation $>_{FA}$ is the partial order on $\mathbb{R}^2$ satisfying $(e_r, e_b) >_{FA} (e'_r, e'_b)$ 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.[16]

Definition 3. A *fairness-accuracy (FA) preference* $\succeq$ is any total order on $\mathbb{R}^2$ such that $e \succ e'$ whenever $e >_{FA} e'$.

It is straightforward to see that these orders are unchanged if $|e_r - e_b|$ is replaced with $\phi(|e_r - e_b|)$ where ϕ is a strictly increasing function. In Section 5, we discuss an extension of the fairness criterion to any $|\phi(e_r) - \phi(e_b)|$ where ϕ is continuous and strictly increasing, which includes the ratio of error rates as a special case (setting $\phi(e) = \log(e)$). We also discuss in Section 5 an extension of the FA dominance relation when fairness and accuracy are evaluated using different loss functions.

We interpret the class of FA preferences as encompassing the broad range of views on how to trade off fairness and accuracy that could be relevant in practice, including the following special cases that have been proposed in the preceding literature.

Example 4 (Utilitarian/Bayes-Optimal). 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.) A designer with preferences $\succeq_u$ is called *Utilitarian* (Harsanyi, 1953, 1955).

Example 5 (Rawlsian/Group-DRO). 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. This is also known as *group distributionally robust optimization* (Sagawa et al., 2020; Hansen et al., 2022). A designer with preferences $\succeq_r$ is called *Rawlsian* (Rawls, 1971).

For example, perfect prediction of criminal offense (Y) by the algorithm for both groups does not address historical inequities that have shaped differential base rates of Y across groups. Moreover, 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 leave to future work the interesting question of how these algorithmic design decisions might impact decisions in a larger dynamic game.

[16] This relation is similar to the property of a *strict improvement* in Kleinberg and Mullainathan (2019), where an algorithm is a strict improvement if it improves both efficiency and equity. Their definitions of efficiency and equity are based on two loss functions that are different from one another, so this corresponds to a special case of our FA-dominance relation in Online Appendix O.1.

Example 6 (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. Note that this includes a designer who solves the constrained optimization problem

$$\min_{a \in \mathscr{A}_X} p_r e_r(a) + p_b e_b(a) \text{ s.t. } e_r(a) = e_b(a)$$

when equal error rates are feasible, as used for example in Hardt et al. (2016).[17] A designer with preferences $\succeq_e$ is called *Egalitarian* (Parfit, 2002).

Example 7 (Constrained Optimization). The designer evaluates errors $e = (e_r, e_b)$ according to

$$w_c(e) = (1 - \lambda) w_u(e) + \lambda w_e(e)$$

for some $\lambda \in [0, 1]$ (using $\succeq_e$ to break ties for $\lambda = 1$). The optimal choices here correspond to the solutions of the following constrained optimization problem

$$\min_{a \in \mathscr{A}_X} p_r e_r(a) + p_b e_b(a) \text{ s.t. } |e_r(a) - e_b(a)| \leq c$$

when they exist, as used for example in Ferry et al. (2022).[18]

Fixing any covariate vector X, we define the feasible set of group error pairs to be those that can be implemented by some algorithm that takes X as input. The fairness-accuracy frontier is the set of all group error pairs that are FA-undominated in the feasible set.

Definition 4. The *feasible set* given covariate vector X is

$$\mathcal{E}(X) \equiv \{(e_r(a), e_b(a)) : a \in \mathscr{A}_X\}$$

Definition 5. The *fairness-accuracy (FA) frontier* given X, denoted $\mathscr{F}(X)$, is the set of all error pairs $e \in \mathcal{E}(X)$ that are FA-undominated, i.e. there does not exist an error pair $e' \in \mathcal{E}(X)$ satisfying $e' >_{FA} e$.

[17] The optimization problem considered in Hardt et al. (2016) includes this case but is more general, as they allow for the loss function used to evaluate accuracy to differ from the loss function used to evaluate fairness. See Section 5 for an extension of our approach for different loss functions.

[18] The constant in w_c is simply the Lagrangian for such an optimization problem where λ corresponds to the multiplier. Note that while the preference $\succeq_c$ is complete, the constrained optimization yields an incomplete ordering e.g. two errors that are both not feasible are not comparable.

The FA frontier consists of all the points that are optimal under some FA preference. Furthermore, it is minimal in the sense that for every point in the FA frontier, there exists some FA preference such that the point is uniquely optimal, so we cannot exclude any points without hurting some designer. We discuss these alternate characterizations in Appendix O.5.

3 The Fairness-Accuracy Frontier

In Section 3.1, we define the property of *group-balance* that will play a key role in several of our results. In Section 3.2, we characterize the frontier and its implications for the kinds of fairness-accuracy tradeoffs that emerge. In Section 3.3, we provide further results for the important special case where group identity is an input in the algorithm.

3.1 Key Property: Group-Balance

For all covariate vectors X, the feasible set $\mathcal{E}(X)$ is closed and convex (Lemma B.1). Certain important feasible points include:

Definition 6 (Group Optimal Points). For any covariate vector X, define

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

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

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

to be the feasible point that minimizes group b's error. In both cases, if the minimizer is not unique, we break ties by choosing the point that minimizes the other group's error. We use G_X to 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 decision for group r for each realization of X (breaking ties in favor of group b).[19] R_X is then the error pair resulting from this algorithm.

Definition 7 (Fairness Optimal Point). For any covariate vector X, define

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

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

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.[20]

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 X is a binary score where the conditional distribution $(X, Y) \mid G$ is described by:

	$X=0$	$X=1$
$Y=0$	3/8	1/8
$Y=1$	1/8	3/8

$G = r$

	$X=0$	$X=1$
$Y=0$	1/3	1/6
$Y=1$	1/6	1/3

$G = b$

Let the loss function ℓ be the misclassification rate; that is, $\ell(d, y, g) = \mathbb{1}(d \neq y)$. Then the b-optimal point B_X is achieved by the algorithm that maps $X = 1$ to $d = 1$ and $X = 0$ to $d = 0$, which leads to a *higher* error for group b than group r (compare 1/3 to 1/4). Thus, using X to minimize errors for group b results in a higher error for group b than group r. The property of group-balance rules this out.

Definition 8. Covariate vector 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 either group g, then we say it is *group-skewed*.

In words, X is r-skewed if group r's error is smaller than group b's error not only at the r-optimal point R_X, but also at the b-optimal point B_X. Geometrically, this means that R_X and B_X fall to the same side of the 45 degree line. In contrast, the covariate vector X is group-balanced if at each group's optimal point, its error is lower than that of the other group, implying that R_X and B_X fall to opposite sides of the 45 degree line. In practice, one reason that covariate vectors may be group-skewed (and hence, generate strong fairness-accuracy conflicts) is if they are recorded or measured more accurately for one group than another (see Example 3).

3.2 Characterization of the Frontier

Depending on whether the covariate vector X is group-balanced or group-skewed, the fairness-accuracy frontier $\mathscr{F}(X)$ takes either of two forms. In the result below, we use *lower boundary*

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

between two points to mean the part of the boundary of the set that lies between the two points and below the line segment connecting the two.

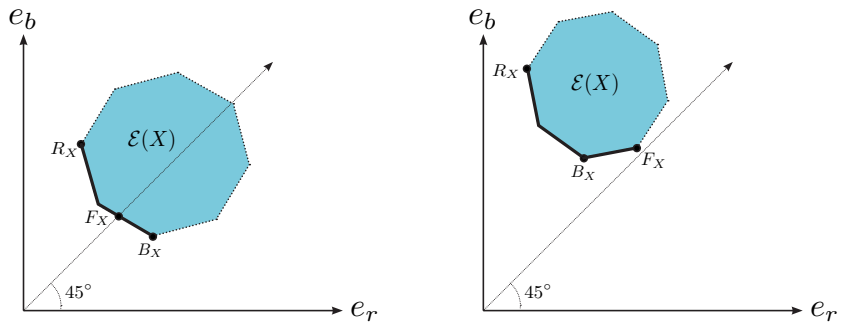
Theorem 1. *The fairness-accuracy frontier $\mathscr{F}(X)$ is the lower boundary of the feasible set $\mathcal{E}(X)$ between*

(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, the two points fall on opposite sides of the 45-degree line (Panel (a)), and the fairness-accuracy frontier is that part of the lower boundary of the feasible set connecting these two points. This corresponds precisely to the usual Pareto frontier of the feasible set, i.e. the set of all points (e_r, e_b) such that no other feasible point (e'_r, e'_b) is component-wise smaller. When X is r-skewed (Panel (b)), then both R_X and B_X fall on the same side of the 45-degree line, and the fairness-accuracy frontier consists not only of the usual Pareto frontier connecting R_X to B_X, but additionally a positively sloped line segment connecting the Pareto frontier to F_X.[21]

Thus, the usual Pareto frontier and the fairness-accuracy frontier differ if and only if the covariate vector is group-skewed, implying the following corollary.



(a) X is group-balanced (b) X is r-skewed

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

Corollary 1. *Suppose F_X is distinct from R_X and B_X. Then if and only if X is group-skewed, there are points $e, e' \in \mathscr{F}(X)$ satisfying $e_r \leq e'_r$ and $e_b \leq e'_b$ with at least one inequality strict.*

[21] When X is group-skewed, the fairness-optimal point F_X may not lie on the 45 degree line.

This corollary says that if the covariate vector is group-balanced, then no two points on the fairness-accuracy frontier can be Pareto-ranked. Thus, a policy proposal that increases errors for both groups but reduces the gap between group errors, cannot be optimal under any fairness-accuracy preference. On the other hand, if inputs are group-skewed, the frontier has a positively-sloped segment of the frontier along which every pair of points can be Pareto-ranked. Along this segment, the only way to decrease the gap in errors is to increase errors for both groups.

3.3 Group Identity as an Input

In the important case where group identity is an algorithmic input, the feasible set and fairness-accuracy frontier simplify further. (In Section O.4, we generalize our results to the case where G is conditionally independent of Y given X.)

Definition 9. 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 the fairness-accuracy frontier $\mathscr{F}(X)$ is the line segment from $R_X = B_X$ to F_X.*

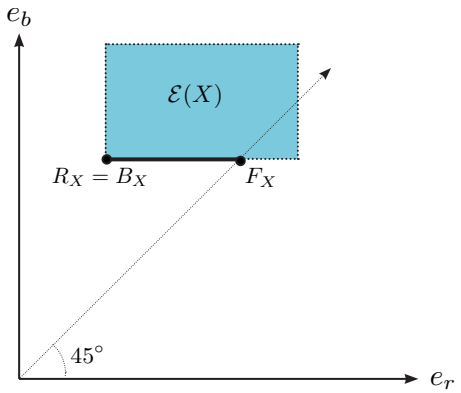


Figure 4: Example feasible set and fairness-accuracy frontier when X reveals G.

An example of such a feasible set and fairness-accuracy frontier are depicted in Figure 4. One endpoint, the Utilitarian-optimal point labeled $R_X = B_X$, gives both groups their minimal feasible error. The other endpoint, the Egalitarian-optimal F_X, maximizes fairness. Everywhere along the fairness-accuracy frontier $\mathscr{F}(X)$, the worse-off (higher error) group

receives its minimal feasible error, so every point on the frontier is optimal for a Rawlsian designer.

Intuitively, when the algorithm is not given access to G, it must assign each realization of X to the same decision for members of both groups, tying together the feasible group errors. When the algorithm is instead given access to group identity, then the designer can set a different rule for each group and thus flexibly adjust one group's error without affecting the other. This gives the feasible set its rectangular shape, and the FA-dominance relation further pins down the fairness accuracy frontier to be a part of the edge that minimizes errors for the group with the higher error.

Indeed, we can say more about the implications for this group. Formally, for any covariate vector X, say that *group g is disadvantaged* if its error at its optimal point G_X is larger than the other group g''s error at its optimal point G'_X; that is, the minimal achievable error for group g (given X) is larger than that for the other group. (In the case of a group-skewed covariate vector X, the disadvantaged group receives the higher error at every point on the fairness-accuracy frontier.) We say a point is $\succeq$-optimal for a FA preference $\succeq$ if it is the best point in the feasible set.

Definition 10. Given any FA preference $\succeq$, $e^* \in \mathcal{E}(X)$ is $\succeq$-optimal if $e^* \succeq e$ for all $e \in \mathcal{E}(X)$.

Let (X, G) denote the covariate vector when group identity is included in addition to X.

Corollary 2. *Suppose group g is disadvantaged given X and fix any FA preference $\succeq$. Then for any $\succeq$-optimal $e^* \in \mathcal{E}(X)$ and $\succeq$-optimal $e^{**} \in \mathcal{E}(X, G)$, it holds that $e^{**}_g \leq e^*_g$.*

That is, the disadvantaged group's error at any designer's optimal point given (X, G) must be weakly smaller than its error at the designer's optimal point given X only.[22] The corresponding statement for the advantaged group is not true. For example, in Panel (b) of Figure 6 below, the Egalitarian designer chooses to increase group r's error when given more information about G.[23]

[22]It is straightforward to see that Corollary 2 holds even if we compare a $\succeq$-optimal point on $\mathcal{E}(X)$ with a $\succeq'$-optimal point on $\mathcal{E}(X, G)$, where $\succeq$ and $\succeq'$ are different FA-preferences. That is, giving the algorithm access to G will reduce the disadvantaged group's error even if the designer choosing the algorithm has changed.

[23]Corollary 2 does not necessarily imply that the disadvantaged group's *welfare* increases when group identity is used, since it could be that the designer cares about inaccuracies (e.g., measuring error using the loss function $\ell(d, g, y) = \mathbb{1}(d \neq y)$), while the individuals care about receiving a specific decision outcome (e.g., measuring welfare using loss function $\ell(d, g, y) = \mathbb{1}(d \neq 1)$). See further discussion in Section 5.

4 Input Design

We have so far assumed that the designer directly chooses the best algorithm to maximize a preference that (weakly) responds to both fairness and accuracy. This is a good description of some settings; for example, a company may internalize fairness concerns in its hiring algorithm. But often the algorithm is set by an agent who does not care about fairness across groups, while the inputs used by the algorithm are constrained by a designer who does. 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. Or, a bank (agent) may seek to maximize profit from loan issuance, while a regulator (designer) may require that the rate at which individuals are incorrectly denied loans does not differ too much across groups. In these settings, the designer can often influence the algorithm indirectly by passing regulation that constrains the algorithm's inputs, for example by excluding the use of specific covariates available to the algorithm.

In Section 4.1, we model this interaction as an information design problem in which the designer constrains the inputs of the algorithm, while the algorithm is chosen by an accuracy-minded agent. In Section 4.2, we ask whether the designer should completely exclude an input such as group identity.

4.1 Input Design Versus Algorithm Design

A 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 (assumed without loss to be finite).[24] Examples include:

Example 8 (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 9 (Adding Noise). $T(x) = x + \varepsilon$ where the noise term ε takes value $+1$ or -1 with equal probability.

Example 10 (Coarsening the Input). The set of realizations $\mathcal{X} = \{1, 2, 3, 4\}$ is partitioned into $\{\{1, 2\}, \{3, 4\}\}$, and $T(x)$ reports (with probability 1) the partition element to which x belongs.

We view these garblings as information policies that the designer can commit to by law. For example, the "ban-the-box" campaign (Agan and Starr, 2018) restricted employers from

[24] This corresponds to a constrained version of the information design problem, where the designer has access to garblings of a given information structure X only.

using criminal history as an input into hiring decisions (similar to Example 8), and the College Board coarsens a test-taker's answers into an integer-valued score between 400 and 1600 (similar to Example 10).[25]

The agent chooses an algorithm $a : \mathcal{T} \to \Delta(\mathcal{D})$ that takes as input the garbling chosen by the designer. The agent's utility function is

$$-\alpha_r \cdot e_r(a) - \alpha_b \cdot e_b(a)$$

for some constants $\alpha_r, \alpha_b \geq 0$, with the special case $\alpha_g = p_g$ returning Utilitarian preferences.[26,27] (We prove additional results in Appendix O.3 for the case in which some coefficient α_g is negative, so that the agent is adversarial and prefers to increase error for one of the two groups.) Since we can rewrite this utility as

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

where p_t is the probability of $T = t$, the agent's problem of minimizing ex-ante error is equivalent to the following ex-post problem[28]

$$a(t) \in \arg\min_{d \in \mathcal{D}} \sum_{y,g} \frac{\alpha_g}{p_g} \cdot \mathbb{P}\left(Y = y, G = g \mid T = t\right) \cdot \ell\left(d, y, g\right). \tag{1}$$

Definition 11. The pair of group errors (e_r, e_b) is *implemented by* T if there exists an algorithm a_T satisfying (1) such that $(e_r, e_b) = (e_r(a_T), e_b(a_T))$.

Definition 12. The *input-design feasible set* given X consists of all error pairs that the

[25] A related example—Chan and Eyster (2003) report a law school admission process that used only a coarsened version of the candidates' LSAT scores: "Nor does [Boalt Hall, UC Berkeley's law school] consider candidates' exact LSAT scores; instead, LSAT scores are partitioned into intervals, and the admissions committee only learns which interval contains the candidate's score."

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

[27] We view the most practically relevant settings as those where the agent cares about improving accuracy. The case in which the agent additionally values fairness is also interesting, but introduces novel technical complications (see Section 5 for further discussion).

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

designer can implement using a garbling of X:

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

The *input-design fairness-accuracy frontier* $\mathscr{F}^*(X)$ is the set of error pairs $e \in \mathcal{E}^*(X)$ with the property that no other $e' \in \mathcal{E}^*(X)$ satisfies $e' >_{FA} e$.

Under relatively weak conditions, it turns out to be without loss to have control only of the algorithm's inputs: Any error pair that a designer would choose to implement in the unconstrained problem (i.e. when the designer can choose the algorithm) can also be achieved under input design. To state the result, we define

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

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 no information.

Proposition 2 (When Input Design is Without Loss)**.** *The following hold:*

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

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

This result follows from the subsequent lemma, which says that the input-design feasible set is equal to the intersection of the unconstrained feasible set and H, with an analogous statement relating the fairness-accuracy frontiers. A version of this lemma has been demonstrated in Alonson and Câmara (2016) and Ichihashi (2019), although we provide an independent argument in Appendix 1 for completeness.

Lemma 1. *For every covariate vector X, the input-design feasible set is $\mathcal{E}^*(X) = \mathcal{E}(X) \cap H$ and the input-design fairness-accuracy frontier is $\mathscr{F}^*(X) = \mathscr{F}(X) \cap H$.*

Clearly the agent's payoff cannot be made worse off than if the agent were given no information, so $\mathcal{E}^*(X) \subseteq \mathcal{E}(X) \cap H$. To show that every point in $\mathcal{E}(X) \cap H$ can be implemented, consider the garbling of X into recommendations of decisions. The obedience constraints turn out to reduce precisely to the condition that the agent's payoff is improved relative to

no information, implying the lemma. Figure 5 provides an illustration of how Proposition 2 is implied by Lemma 1.

These results tell us that input design is always sufficient to recover part of the original fairness-accuracy frontier. Moreover, so long as certain points (R_X and B_X in the case of a group-balanced X, or G_X and F_X in the case of a g-skewed X) improve the agent's payoffs relative to no information, then the designer can induce the agent to choose the designer's most preferred outcome even without explicit control of the algorithm. Conversely, when these conditions do not hold, then input design is limiting; designers with certain preferences are unable to achieve their most preferred outcome.

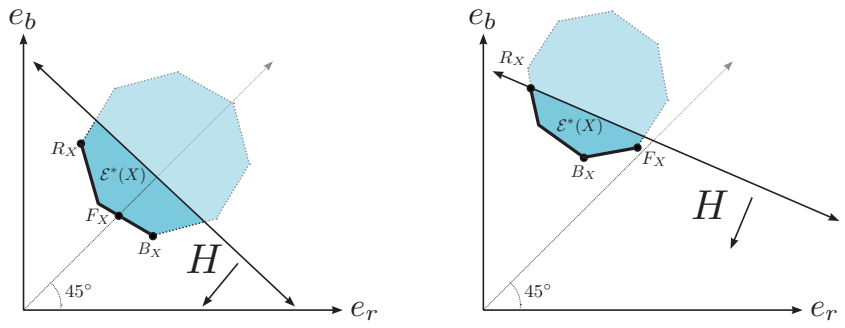


Figure 5: Depiction of an example input-design fairness-accuracy frontier for (a) a group-balanced covariate vector X and (b) an r-skewed covariate vector X.

4.2 Excluding a Covariate

In practice, constraints on algorithmic inputs sometimes take the form of a complete ban on use of a specific covariate. For example, protected group identities such as race, religion and gender are illegal inputs into lending and hiring decisions,[29] and the University of California university system recently excluded consideration of standardized test scores from their admissions decisions.[30]

Since the designer and agent have (potentially) misaligned preferences, giving the agent more information is not always better for the designer. But for two important classes of inputs, we show that excluding the input is worse for all FA-preferences.

[29]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.

[30]See for reference: `https://www.nytimes.com/2021/05/15/us/SAT-scores-uc-university-of-california.html` and Garg et al. (2021).

Definition 13. Say that *excluding covariate X' over X uniformly worsens the (input design) frontier* if every point in $\mathscr{F}^*(X)$ is FA-dominated by a point in $\mathscr{F}^*(X, X')$.

To interpret this condition, recall that $\mathscr{F}^*(X)$ is the frontier of error pairs that can be implemented by some garbling of X, while $\mathscr{F}^*(X, X')$ is the frontier of error pairs that can be implemented by some garbling of (X, X'). So any point that belongs to $\mathscr{F}^*(X, X')$ but not to $\mathscr{F}^*(X)$ can only be implemented if the garbling chosen by the designer includes some information about X'. We view garblings as a practically useful tool for the designer, and thus choose to compare the full set of garblings given X versus (X, X'), rather than the more limited information policy of fully revealing X versus (X, X'). Note that uniform worsening of the frontier does not imply a direct comparison between fully revealing X or fully revealing (X, X').

4.2.1 Excluding Group Identity

First compare garblings of X with garblings of (X, G). The property of group balance (suitably strengthened) turns out to be critical:

Definition 14. Say that X is *strictly group-balanced* if $e_r < e_b$ at R_X and $e_b < e_r$ at B_X.

Relative to group-balance, strict group-balance rules out covariate vectors X for which $R_X = B_X = F_X$.

Proposition 3. *Suppose $R_X, B_X \in H$. Then, excluding G over X uniformly worsens the frontier if and only if X is strictly group-balanced.*[31]

To show this result, we first demonstrate that the minimal (and maximal) feasible error for both groups is the same given X and given (X, G). Geometrically, this means that the feasible set given (X, G) is the smallest rectangle containing the feasible set given X. When X is group-balanced, then $\mathscr{F}^*(X)$ is characterized by Part (a) of Theorem 1 while $\mathscr{F}^*(X, G)$ is characterized by Proposition 1 (using the equivalence in Proposition 2 for both cases). As depicted in Panel (a) of Figure 6, the fairness-accuracy frontier given X does not intersect with the frontier given (X, G), so every point on the new frontier (after excluding G) is FA-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).

Proposition 3 implies that for a large class of covariate vectors (any X that is strictly group-balanced), *every* designer can strictly improve their payoffs by choosing a garbling

[31] 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 3 that even when this assumption fails, strict group-balance is a sufficient condition for the frontier to uniformly worsen when excluding G.

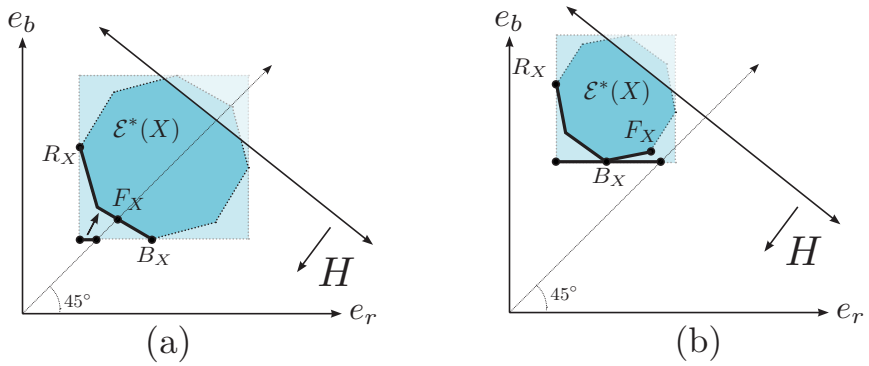


Figure 6: (a) X is strictly group-balanced and excluding G over X uniformly worsens the input-design frontier; (b) X is r-skewed and excluding G over X does not uniformly worsen the input-design frontier.

that uses information about G (beyond what was already known about G from X).[32] Conditioning on G allows the designer to use garblings of X that differ across groups. Thus when X is strictly group-balanced, even arbitrarily fairness-minded designers strictly prefer to implement noisy transformations that are asymmetric between the two groups. Such policies may be unfair in terms of of disparate *treatment* (i.e., whether the policy discriminates between individuals on the basis of group identity), but may be necessary to impose fairness in terms of disparate *impact* (i.e., whether the adverse effects of the policy are disproportionately borne by members of a specific group).[33] Our analysis helps to formalize the tension between these goals.

Proposition 3 is related to results in the algorithmic fairness literature (e.g, Kleinberg and Mullainathan (2019) and Kleinberg et al. (2018)) and the earlier statistical discrimination literature; for example, Chan and Eyster (2003) presents a model of college admissions in which restricting ability to condition on race results in poorer student quality, and Lundberg (1991) presents a model in which prohibiting firms to condition wages on group identity reduces efficiency. Our Proposition 3 relies on similar forces, but is derived for a broad class of designer criteria. Additionally, the link that we demonstrate between the consequences of banning group identity and the property of group-balance emerges from our consideration of the full fairness-accuracy frontier.

[32] We show in Appendix O.3 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.

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

4.2.2 Excluding a Covariate When Group Identity is Known

Next compare garblings of X with garblings of (X, X') when X reveals G. First, a definition.

Definition 15. 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

$$\{1\} = \arg\min_{d \in \mathcal{D}} \mathbb{E}[\ell(a, Y, g) \mid X = x, X' = x', G = g]$$

while

$$\{0\} = \arg\min_{d \in \mathcal{D}} \mathbb{E}[\ell(a, Y, g) \mid X = x, X' = \tilde{x}', G = g].$$

This weak condition requires only that there is some individual in group g for whom the decision that maximizes (expected) accuracy is different given X than that given (X, X'), so that the information in X' is decision-relevant for that individual.

Proposition 4. *Suppose X reveals G and choose an arbitrary covariate vector X'.*

(a) If X is g-skewed, then excluding X' over 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' over X uniformly worsens the frontier if and only if X' is decision-relevant over X for both groups.

When X' is decision-relevant over X for the disadvantaged group, then the minimal feasible error for that group given (X, X') is strictly lower than the minimal feasible error given X only. So the fairness-accuracy frontier is pushed towards the origin (either downwards or the left), as in 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 fairness-accuracy frontier must remain a line that overlaps with the previous frontier (see Panel (b) of Figure 7), so there is some FA preference for which excluding X' is at least weakly (and possibly strictly) worse. This yields part (a) of the result.

The case of a group-balanced X is very special under the assumption that X reveals G: it must be that the minimal feasible error is the same for both groups. This minimal feasible error is reduced through access to X' only when X' is decision-relevant for both groups, yielding part (b) of the result.

Suppose X' is a test score. The condition of decision-relevance does not depend on whether the covariate X' is biased, in the sense of being systematically lower-valued (as in Example 2) or less informative for either group (as in Example 3), but only on whether

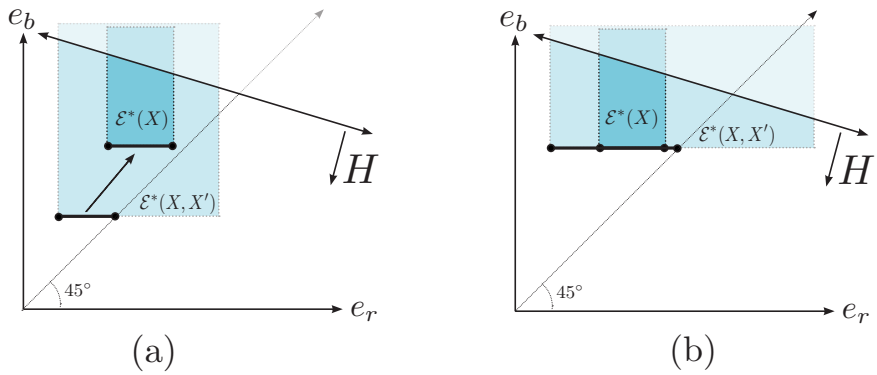


Figure 7: (a) Example in which X' is decision-relevant for group b, and excluding X' uniformly worsens the frontier; (b) Example in which X' is not decision-relevant for group b, and excluding X' does not uniformly worsen the frontier.

the test score is informative. We expect that test scores are decision-relevant in practice.[34] Our result thus suggests the following: So long as group identities are permissible inputs for college admission decisions (as is presently true in most states in the 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 the optimal garbling of covariates for a sufficiently fairness-minded designer may involve completely excluding that covariate. We conclude with a simple example to illustrate such a possibility.

Example 11. Suppose $\mathcal{Y} = \{0, 1\}$ and Y and G are 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 $\mathbb{P}(X' \mid Y, G)$: [35]

	$X'=1$	$X'=0$
$Y=1$	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 informa-

[34] Rambachan et al. (2021) study a screening model and demonstrate that any informative covariate, however biased, will be optimally used by a social planner with control of the algorithm. We show that this insight extends *when group identity is available* even when the social planner chooses only the inputs of the algorithm (while another agent chooses the algorithm), but can fail when group identity is not available.

[35] In this example, neither covariates X nor X' reveal group identity. Thus, this example falls outside of the settings considered in the previous two subsections.

tive about those in group b. Suppose $\ell(d, y, g) = \mathbb{1}(d \neq y)$ is the misclassification rate, and the agent is Utilitarian ($\alpha_r = p_r = 1/2$ and $\alpha_b = p_b = 1/2$).

As we compute in Appendix B.10, the input-design feasible set $\mathcal{E}^*(X, X')$ is the line segment connecting $(0, 0.4)$ with $(0.5, 0.5)$. This entire line segment is also the fairness-accuracy frontier $\mathscr{F}^*(X, X')$, as illustrated in Figure 8:

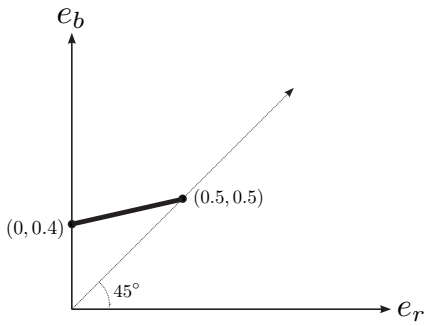


Figure 8: The input-design fairness-accuracy frontier given (X, X')

For an Egalitarian designer, sending the null signal X leads to the point $(0.5, 0.5)$ which yields a payoff of 0. If the designer chooses any nontrivial garbling of (X, X'), then the agent will use this additional information about X' to maximize aggregate accuracy. Since this information is inevitably more informative about group r than about group b, conditioning decisions on this information increases the gap between the two group errors, reducing the designer's payoff. While we assume an Egalitarian designer here for simplicity, a similar construction is possible for any designer who places sufficient weight on fairness considerations.

5 Extensions

Different loss functions for evaluating fairness and accuracy. When defining the partial order $>_{FA}$ we use the same loss function to evaluate accuracy and fairness. In some cases, the designer may wish to evaluate accuracy using one loss function and fairness using another. (For example, the designer may wish to minimize the misclassification rate subject to equality of false positive rates.) In Appendix O.1 we develop a more general version of our framework that allows for different loss functions, and extend Theorem 1 under an assumption that the accuracy and fairness loss functions are not "directly opposed" to one another. In this result, our group-balance condition is generalized to a condition of whether the fairness-maximizing point F_X belongs to usual Pareto frontier. When this condition is satisfied, then the fairness-accuracy frontier is identical to the usual Pareto

frontier; otherwise, the fairness-accuracy frontier is the union of the Pareto frontier and a positively-sloped sequence of lines, along which every pair of points has the property that one point involves higher errors for both groups but greater fairness.

Beyond absolute difference for evaluating fairness. Our main analysis assumes that (un)fairness is evaluated according to the absolute difference of errors between the two groups, i.e. $|e_r - e_b|$. A natural extension is to consider $|\phi(e_r) - \phi(e_b)|$ where ϕ is some continuous strictly increasing function. For instance, if ϕ is log, then this corresponds to evaluating fairness using the ratio of errors rather than their difference. Our main characterization (Theorem 1) holds for any such ϕ with the fairness optimal point F_X suitably defined.[36] We further demonstrate various comparative statics in ϕ in Appendix O.2, showing that the frontier becomes larger (smaller) whenever ϕ is concave (convex). Thus, for example, evaluating fairness using ratios instead of absolute difference results in a larger frontier, although the qualitative properties of this frontier are unchanged.

Other agent preferences in the input design problem. Section 4 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 O.3, 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 group's error to be higher. We generalize several results from Section 4 and show that, perhaps surprisingly, even if the agent is negatively biased, it can still be optimal for the designer to provide information about group identity so long as the bias is not too extreme.

Two other potential generalizations would be to permit the agent and designer to have different loss functions, or to permit the agent to care about fairness.[37]. In both cases, the set of points that the agent prefers over the prior (what we defined to be H) is no longer a halfspace from the designer's perspective. Moreover, non-linearities in the agent's objective function imply that the agent's ex-ante and ex-post problems may be different, and so it is relevant whether the agent commits to the algorithm or chooses the decision after the realization of the garbling. We consider these problems beyond the scope of the present paper, and leave them as open questions for future work.

[36] To see why, first note that no interior point can be on the frontier. Otherwise, we can always find some $\epsilon_1, \epsilon_2 > 0$ such that $|\phi(e_r - \epsilon_1) - \phi(e_b - \epsilon_2)| \leq |\phi(e_r) - \phi(e_b)|$ so $(e_r - \epsilon_1, e_b - \epsilon_1) >_{FA} (e_r, e_b)$ yielding a contradiction. The rest of the proof follows as in Theorem 1.

[37] 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 4

Capacity constraints. In our main model, we allow the designer unconstrained choice of any algorithm. In a few of the applications of interest, there may be an additional capacity constraint on the algorithm, e.g., in admissions decisions, only a fixed number of students can be admitted. One way to formulate a capacity constraint is a restriction on the ex-ante probability of assignment of decision $d = 1$ (e.g., admit). In this case, the set of error pairs satisfying the constraint can be shown to be a convex set, so the feasible set is simply the intersection between the feasible set (as we have defined) and the convex set of error pairs that satisfy this capacity constraint. Our Theorem 1 then applies for this new feasible set, although the fairness-accuracy frontier as characterized in Proposition 1 may no longer be a horizontal line.

More than two decisions. We have assumed that there are two decisions $\mathcal{D} = \{0, 1\}$. All of our results in Section 3 about the unconstrained problem directly extend for any finite $\mathcal{D}$. However, Lemma 1 (the relationship between the input-design fairness-accuracy frontier and the unconstrained fairness-accuracy frontier) relies on the assumption of two decisions. We leave a characterization of the input design frontier analysis for this more general case to future work.

More than two groups. We have assumed that there are two groups $\mathcal{G} = \{r, b\}$. Some of our results, such as Proposition 2, 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 FA-dominance relationship is to say that a vector of group errors $(e_g)_{g\in\mathcal{G}}$ FA-dominates another vector $(e'_g)_{g\in\mathcal{G}}$ if $e_g \leq e'_g$ for every group g, and also $|e_g - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} e_g| \leq |e'_g - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} e'_g|$ 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. We expect our characterization in Theorem 1 to extend qualitatively in this case.

A Fairness Criteria in the Literature

We review here certain fairness criteria that have appeared in the literature, and explain how these criteria can be accommodated within our framework.

Statistical Parity. This criterion seeks equality in decisions, namely that the proportion of either group receiving the two decisions is the same (Dwork et al., 2012). Formally, an

algorithm a satisfies statistical parity if

$$\mathbb{E}(a(X) = 1 \mid G = r) - \mathbb{E}(a(X) = 1 \mid G = b) = 0$$

The loss function

$$\ell(d, y, g) = \begin{cases} 1 & \text{if } d = 1 \\ 0 & \text{otherwise} \end{cases}$$

returns a relaxed version of this criterion, since

$$e_g(a) = \mathbb{E}\left[\ell(a(X), Y, g) \mid G = g\right] = \mathbb{E}\left[a(X) = 1 \mid G = g\right]$$

so $|e_r(a) - e_b(a)|$ is the absolute difference in the probability that a group-r individual and a group-b individual receive the decision $d = 1$.

False Positives. Another common fairness criterion is equality of false positives across two groups (Chouldechova, 2017; Kleinberg et al., 2017). For example, among borrowers who would not have defaulted on their loan if approved, prediction of default should be equal across the two groups. Formally, an algorithm a satisfies equality of false positive rates if

$$\mathbb{E}(a(X) = 1, Y = 0 \mid G = r) - \mathbb{E}(a(X) = 1, Y = 0 \mid G = b) = 0$$

The loss function

$$\ell(d, y, g) = \begin{cases} 1 & \text{if } (d, y) = (1, 0) \\ 0 & \text{otherwise} \end{cases}$$

returns a relaxed version of this criterion, since

$$e_g(a) = \mathbb{E}\left[\ell(a(X), Y, g) \mid G = g\right] = \mathbb{E}\left[a(X) = 1, Y = 0 \mid G = g\right]$$

is the false-positive rate for group g, and so $|e_r(a) - e_b(a)|$ is the absolute difference in false positive rates. A fairness criterion based on the difference in false negative rates can be accommodated similarly.

Equalized Odds. Another popular fairness criterion asks for equalized odds (Hardt et al., 2016), which an algorithm a satisfies if

$$\mathbb{E}_Y[\mathbb{E}_X[a(X) \mid G = r, Y] - \mathbb{E}_X[a(X) \mid G = b, Y]] = 0 \tag{A.1}$$

The inner difference compares the average decision for group-r and group-b individuals who share the same type Y, and the outer expectation averages over those values of Y.

The loss function

$$\ell(d, y, g) = \begin{cases} \frac{P(Y=y)}{P(Y=y|G=g)} & \text{if } d = 1 \\ 0 & \text{otherwise} \end{cases}$$

returns a relaxed version of this criterion, since

$$\begin{aligned}
\mathbb{E}[\ell(d, y, g) \mid G = r] &= P(Y = 0 \mid G = r) \times \mathbb{E}\left[\frac{P(Y = 0)}{P(Y = 0 \mid G = r)} \times \mathbb{1}(d = 1) \mid G = r, Y = 0\right] \\
&\quad + P(Y = 1 \mid G = r) \times \mathbb{E}\left[\frac{P(Y = 1)}{P(Y = 1 \mid G = r)} \times \mathbb{1}(d = 1) \mid G = r, Y = 1\right] \\
&= P(Y = 0) \times \mathbb{E}[\mathbb{1}(d = 1) \mid G = r, Y = 0] \\
&\quad + P(Y = 1) \times \mathbb{E}[\mathbb{1}(d = 1) \mid G = r, Y = 1]
\end{aligned}$$

so $|\mathbb{E}[\ell(a(X), Y, G) \mid G = r] - \mathbb{E}[\ell(a(X), Y, G) \mid G = b]|$ is exactly the LHS of (A.1).

B Proofs for Main Text Results

B.1 Characterization of the Feasible Set

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

Proof. Given algorithm a, we slightly abuse notation to let $a(x)$ denote the probability of choosing decision $d = 1$ at covariate vector 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:

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

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

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

If we let

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

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. □

B.2 Proof of Theorem 1

First observe that the FA frontier must be part of the boundary of the feasible set $\mathcal{E}(X)$, because any interior point (e_r, e_b) is FA-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 since it must be on the 45-degree line. In this case it is clear that the fairness-accuracy frontier consists of that single point, which FA-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 FA-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 FA-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 $Q' >_{FA} Q$. 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 FA-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 FA-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* FA-dominated. R_X (and symmetrically B_X) cannot be FA-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 FA-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 FA-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 FA-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 FA-dominates Q. If Q^* lies on the extension of the ray $F_X R_X$, then as before it can be shown that $R_X >_{FA} Q$. 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 FA-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 FA-undominated.

By essentially the same argument, we know that the lower boundary from R_X to B_X is FA-undominated. As for the lower boundary from B_X to F_X, note that if some point Q here is FA-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_XQ. 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 FA-dominate Q, completing the proof.

B.3 Proof of Corollary 1

Suppose X is group-balanced, then by Theorem 1 the fairness-accuracy 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_XL_X and B_XL_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 $\mathscr{F}(X)$ 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 fairness-accuracy 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_XF_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.

B.4 Proof of Proposition 1

We recall the proof of Lemma B.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 $\left(\sum_y \frac{x_{y,r}}{p_r} \ell(1,y,r), \sum_y \frac{x_{y,b}}{p_b} \ell(1,y,b)\right)$ and $\left(\sum_y \frac{x_{y,r}}{p_r} \ell(0,y,r), \sum_y \frac{x_{y,b}}{p_b} \ell(0,y,b)\right)$. 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 fairness-accuracy 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 fairness-accuracy frontier is the horizontal line segment from $R_X = B_X$ to F_X. This proves the result.

B.5 Proof of Corollary 2

Suppose group b is disadvantaged given covariate vector X. Then the covariate vector (X, G) must be r-skewed, since the point $B_{X,G} = R_{X,G}$ 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, and similarly for group r). Choose an arbitrary FA preference $\succeq$, and let e^* and e^{**} respectively be $\succeq$-optimal points on $\mathcal{E}(X)$ and $\mathcal{E}(X,G)$. Since (X,G) is r-skewed, e_b^{**} is 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 $(e_r, e_b) \in \mathcal{E}(X)$, and thus certainly also $e_b^{**} \leq e_b^*$.

B.6 Proof of Lemma 1

We first characterize the input-design feasible set, and later study the input-design fairness-accuracy frontier. 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 $\alpha_r e_r + \alpha_b e_b \leq e_0$ can be implemented by some garbling T. Consider a garbling T that maps X to $\Delta(\mathcal{D})$, with the interpretation that the realization of $T(x)$ is the recommended

decision for the agent. If we abuse notation to let $a(x)$ denote the probability that the recommendation is $d = 1$ at covariate vector x, then this algorithm a needs to satisfy the following obedience constraint for $d = 1$:[38]

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

The above is just equation (1) 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 $\sum_x p_{x,y,g} \cdot a(x)$, 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

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

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

$$\sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot (a(x)\ell(1, y, g) + (1 - a(x))\ell(0, y, g)) \leq \sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot \ell(0, y, g). \tag{B.1}$$

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

On the other hand, the RHS in (B.1) can be seen to be the agent's expected loss when taking the decision $d = 0$ regardless of the designer's recommendation. Thus, we deduce that the obedience constraint for the recommendation $d = 1$ is equivalent to (B.1), which simply says that the agent's payoff under the designer's recommendation should be weakly better than the constant decision $d = 0$ ignoring the recommendation. Symmetrically, the other obedience constraint for the recommendation $d = 0$ is equivalent to the agent's payoff being better than the constant decision $d = 1$. Put together, these obedience constraints thus reduce 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 a garbling T that implements it by recommending the desired decision. 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.

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

Finally we turn to the fairness-accuracy frontier and argue that $\mathscr{F}^*(X) = \mathscr{F}(X) \cap H$. In one direction, if an error pair is FA-undominated in $\mathcal{E}(X)$ and implementable under input design, then it is also FA-undominated in the smaller set $\mathcal{E}^*(X)$. This proves $\mathscr{F}(X) \cap H \subseteq \mathscr{F}^*(X)$. In the opposite direction, suppose for contradiction that a certain point $(e_r, e_b) \in \mathscr{F}^*(X)$ does not belong to $\mathscr{F}(X) \cap H$. Since $\mathscr{F}^*(X) \subseteq \mathcal{E}^*(X) \subseteq H$, we know that (e_r, e_b) must not belong to $\mathscr{F}(X)$. Thus by definition of $\mathscr{F}(X)$, (e_r, e_b) is FA-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 $(e_r, e_b) \in \mathscr{F}^*(X) \subseteq \mathcal{E}^*(X)$). It follows that the FA-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 FA-undominated in $\mathcal{E}^*(X)$. Such a contradiction completes the proof.

B.7 Proof of Proposition 2

We now deduce Proposition 2 from Lemma 1. If X is group-balanced, then by Theorem 1 we know that $\mathscr{F}(X)$ is the part of the boundary of $\mathcal{E}(X)$ that connects R_X to B_X from below. Clearly, $\mathscr{F}^*(X) = \mathscr{F}(X)$ can only hold if $R_X, B_X \in \mathscr{F}^*(X) \subseteq H$, 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 $\mathscr{F}(X)$ 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 $\mathscr{F}(X) \subseteq H$. Hence whenever $R_X, B_X \in H$, we have by Lemma 1 that $\mathscr{F}^*(X) = \mathscr{F}(X) \cap H = \mathscr{F}(X)$. This argument proves Proposition 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.[39] 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

[39]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 positively sloped because B_X has minimum e_b. The final edge of this boundary must also be positively sloped, since otherwise the starting vertex of this edge would be closer to the 45-degree line than F_X. It follows by convexity that the entire boundary from B_X to F_X has positive slopes.

B_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.

B.8 Proof of Proposition 3

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

Lemma B.2. *Excluding covariate X' over X uniformly worsens the frontier if and only if $\mathscr{F}^*(X)$ does not intersect with $\mathscr{F}^*(X, X')$.*

The proof of this lemma is straightforward: If there exists a point in $\mathscr{F}^*(X)$ that also belongs to $\mathscr{F}^*(X, X')$, then this point is not FA-dominated by any point in $\mathscr{F}^*(X, X')$, so that the frontier does not uniformly worsen when excluding X'. On the other hand, suppose no point in $\mathscr{F}^*(X)$ belongs to $\mathscr{F}^*(X, X')$. Note that any point in $\mathscr{F}^*(X)$ 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 FA-optimal in this set, it must be FA-dominated by some FA-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 B.2 to deduce Proposition 3. 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 $R_X, B_X \in H$, Proposition 2 tells us that the input-design fairness-accuracy frontier $\mathscr{F}^*(X)$ is the same as the unconstrained fairness-accuracy frontier $\mathscr{F}(X)$, 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 B.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 fairness-accuracy frontier $\mathscr{F}^*(X, G)$ given (X, G). To characterize the latter frontier, let $L_X = R_{X,G} = B_{X,G}$ denote 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 fairness-accuracy frontier $\mathscr{F}(X, G)$ 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 fairness-accuracy frontier $\mathscr{F}^*(X,G)$ 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 fairness-accuracy frontier $\mathscr{F}(X,G)$ 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, $B_X \in \mathscr{F}(X)$ and $B_X \in \mathscr{F}(X,G)$. But by assumption, B_X also belongs to H. So Lemma 1 tells us that B_X belongs to the input-design fairness-accuracy frontiers $\mathscr{F}^*(X)$ and $\mathscr{F}^*(X,G)$. This shows that the two frontiers $\mathscr{F}^*(X)$ and $\mathscr{F}^*(X,G)$ intersect, which completes the proof by Lemma B.2.

B.9 Proof of Proposition 4

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

Lemma B.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 outcome for group g, i.e.,

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

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

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

By optimality of a^*_g,

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

$$\leq \mathbb{E}\left[\ell(a_g(x), Y, g) \mid G = g, X = x, X = x'\right] \quad \forall x \in \mathcal{X}, \forall x' \in \mathcal{X}'. \quad \text{(B.2)}$$

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

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

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

We now use Lemma B.2 and B.3 to prove Proposition 4. 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 fairness-accuracy frontier $\mathscr{F}(X)$ 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 B.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 fairness-accuracy frontier $\mathscr{F}(X, X')$ 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 fairness-accuracy frontier $\mathscr{F}(X, X')$ is a horizontal line segment that is a superset of the line segment $\mathscr{F}(X)$. 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 fairness-accuracy frontiers $\mathscr{F}^*(X)$ and $\mathscr{F}^*(X, X')$ ($R_X = B_X$ 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 B.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 B.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 fairness-accuracy frontier $\mathscr{F}(X, X')$ 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 fairness-accuracy frontier $\mathscr{F}(X)$. Thus $\mathscr{F}(X)$ does not intersect $\mathscr{F}(X, X')$. As their subsets, the input-design fairness-accuracy frontiers $\mathscr{F}^*(X)$ and $\mathscr{F}^*(X, X')$ also do not intersect. Thus by Lemma B.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 fairness-accuracy frontier $\mathscr{F}(X, X')$ 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 $\mathscr{F}(X, X')$ has strictly smaller e_b compared to any point on $\mathscr{F}(X)$. Once again these two unconstrained frontiers do not intersect, and nor do the input-design frontiers. This proves Proposition 4 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 fairness-accuracy frontiers $\mathscr{F}(X)$ and $\mathscr{F}^*(X)$ are both this singleton point. If X' is not decision-relevant over X for group b, then Lemma B.3 tells us that $\underline{e}_b^* = \underline{e}_b = \underline{e}_r \geq \underline{e}_r^*$. When equality holds the fairness-accuracy frontiers $\mathscr{F}(X, X')$ and $\mathscr{F}^*(X, X')$ 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 fairness-accuracy frontier $\mathscr{F}(X, X')$ 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 fairness-accuracy frontier $\mathscr{F}^*(X, X')$, showing that $\mathscr{F}^*(X)$ and $\mathscr{F}^*(X, X')$ intersect. Uniform worsening of the frontier does not occur either way.

B.10 Details of Example 11

In this section, we compute the input-design feasible set and fairness-accuracy frontier for Example 11. Since X is a null signal, garblings of (X, X') are the same as garblings of X'. Without loss, we can restrict attention to garblings of X' that take two values, $d = 1$ and $d = 0$, which correspond to the designer's decisions 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 $d = 1$, and β is the probability with which $X' = 0$ is mapped into $d = 1$. It is easy to check that the agent's obedience constraint reduces to the simple inequality $\alpha \geq \beta$, which intuitively requires the agent to choose $d = 1$ more often when $X' = 1$.

For any pair (α, β), the two groups' errors can be calculated as

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

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

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 fairness-accuracy frontier $\mathscr{F}^*(X, X')$, as illustrated in Figure 8 in the main text.

For an Egalitarian designer, sending the null signal X leads to the point $(0.5, 0.5)$ and yields a payoff of 0. In contrast, we say that the designer "makes use of X' over X" if the garbling T is *not* independent of X' conditional on X (in this example the conditioning is irrelevant since X is null). Whenever T is not independent of X', then for some realizations of T the agent believes $X' = 1$ is more likely, which makes $d = 1$ strictly optimal. Thus, whenever the designer makes use of X' in the garbling, the agent is strictly better off compared to the null signal, and the resulting error pair must be distinct from $(0.5, 0.5)$. But given the shape of the implementable set, this means that the designer is strictly worse off when any information about X' is provided to the agent.

Conversely, suppose X' is decision-relevant over X for both groups. Then by Proposition 1, the unconstrained frontier $\mathscr{F}(X, X')$ 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 $\mathscr{F}(X)$ does not intersect with $\mathscr{F}(X, X')$. Hence $\mathscr{F}^*(X)$ and $\mathscr{F}^*(X, X')$ also do not intersect, and by Lemma B.2 we know that there is uniform worsening of the frontier. This completes the entire proof of Proposition 4.

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.

Alonson, R. and O. Câmara (2016): "Persuading Voters," *American Economic Review*, 106, 3590–3605.

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.

Bertrand, M. and E. Kamenica (2020): "Coming apart? Cultural distances in the United States over time," Working Paper.

Chan, J. and E. Eyster (2003): "Does Banning Affirmative Action Lower College Student Quality?" *American Economic Review*, 93, 858–872.

Chohlas-Wood, A., M. Coots, E. Brunskill, and S. Goel (2021): "Learning to be Fair: A Consequentialist Approach to Equitable Decision-Making," Working Paper.

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.

Fehr, E. and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

Ferry, J., U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala (2022): "Improving Fairness Generalization Through a Sample-Robust Optimization Method," *Machine Learning*.

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.

Hansen, V. P. B., A. T. Neerkaje, R. Sawhney, L. Flek, and A. Søgaard (2022): "The Impact of Differential Privacy on Group Disparity Mitigation," *ArXiv*, abs/2203.02745.

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.

Ichihashi, S. (2019): "Limiting Sender's Information in Bayesian Persuasion," *Games of Economic Behavior*, 117, 276–288.

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. and S. Mullainathan (2019): "Simplicity Creates Inequity: Implications for Fairness, Steretotypes, and Interpretability," Working Paper.

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.

Lundberg, S. J. (1991): "The Enforcement of Equal Opportunity Laws Under Imperfect Information: Affirmative Action and Alternatives," *The Quarterly Journal of Economics*, 106, 309–326.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2022): "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, 54, 1–35.

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.

Rawls, J. (1971): *A Theory of Justice*, Harvard University Press.

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

Sagawa, S., P. W. Koh, T. B. Hashimoto, and P. Liang (2020): "Distributionally

Robust Neural Networks," in *International Conference on Learning Representations*.

Wei, S. and M. Nietham mer (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.

Online appendix to the paper

Algorithmic Design: Fairness and Accuracy

Annie Liang Jay Lu Xiaosheng Mu

August 16, 2022

O.1 Different Loss Functions

In this section, we generalize Theorem 1 when fairness and accuracy are evaluated using loss functions that are possibly different but not "directly opposed."

As in the main text, let $a : \mathcal{X} \to \Delta(D)$ describe a generic algorithm and let $\mathscr{A}_X$ be the set of all algorithms. Different from the main text, we have two loss functions—an accuracy loss function $\ell^A : \mathcal{D} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$ and a fairness loss function $\ell^F : \mathcal{D} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$. For either group $g \in \{r, b\}$, let

$$e^A_g(a) = \mathbb{E}_{D\sim a(X)}[\ell^A(D, Y, g) \mid G = g] \qquad \forall a \in \mathscr{A}_X$$
$$e^F_g(a) = \mathbb{E}_{D\sim a(X)}[\ell^F(D, Y, g) \mid G = g] \qquad \forall a \in \mathscr{A}_X$$

be group errors defined using the respective loss functions. We use $e^A(a) \equiv (e^A_r(a), e^A_b(a))$ to denote the error pairs evaluated by the accuracy loss function, and

$$\mathcal{E}(X) = \{e^A(a) : a \in \mathscr{A}_X\}$$

to denote the set of feasible (accuracy) error pairs. Also define

$$\Delta(a) = |e^F_r(a) - e^F_b(a)| \qquad \forall a \in \mathscr{A}_X$$

to be the gap between group errors evaluated by the fairness loss function, i.e., the "unfairness" of algorithm a. The function $u : \mathcal{E}(X) \to \mathbb{R}$ satisfying

$$u(e) = \min_{a\in\mathcal{A}_X} \{\Delta(a) : e^A(a) = e\}$$

maps each (accuracy) error pair to the minimal achievable unfairness value. This function is well-defined as $\Delta(\cdot)$ is continuous and $e^A(\cdot)$ is linear.

We now extend the definitions of FA-dominance and the fairness-accuracy frontier.

Definition O.1. Let $>_{FA}$ be the partial order on $\mathcal{E}(X)$ satisfying $(e_r, e_b) >_{FA} (e'_r, e'_b)$ if $e_r \leq e'_r$, $e_b \leq e'_b$, and $u(e) \leq u(e')$, with at least one of these inequalities strict.

Definition O.2. $\mathscr{F}(X)$ is the set of all pairs $e \in \mathcal{E}(X)$ that are FA-undominated, i.e. no $e' \in \mathcal{E}(X)$ exists that satisfies $e' >_{FA} e$.

When $\ell^F = \ell^A$ then we can express u directly as a function of the (accuracy) error-pairs, $u(e) = |e_r - e_b|$, and so these definitions reduce to Definitions 2, 3 and 5.

Lemma O.4. *$u(\cdot)$ is piecewise linear and convex.*

Proof. Since $\mathcal{X}$ is finite, $e^A(\cdot)$ is linear and $\Delta(\cdot)$ is piecewise linear, $u(\cdot)$ must be piecewise linear. We now prove convexity. Fix any $e_1, e_2 \in \mathcal{E}(X)$. Since these error pairs are feasible, there exist algorithms $a_1, a_2 \in \mathcal{A}_X$ that implement them, i.e., $u(e_i) = u\left(e^A(a_i)\right) = \Delta(a_i)$ for each $i = 1, 2$. Let $a = \lambda a_1 + (1-\lambda) a_2$ for $\lambda \in [0,1]$ and note that since $e^A(\cdot)$ is linear, $e^A(a) = \lambda e_1 + (1-\lambda) e_2$. Thus,

$$\begin{aligned} u(\lambda e_1 + (1-\lambda) e_2) &= u\left(e^A(a)\right) \leq \Delta(a) = \left|e_r^F(a) - e_b^F(a)\right| \\ &= \left|\lambda e_r^F(a_1) + (1-\lambda) e_r^F(a_2) - \left(\lambda e_b^F(a_1) - (1-\lambda) e_b^F(a_2)\right)\right| \\ &\leq \lambda \left|e_r^F(a_1) - e_b^F(a_1)\right| + (1-\lambda)\left|e_r^F(a_2) - e_b^F(a_2)\right| \\ &\leq \lambda u(e_1) + (1-\lambda) u(e_2) \end{aligned}$$

as desired. □

Given Lemma O.4, the directional derivatives of u are well-defined in the interior of $\mathcal{E}$. We generalize Theorem 1 under the following assumption.

Assumption 1. *There does not exist $e \in \mathcal{E}(X)$ such that $D_{(1,0)}u(e) < 0$ and $D_{(0,1)}u(e) < 0$.*

This assumption says that, for at least one group, increasing error must hurt fairness. It rules out the case when fairness and accuracy are directly opposed, in the sense that increasing errors in both groups improves fairness. Since we are primarily interested in the tradeoffs between fairness and accuracy due to informational constraints rather than the definitions of fairness and accuracy being intrinsically in conflict, we view this assumption as a natural one for our purposes. In the case when both loss functions are the same so $u(e) = |e_r - e_b|$, this assumption is always satisfied.[40]

[40] In the case where $Y \in \{0,1\}$, the accuracy loss function is the misclassification rate $\ell^A(d,y) = \mathbb{1}(d \neq y)$, and the fairness loss function is $\ell^F(d,y) = \mathbb{1}(d = 1)$, a sufficient condition for Assumption 1 to hold is existence of $x, x' \in \mathcal{X}$ such that $\mathbb{E}(Y = 1 \mid X = x, G = g) < 1/2$ and $\mathbb{E}(Y = 1 \mid X = x', G = g) < 1/2$ for both g (so that the Bayes-optimal assignment at both x and x' is 0 for members of either group), and also $\mathbb{P}(X = x \mid G = r) > \mathbb{P}(X = x \mid G = b)$ while $\mathbb{P}(X = x' \mid G = r) < \mathbb{P}(X = x' \mid G = b)$. Details are available upon request. We leave to future work the derivation of other conditions on primitives for specific (ℓ^A, ℓ^F) pairings.

We now define the fairness-optimal set. First, let

$$\underline{\Delta} := \min_{e \in \mathcal{E}(X)} u(e)$$

be the minimal achievable level of unfairness.

Definition O.3 (Pareto Frontier). For any set $E \subseteq \mathbb{R}^2$, let $\mathcal{P}(E)$ denote the usual Pareto frontier of E, i.e., all points $(e_r, e_b) \in E$ where no $(e'_r, e'_b) \in E$ is weakly smaller in each entry and strictly smaller in at least one.

Definition O.4. The fairness-optimal set is

$$F_X \equiv \mathcal{P}\{e \in \mathcal{E}(X) \,:\, u(e) = \underline{\Delta}\}$$

It's easy to see that F_X is always a subset of the fairness-accuracy frontier.

Theorem O.1. *Under Assumption 1, the following hold:*

1. *If $F_X \subseteq \mathcal{P}(\mathcal{E}(X))$, then $\mathscr{F}(X) = \mathcal{P}(\mathcal{E}(X))$*
2. *If $F_X \nsubseteq \mathcal{P}(\mathcal{E}(X))$, then F_X is a singleton and $\mathscr{F}(X)$ is the union of $\mathcal{P}(\mathcal{E}(X))$ and a connected sequence of positively-sloped line segments ending at F_X*

Thus, the condition that the fairness-maximizing point F_X belongs to the Pareto frontier generalizes group-balance. That is, when this condition is satisfied, we can restrict attention to the usual Pareto frontier without loss. Moreover, no two points on the fairness-accuracy frontier can be Pareto-ranked. When (generalized) group-balance fails, then the frontier consists of two parts: the Pareto frontier, and a positively-sloped sequence of lines, along which every pair of points has the property that one point involves higher errors for both groups but greater fairness. Corollary 1 thus extends directly under this generalized notion of group-balance.

Proof of Theorem O.1. To save on notation we suppress dependence on X in what follows, using $\mathscr{F}$ for the fairness-accuracy frontier and $\mathcal{E}$ for the feasible set. We first show that the fairness-accuracy frontier is the union of the Pareto frontiers of the unfairness sublevel sets.

Definition O.5. For any $\Delta \in \mathbb{R}$, let $\mathcal{E}_{\leq\Delta} = \{e \in \mathcal{E} \mid u(e) \leq \Delta\}$ be u's Δ-sublevel set.

Lemma O.5. $\mathscr{F} = \bigcup_\Delta \mathcal{P}(\mathcal{E}_{\leq\Delta})$.

Proof. Fix any unfairness level Δ and point $e \in \mathcal{P}(\mathcal{E}_{\leq\Delta})$. We will show that e must belong to the fairness-accuracy frontier $\mathscr{F}$. Suppose to the contrary that there exists $e' \in \mathcal{E}$ such that $e'_r \leq e_r$, $e'_b \leq e_b$, and $u(e') \leq u(e)$ with at least one inequality strict. Since $u(e') \leq u(e) \leq \Delta$, the error pair e' must belong to $\mathcal{E}_{\leq\Delta}$. But since e belongs to the Pareto frontier $\mathcal{P}(\mathcal{E}_{\leq\Delta})$, there cannot exist a point $e' \in \mathcal{E}_{\leq\Delta}$ satisfying $e'_r \leq e_r$ and $e'_b \leq e_b$ with either inequality strict. Thus

$$e'_r = e_r \quad e'_b = e_b \quad u(e') < u(e)$$

in contradiction of the definition of $u(e)$.

In the other direction, consider any $e \in \mathscr{F}$ and set $\Delta \equiv u(e)$ so that $e \in \mathcal{E}_{\leq\Delta}$. We will show that $e \in \mathcal{P}(\mathcal{E}_{\leq\Delta})$. Suppose not. Then there exists $e' \in \mathcal{E}_{\leq\Delta}$ such that $e'_r \leq e_r$ and $e'_b \leq e_b$ with at least one inequality strict. But since also $u(e') \leq u(e) = \Delta$, it must be that $e' >_{FA} e$, and we have the desired contradiction. □

Assumption 1 implies that the Pareto frontiers $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ takes either of two forms:

Lemma O.6. *For every $\Delta \geq \underline{\Delta}$,*

(a) If $\mathcal{E}_{\leq\Delta} \cap \mathcal{P}(\mathcal{E}) \neq \emptyset$, then $\mathcal{P}(\mathcal{E}_{\leq\Delta}) = \mathcal{E}_{\leq\Delta} \cap \mathcal{P}(\mathcal{E})$

(b) If $\mathcal{E}_{\leq\Delta} \cap \mathcal{P}(\mathcal{E}) = \emptyset$, then $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ is a singleton.

That is, if the sublevel set $\mathcal{E}_{\leq\Delta}$ has nonempty intersection with the Pareto frontier $\mathcal{P}(\mathcal{E})$, then the Pareto frontier of $\mathcal{E}_{\leq\Delta}$ is precisely this intersection. Otherwise, the Pareto frontier $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ must be a singleton.

Proof. Suppose $\mathcal{E}_{\leq\Delta}$ has nonempty intersection with the accuracy frontier $\mathcal{P}(\mathcal{E})$. This intersection $\mathcal{E}_{\leq\Delta} \cap \mathcal{P}(\mathcal{E})$ must be part of the frontier $\mathcal{P}(\mathcal{E}_{\leq\Delta})$, since if a point e is Pareto undominated within $\mathcal{E}$, it must also be Pareto undominated within the smaller set $\mathcal{E}_{\leq\Delta}$.

Suppose $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ includes a point that does not belong to $\mathcal{P}(\mathcal{E})$. Since the sublevel sets are nested convex polygons (by Lemma O.4), $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ must include an entire line segment not included in $\mathcal{P}(\mathcal{E})$. This line segment must further be negatively sloped, since any Pareto frontier consists exclusively of negatively sloped lines. Choose any point e in the interior of this line segment. Since e is not in $\mathcal{P}(\mathcal{E})$, it must be Pareto dominated by some other point $e' \in \mathcal{E}$. Consider a point e^* between e' to e and arbitrarily close to e. Since the line segment must have negative slope, it must be that $D_{(1,0)}u(e^*) < 0$ and $D_{(0,1)}u(e^*) < 0$. But this contradicts Assumption 1. So $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ cannot include any points outside of $\mathcal{P}(\mathcal{E})$, and we conclude that $\mathcal{P}(\mathcal{E}_{\leq\Delta}) = \mathcal{E}_{\leq\Delta} \cap \mathcal{P}(\mathcal{E})$ as desired.

Now suppose $\mathcal{E}_{\leq\Delta} \cap \mathcal{P}(\mathcal{E}) = \emptyset$. Suppose towards contradiction that $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ is not a singleton. Then $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ consists of negatively sloped line segments. Choose some point e

in the interior of one such line segment. Since $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ does not intersect with $\mathcal{P}(\mathcal{E})$, e must be Pareto dominated by some point $e' \in \mathcal{E}$. By the same argument above, this contradicts Assumption 1. □

Now we can complete the proof of Theorem O.1. First suppose F_X belongs to the accuracy frontier $\mathcal{P}(\mathcal{E})$. Since every $\mathcal{E}_{\leq\Delta}$ for $\Delta \geq \underline{\Delta}$ includes F_X, each sublevel set must have nonempty intersection with $\mathcal{P}(\mathcal{E})$. Applying Lemma O.5 and Part (a) of Lemma O.6, each $\mathcal{P}(\mathcal{E}_{\leq\Delta})$ is a subset of $\mathcal{P}(\mathcal{E})$, and we recover all of $\mathcal{P}(\mathcal{E})$ as we vary over Δ. So $\mathscr{F} = \mathcal{P}(\mathcal{E})$.

Next suppose F_X does not belong to the accuracy frontier $\mathcal{P}(\mathcal{E})$. Define

$$U \equiv \{\Delta \mid \mathcal{E}_{\leq\Delta} \cap \mathcal{P}(\mathcal{E}) = \emptyset\}$$

be the unfairness levels Δ for which the sublevel set $\mathcal{E}_{\leq\Delta}$ have empty intersection with $\mathcal{P}(\mathcal{E})$. For any $\Delta \in U^c$, the previous arguments apply and show that the full accuracy frontier $\mathcal{P}(\mathcal{E})$ is again recovered as part of the fairness-accuracy frontier $\mathscr{F}$.

For any $\Delta \in U$, Part (b) of Lemma O.6 implies that the accuracy frontier in this sublevel set is a singleton, and hence can be characterized as the point $A_\Delta = \arg\min_{e\in\mathcal{E}_{\leq\Delta}} e_r$, where the choice of group r is arbitrary. The sublevel set $\mathcal{E}_{\leq\Delta}$ is convex and compact for each $\Delta \in U$, and $\mathcal{E}_{\leq\Delta}$ is continuous at each $\Delta \in U$ by continuity of $u(e)$. By the theorem of the maximum, A_Δ is continuous in Δ, so the set $\{A_\Delta\}_{\Delta\geq 0}$ is connected. By Lemma O.4, this path consists of a sequence of line segments. Moreover, since the sets $\mathcal{E}_{\leq\Delta}$ are nested, and the point A_Δ simultaneously minimizes e_r and e_b within the set $\mathcal{E}_{\leq\Delta}$, these points must move weakly down and left as Δ increases, so the path consists of a sequence of positively sloped line segments. Thus the fairness-accuracy frontier $\mathscr{F}$ is the union of the accuracy frontier $\mathcal{P}(\mathcal{E})$ and a sequence of positively sloped line segments connecting $\mathcal{P}(\mathcal{E})$ to F_X, as desired.

O.2 General Fairness Criteria

In this section, we consider the general case where fairness is evaluated using $|\phi(e_r) - \phi(e_b)|$ for some strictly increasing continuous function ϕ. For instance, if ϕ is log, then this reduces to using the ratio of error rates as a measure of fairness. The characterization of the fairness-accuracy frontier remains the same except the fairness optimal point F_X may now be different. Whether it expands or contracts depends on the curvature of ϕ as the following Proposition demonstrates.[41]

[41] We assume that the accuracy and fairness loss functions are the same but can generalize the results in this section via the same methodology as in Section O.1.

Proposition O.1. *Let $\mathscr{F}'(X)$ denote the fairness-accuracy frontier where fairness is evaluated using*

$$|\phi(e_r) - \phi(e_b)|$$

for strictly increasing $\phi : \mathbb{R} \to \mathbb{R}$. Then

1. *$\mathscr{F}(X) = \mathscr{F}'(X)$ if X is group-balanced*
2. *$\mathscr{F}(X) \subset \mathscr{F}'(X)$ if X is group-skewed and ϕ in concave*
3. *$\mathscr{F}(X) \supset \mathscr{F}'(X)$ if X is group-skewed and ϕ in convex*

Proof. Let $\mathcal{E}(X)$ and $\mathcal{E}'(X)$ denote the feasible sets where fairness is defined using $|e_r - e_b|$ and $|\phi(e_r) - \phi(e_b)|$ respectively. Let F_X and F'_X denote the corresponding fairness optimal points. First, note that if X is group-balanced, then by the same argument as Theorem 1, $\mathscr{F}(X) = \mathscr{F}'(X)$ is the lower boundary from $R_X = R'_X$ to $B_X = B'_X$.

Now, suppose X is r-skewed without loss. Let e and e' correspond to F_X and F'_X so

$$\begin{aligned} e_b - e_r &\leq e'_b - e'_r \\ \phi(e'_b) - \phi(e'_r) &\leq \phi(e_b) - \phi(e_r) \end{aligned}$$

First, suppose ϕ is concave. We will show that $e'_r \geq e_r$. Suppose by contradiction that $e'_r < e_r$ so $\phi(e'_r) < \phi(e_r)$. Thus,

$$\phi(e'_b) - \phi(e_b) \leq \phi(e'_r) - \phi(e_r) < 0$$

so $e'_b < e_b$. Thus, we have $e'_r \leq e'_b < e_b$. Note that

$$e'_b = \lambda e_b + (1 - \lambda) e'_r$$

where

$$\lambda := \frac{e'_b - e'_r}{e_b - e'_r}$$

We thus have

$$\begin{aligned} \phi(e_b) - \phi(e_r) + \phi(e'_r) &\geq \phi(e'_b) = \phi(\lambda e_b + (1 - \lambda) e'_r) \\ &\geq \lambda \phi(e_b) + (1 - \lambda) \phi(e'_r) \\ (1 - \lambda)(\phi(e_b) - \phi(e'_r)) &\geq \phi(e_r) - \phi(e'_r) \\ (e_b - e'_b) \frac{\phi(e_b) - \phi(e'_r)}{e_b - e'_r} &\geq \phi(e_r) - \phi(e'_r) \end{aligned}$$

where the second inequality follows from the fact that ϕ is concave. Since $e_r - e'_r \geq e_b - e'_b$, this implies

$$\frac{\phi(e_b) - \phi(e'_r)}{e_b - e'_r} \geq \frac{\phi(e_r) - \phi(e'_r)}{e_r - e'_r}$$

Since X is r-skewed, $e_b \geq e_r > e'_r$. Since ϕ is concave, the above inequality must be satisfied with equality. This means that

$$(e_b - e'_b)\frac{\phi(e_b) - \phi(e'_r)}{e_b - e'_r} \geq \phi(e_r) - \phi(e'_r) = (e_r - e'_r)\frac{\phi(e_b) - \phi(e'_r)}{e_b - e'_r}$$

so $e_b - e'_b = e_r - e'_r$ or $e_b - e_r = e'_b - e'_r$. But e corresponds to F_X and since e' achieves the same fairness as e, it must be that $e_r \leq e'_r$. This contradicts our assumption that $e'_r < e_r$. Thus, $e'_r \geq e_r$ and by the same argument characterizing the FA frontier as in Theorem 1, $\mathscr{F}(X) \subset \mathscr{F}'(X)$. The case for when ϕ is convex is symmetric. □

O.3 Adversarial Agents

We now consider the problem outlined in Section 4, when one of the weights α_r, α_b is negative.[42] 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.

Lemma O.7. *For every covariate vector X, $\mathcal{E}^*(X) = \mathcal{E}(X) \cap H$.*

But the analogous equivalence for the FA 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 O.6. Covariate vector X is:

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

- *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 O.2. *The input-design fairness-accuracy (FA) frontier $\mathscr{F}^*(X)$ is the lower boundary of the input-design feasible 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 4.2.1.

Definition O.7. X is *strictly input-design-group-balanced* if $e_r < e_b$ at R_X^* and $e_b < e_r$ at B_X^*.

Proposition O.2. *Suppose $\alpha_r > 0 > \alpha_b$ and X is strictly input-design group-balanced. Then excluding G over X uniformly worsens the frontier.*

This result says that, perhaps 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. The notion of group-balanced covariate vectors, suitably adapted to the input design setting, again serves as a sufficient condition for uniform worsening of the frontier when excluding G.

Proof. By assumption that X is strictly input-design group-balanced, the input-design FA frontier given X is the lower boundary of $\mathcal{E}^*(X)$ from R_X^* to B_X^*, which consists of negatively sloped edges. We will show that every point on this frontier is FA-dominated by some point in $\mathcal{E}^*(X, G)$.

If this point (e_r, e_b) is distinct from B_X^* and R_X^*, then we claim that for sufficiently small positive ϵ, the point $(e_r - \epsilon, e_b - \epsilon)$ belongs to $\mathcal{E}^*(X, G)$. Indeed, $(e_r - \epsilon, e_b - \epsilon)$ belongs to the unconstrained feasible set $\mathcal{E}(X, G)$ because this feasible set is a rectangle, and $e_r - \epsilon$, $e_b - \epsilon$ are within the minimal and maximal group errors achievable given X. Moreover, (e_r, e_b) must have smaller group-r error and larger group-b error compared to B_X^*, which means the same is true for $(e_r - \epsilon, e_b - \epsilon)$. Since $\alpha_r > 0 > \alpha_b$, the point $(e_r - \epsilon, e_b - \epsilon)$ must belong to H given that B_X^* does. Hence when (e_r, e_b) differs from B_X^* and R_X^*, it is FA-dominated by $(e_r - \epsilon, e_b - \epsilon) \in \mathcal{E}^*(X, G)$.

Suppose now that $(e_r, e_b) = B_X^*$. Then by similar argument it is FA-dominated by $(e_r - \epsilon, e_b) \in \mathcal{E}^*(X, G)$. Finally if $(e_r, e_b) = R_X^*$, then it is FA-dominated by $(e_r, e_b - \epsilon) \in \mathcal{E}^*(X, G)$. In all these cases the FA frontier uniformly worsens when excluding G, completing the proof. □

O.4 Conditional Independence

Section 3.3 considers the case where group identity is an input. In this section, we consider the case where covariates satisfy a more general property called conditional independence. Our results in this section hold for all loss functions that are group independent, i.e. $\ell(d, y, g) = \ell(d, y)$.

Definition O.8. X satisfies *conditional independence* if $G \perp\!\!\!\perp Y \mid X$.

Under conditional independence, the covariate vector X contains all of the information in group identity that is relevant for predicting Y. In other words, once the algorithm has conditioned on X, there is no additional predictive value to knowing group identity. Note that if X reveals G, then X is conditionally independent.

We first characterize the fairness-accuracy frontier under conditional independence.

Proposition O.3. *Suppose X is conditionally independent. Then $\mathscr{F}(X)$ is that part of the lower boundary of the feasible set ranging from the point $B_X = R_X$ to the point F_X.*

Proof. We will show that $B_X = R_X$ under conditional independence. Recall from the proof of Lemma B.1 that

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

where

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

Under conditional independence, $x_{y,g} = x_y x_g$ so we have

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

This means that for each realization $x \in \mathcal{X}$, the outcome that gives the lower error for group r also gives the lower error for group b. In other words, when $\sum_y x_y \ell(1, y) \leq$

$\sum_y x_y \ell(0,y)$, then outcome $Y=1$ is optimal for both groups (and vice-versa for the other outcome). Consider the following algorithm:

$$f(x) = \begin{cases} 1 & \text{if } \sum_y x_y \ell(1,y) \leq \sum_y x_y \ell(0,y) \\ 0 & \text{if } \sum_y x_y \ell(1,y) > \sum_y x_y \ell(0,y) \end{cases}$$

This algorithm will deliver the lowest error for both groups and

$$(e_r(f), e_b(f)) = R_X = B_X$$

as desired. □

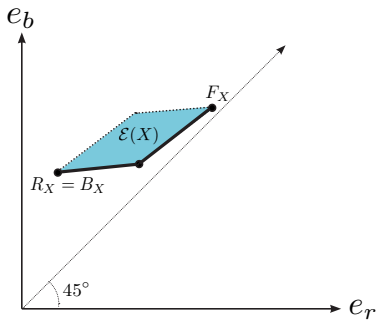


Figure 9: Depiction of the fairness-accuracy frontier under assumption of conditional independence of G and Y.

Figure 9 depicts an example of a fairness-accuracy frontier for a covariate vector satisfying Conditional Independence. 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 fairness-accuracy 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. We generalize this in the corollary below.[43]

[43] In the special case when $R_X = B_X = F_X$, the fairness-accuracy frontier is just a singleton, and there is no strong fairness-accuracy conflict. (Corollary O.1 is vacuous in this case, since there are no two distinct points on the fairness-accuracy frontier.)

Corollary O.1. *Suppose X is conditionally independent. Then any two points in $\mathscr{F}(X)$ exhibit 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 fairness-accuracy 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 fairness-accuracy frontier is the lower boundary of $\mathcal{E}(X)$ from R_X to F_X. Since $R_X = B_X$, the fairness-accuracy 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. □

Finally, we consider another special case of conditional independence when covariate vectors satisfy the following strong independence condition:

Definition O.9. Say that X satisfies *strong independence* if for both groups g,

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

In this case, the feasible set turns out to be a line segment on the 45-degree line, and the fairness-accuracy frontier is a single point, as depicted in Figure 10.

Proposition O.4. *Suppose X is strongly independent. Then the fairness-accuracy frontier is a single point on the 45-degree line.*

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

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

Thus $\frac{x_{y,r}}{p_r} = \frac{x_{y,b}}{p_b}$ for all x, y. It follows that the line segment $E(x)$, which connects the two points $\left(\sum_y \frac{x_{y,r}}{p_r}\ell(1,y), \sum_y \frac{x_{y,b}}{p_b}\ell(1,y)\right)$ 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 FA 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 FA-dominance principle outlined in Definition 2 prefer the same policy.

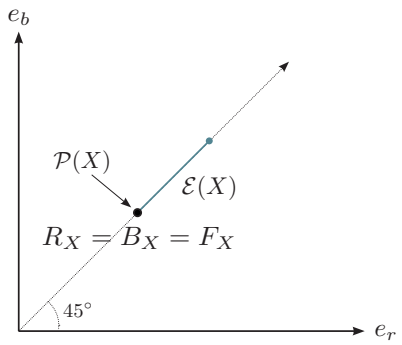


Figure 10: Depiction of the fairness-accuracy frontier under assumption of strong independence

O.5 Microfoundations for the FA frontier

We now provide a foundation of our FA frontier as the optimal points for different classes of FA preferences.[44] First, consider 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 the corresponding preference of this utility *simple*. Simple preferences are FA preferences. For example, both the Utilitarian and Rawlsian preferences 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 $\alpha_r = \alpha_b = \alpha_f = -1$.

Given any FA preference $\succeq$, let

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

denote the set of $\succeq$-optimal points. We now provide the following characterizations of the FA frontier.[45]

[44] Note that we could have alternatively defined FA preferences to be weakly decreasing in e_r, e_b and $|e_r - e_b|$. The equivalence of (1), (3) and (4) in Proposition O.5 would still hold.

[45] The proof of the equivalence of (1) and (4) in Proposition O.5 relies on finite X. The other parts do not.

Proposition O.5. *The following are equivalent:*

1. $e \in \mathscr{F}(X)$
2. $e \in \mathscr{F}_{\succeq}(X)$ *for some FA preference* $\succeq$
3. $\{e\} = \mathscr{F}_{\succeq}(X)$ *for some FA preference* $\succeq$
4. $e \in \mathscr{F}_{\succeq}(X)$ *for some simple FA preference* $\succeq$

The above result shows that our FA frontier is the set of all optimal points for all FA preferences. Moreover, $\mathscr{F}(X)$ is minimal in the sense that we cannot exclude any points from $\mathscr{F}(X)$ without hurting some designer. This is because for every point $e \in \mathscr{F}(X)$, there exists some FA preference $\succeq$ such that e is the *unique* optimal error pair given $\succeq$ within the feasible set $\mathcal{E}(X)$. Finally, our FA frontier also corresponds to the optimal points for all simple FA preferences.

Proof. We will first show that (3) implies (2) implies (1) implies (3). Note that (3) implies (2) is trivial. To see why (2) implies (1), suppose $e \in \mathscr{F}_{\succeq}(X)$ for some FA preference $\succeq$ but $e \notin \mathscr{F}(X)$. Thus, there exists some $e' >_{FA} e$ so $e' \succ e$ yielding a contradiction.

We now prove that (1) implies (3). Fix some $e^* \in \mathscr{F}(X)$ and let $h : \mathbb{R} \to (0,1)$ be a strictly decreasing function. Define

$$w(e) = \begin{cases} 1 + h(e_r + e_b) & \text{if } e = e^* \text{ or } e >_{FA} e^* \\ h(e_r + e_b) & \text{otherwise} \end{cases}$$

and let $\succeq$ be the corresponding preference. We will show that $\succeq$ is an FA preference. Suppose $e >_{FA} e'$ so $h(e_r + e_b) > h(e'_r + e'_b)$. If both points FA-dominate e^* or neither do, then $w(e) > w(e')$. The only remaining case is when $e >_{FA} e^*$ but e' does not FA-dominate e^*, in which case

$$w(e) = 1 + h(e_r + e_b) > 1 > h(e'_r + e'_b) = w(e')$$

Thus, $\succeq$ is an FA preference. Now, since $e^* \in \mathscr{F}(X)$, there exists no other $e \in \mathcal{E}(X)$ such that $e >_{FA} e^*$. That means that for all $e \in \mathcal{E}(X) \setminus \{e^*\}$, $w(e^*) > w(e)$ so $\{e^*\} = \mathscr{F}_{\succeq}(X)$. This proves (3).

Finally, we show the equivalence of (1) and (4). Note that (4) implies (2) which implies (1) from above. We now show that (1) implies (4). Fix some $e^* \in \mathscr{F}(X)$, so 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 $e^* \in \mathscr{F}(X)$ 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. □