ALGORITHM DESIGN: A FAIRNESS-ACCURACY FRONTIER

ANNIE LIANG†, JAY LU§, XIAOSHENG MU‡, AND KYOHEI OKUMURA*

June 26, 2025

Abstract. Algorithm designers increasingly optimize not only for accuracy, but also for the fairness of the algorithm across pre-defined groups. We study the tradeoffs between fairness and accuracy induced by a given set of inputs to the algorithm. We formalize these tradeoffs as a fairness-accuracy frontier, defined as the set of outcomes that cannot be simultaneously improved upon in both fairness and accuracy. Our results show that the shape of the frontier is determined by a simple property of the inputs, which we call *group-skew*. Group-skewed inputs inherently advantage one group, resulting in lower errors for that group even when the algorithm is optimized for the other. We show that decreasing accuracy for both groups in order to increase fairness can be justified by fairness considerations if and only if inputs are group-skewed. We further study an information design problem where a designer flexibly regulates the inputs, but another agent chooses the algorithm. The optimal regulation of inputs generally depends on the designer's preferences, but when inputs are not group-skewed then two implications hold across all designer preferences: (1) banning group identity is strictly suboptimal, and (2) if group identity is available, then banning any informative input is strictly suboptimal.

1. Introduction

Suppose a hospital uses a machine learning algorithm to aid in the diagnosis of a medical condition, where the algorithm makes the correct diagnosis 90% of the time for Red patients but only 50% of the time for Blue patients. Such an outcome—where the consequences of a policy differ systematically across groups—is known as *disparate impact*. Across a wide range of applications, algorithms have been shown to have disparate impact (Arnold, Dobbie, and Hull, 2021; Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2021). For example,

We thank Nageeb Ali, Eric Auerbach, Simon Board, Krishna Dasaratha, Will Dobbie, Alex Frankel, Ben Golub, Sergiu Hart, Peter Hull, Navin Kartik, Yair Livne, Sendhil Mullainathan, Derek Neal, Jose Montiel Olea, Larry Samuelson, Max Tabord-Meehan and five anonymous referees for helpful comments. Annie Liang thanks the National Science Foundation Grant SES-2145352 for financial support. We also thank Andrei Iakovlev and Aristotle Magganas for valuable research assistance on this project.
†Northwestern University, liang.annie.h@gmail.com.
§UCLA, jay@econ.ucla.edu.
‡Princeton University, xmu@princeton.edu.
*Northwestern University, kyohei.okumura@gmail.com.

patients assigned the same risk score by a healthcare algorithm were shown to have substantially different actual health risks depending on their race (Obermeyer, Powers, Vogeli, and Mullainathan, 2019); 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); and the accuracies of facial-recognition technologies vary substantially across demographic groups (Klare et al., 2012). These findings have led algorithm designers to impose group-based fairness constraints in settings ranging from healthcare to bail evaluation to lending (Roth and Kearns, 2019; Hardt et al., 2016).

Policymakers prefer for algorithms to have both lower disparate impact and also higher accuracy. In an ideal world, both goals could be achieved simultaneously; in practice, there may be an intrinsic tradeoff between accuracy (the overall error rate of the algorithm) and fairness (how similar the error rates are across pre-defined groups).[1] What this tradeoff looks like depends on the inputs to the algorithm, which can be observed, manipulated, and regulated—raising the following questions: How does the tradeoff between fairness and accuracy depend on the information available for prediction? Which informational environments create a tension between fairness and accuracy, and which ameliorate it? While the tradeoff between fairness and accuracy has been empirically computed in specific applications (Wei and Niethammer, 2022; Chohlas-Wood et al., 2024; Little et al., 2022), we know substantially less about how the available information shapes the tension between these two objectives in general.

In this paper, we address these questions by defining and studying a *fairness-accuracy frontier*. This frontier consists of all outcomes that cannot be simultaneously improved upon in fairness and accuracy. We prove two types of results about the frontier. First, we identify simple properties of the algorithmic inputs that determine the shape of this frontier. Second, we take an information design perspective to understand how constraints on information can induce certain desired outcomes. Specifically, we consider an interaction between a policymaker who flexibly constrains the inputs and an agent who sets the algorithm. We characterize what part of the fairness-accuracy frontier the designer can achieve through appropriate design of the inputs and examine whether it might be optimal for the designer

[1]Equity-efficiency tradeoffs such as this have been studied in settings as diverse as taxation (Saez and Stantcheva, 2016; Dworczak, Kominers, and Akbarpour, 2021), policing (Persico, 2002; Jung, Kannan, Lee, Pai, Roth, and Vohra, 2020), and college admissions (Chan and Eyster, 2003; Ellison and Pathak, 2021).

to exclude an input altogether (e.g., excluding group identity in the context of medical predictions).

In our model, a designer chooses an algorithm that takes observed covariates as inputs (e.g., medical scans, lab tests, records of prior hospital visits) and outputs a decision (e.g., whether to recommend a medical procedure). The algorithm's consequences for any given individual are evaluated using a loss function, which can be interpreted as the inaccuracy or harm of the decision. We aggregate losses within two groups, group r (Red) and group b (Blue). Each group's *error* is the expected loss for individuals of that group, and the designer evaluates algorithms according to a preference over these group errors.

We define the class of *fairness-accuracy (FA) preferences* to be all preferences over group errors 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).[2]

This strict 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 absolute difference between group errors) among others. Some of these preferences also correspond directly to optimization problems that have been proposed for use in practice.[3] We define the *fairness-accuracy frontier* as the set of all group error pairs that are feasible (i.e., can be implemented by some algorithm given the observed covariates), and are moreover FA-undominated within the feasible set. That is, these error pairs cannot be improved upon simultaneously in 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-skewed* if it inherently advantages one group, giving that group lower errors even when the algorithm is optimized for the other group. Otherwise say that the covariate vector is *group-balanced.*

[2]An alternative definition of FA-dominance might rank one pair of group errors over another if the first involves a smaller aggregate error (weighting each group by its size) and also a smaller absolute difference between group errors. Relative to this, our definition of FA-dominance is consistent with a larger class of preferences. For example, any "generalized utilitarian" rule which ranks pairs of group errors based on positive weights for aggregation (possibly different from population size) will belong to our class of FA preferences but may not belong to this alternative class.

[3]For example, optimizing a Rawlsian preference is equivalent to implementing group distributionally robust optimization (Sagawa et al., 2020), and optimizing an Egalitarian preference is equivalent (on a restricted domain) to maximizing accuracy subject to equality of error rates (as considered in Hardt et al. (2016) among others).

While it may be difficult to anticipate in advance of a comprehensive empirical analysis which of group-balance or group-skew is more typical in practice, one scenario in which the latter may arise is if covariates have the same implications for both groups but are measured more accurately for one group than the other (e.g., medical data is recorded more accurately for high-income patients than low-income patients).

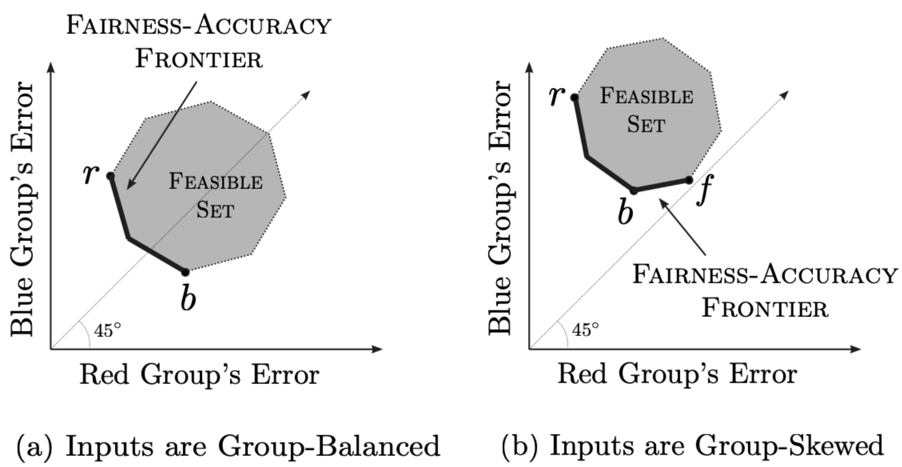


Figure 1. The Fairness-Accuracy Frontier.

We provide two complementary characterizations of the fairness-accuracy frontier. Our first and main characterization says that the fairness-accuracy frontier takes one of two possible forms depending on whether the covariate vector is group-balanced or group-skewed. When inputs are group-balanced, the fairness-accuracy frontier is exactly the standard Pareto frontier, i.e. the set of all feasible error pairs that cannot be simultaneously reduced in both coordinates. Within the feasible set of error pairs, this is the part of the lower boundary of the feasible set that begins at the point that is best for group Red (labeled r) and ends at the point that is best for group Blue (labeled b), as depicted in Panel (a) of Figure 1 above. Here, the tradeoff is between accuracy for one group and accuracy for the other group. On the other hand, when inputs are group-skewed, the fairness-accuracy frontier includes not only the standard Pareto frontier but additionally a positively-sloped part (the segment from b to the fairness-maximizing point f in Figure 1) along which both groups' errors increase but the gap between them decreases. Here, the tradeoff is between the accuracy of *both* groups and fairness, indicating a strong conflict between the two objectives. This characterization implies that increasing errors for both groups cannot be justified by fairness considerations if the covariate vector is group-balanced, but potentially can if the covariate vector is group-skewed. Our second characterization describes a class of algorithms that implement the error pairs along the fairness-accuracy frontier. These algorithms can be described as possessing a simple threshold structure when the covariate vectors are ordered in a particular way, where moving along the fairness-accuracy frontier corresponds to shifting the location of the threshold.

In the second part of the paper, we investigate what happens if an agent different from the designer chooses the algorithm, while the designer regulates the inputs of the algorithm. We suppose that the designer has fairness concerns, while the agent setting the algorithm does not. For example, a healthcare provider (agent) determining treatment may seek to maximize the number of correct diagnoses, while a policymaker (designer) may additionally prefer that the accuracy of the provider's treatments be equitable across certain social groups. In these cases, the policymaker can impose regulations that restrict the inputs available to the algorithm, for example by banning the use of a specific input.

We model the designer's choice of input regulation 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 ask whether the optimal garbling might involve completely excluding a covariate, and demonstrate two results: First, excluding group identity as an algorithmic input is strictly welfare-reducing for all designers (with a preference in our fairness-accuracy class) if and only if the permitted covariates are group-balanced. 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 mild condition that we call *decision-relevance.* This result applies, for instance, to the question of whether to exclude standardized test scores in university admissions decisions. It suggests that when group identity is a permissible input in admission decisions, then excluding test scores reduces welfare for *all* designers with the power to flexibly garble covariates. On the other hand, when group identity is not permitted as an input in college admissions decisions (as is now the case in the United States following the Supreme Court decision in *Students for Fair Admissions, Inc. v. President and Fellows* *of Harvard College*), then the optimal garbling of covariates for some designer preference may indeed involve completely excluding test scores, and we provide a simple example to illustrate this effect (Section 4.3.1).

Finally, we illustrate our key definitions and results using two popular healthcare datasets from the algorithmic fairness literature. We conduct two analyses: First, we evaluate whether the covariates in these datasets are group-balanced or group-skewed (finding that one set of covariates is group-skewed and the other is an interesting case of group-balance in which the group-optimal points lie on the 45 degree line). Second, we depict the fairness-accuracy frontier. This allows us to study the limits of input design for each of the datasets (input design turns out to be without loss for the designer in one dataset but not in the other) and also to examine how the frontier changes when group-specific algorithms are permitted. These empirical illustrations of our framework speak to its potential relevance for assessing fairness-accuracy tradeoffs in practical algorithm design problems.

1.1. **Related Literature.** Our paper is motivated by recent problems regarding algorithmic bias (Section 1.1.3), but adopts a novel perspective on these questions based on approaches from two literatures in economic theory: the literature on information design (Section 1.1.1) and the literature on social preferences and inequality (Section 1.1.2). Building on the former, we model the interaction between a designer flexibly regulating inputs and an agent setting the algorithm. Building on the latter, we focus on understanding equity-efficiency tradeoffs, and consider a wide class of preferences that reflects heterogeneity in social preferences.

1.1.1. *Information Design.* One contribution of our paper is the modeling of the design of algorithmic inputs as an information design problem (see Kamenica (2019) and Bergemann and Morris (2019) for recent surveys). This approach complements previous frameworks for modeling algorithm regulation, in which policymakers communicate information via cheap talk (Cowgill and Stevenson, 2020) or impose restrictions directly on the algorithm (Yang and Dobbie, 2020; Rambachan, Kleinberg, Mullainathan, and Ludwig, 2021; Blattner, Nelson, and Spiess, 2024). It also complements Doval and Smolin (2024)'s characterization of the feasible set of welfare profiles for heterogeneous types in the population as the information policy varies.[4] We view the garbling of inputs as a potentially effective policy tool that can

[4]For example, Doval and Smolin (2024) show that excluding inputs is suboptimal in the sense that more information necessarily increases the feasible set of payoffs. In contrast, in our model it may be strictly optimal for the designer to exclude an input, since a different agent chooses which payoff vector is implemented from among our feasible set.

be implemented through various technological or legal commitments,[5] and which deserves further attention within the context of algorithmic fairness.

Our focus on algorithmic fairness motivates an analysis that departs from typical information design problems in several interesting ways. For example, the Sender in our framework cannot choose a completely flexible information structure, but is instead constrained to garblings of a primitive covariate vector. Additionally, motivated by heterogeneous attitudes toward fairness (Section 1.1.2), we focus on a frontier of solutions with respect to a wide class of Sender preferences. Our results in Section 4.3 describe how the frontier of solutions responds to changes in the underlying information. We focus on special cases of this comparative static that are of interest given our motivation (e.g., adding or removing a covariate), but a more general solution (analogous to Curello and Sinander (2024)'s work on comparative statics with respect to the Sender's utility function) would be an interesting avenue for future work.

Finally, at the broader intersection of information design and algorithms, Ichihashi (2023) considers optimal information acquisition for crime deterrence, and Caplin, Martin, and Marx (2024) draws a connection between different machine learning objectives and costly information design.

1.1.2. *Social Preferences and Inequality.* The literature on social preferences documents substantial heterogeneity in how individuals assess efficiency-equity tradeoffs (Fehr and Schmidt, 1999; Andreoni and Miller, 2002; Fisman, Kariv, and Markovits, 2007; Sullivan, 2023), which is reflected in our broad class of FA-preferences. In this literature, social preferences are preferences over individual payoffs rather than preferences over group errors, but most have analogues in our setting. For example, the "social welfare approach" aggregates individual payoffs using differential weights (Charness and Rabin, 2002; Saez and Stantcheva, 2016; Dworczak, Kominers, and Akbarpour, 2021), and is nested in our class of FA preferences if we interpret individual payoffs as group errors. We additionally allow for a direct penalty for unequal outcomes, as in Loewenstein et al. (1989), Fehr and Schmidt (1999) and Bolton and Ockenfels (2000)'s inequity aversion models.

[5]For example, organizations such as the US Census Bureau, Google, Apple, and Microsoft are committed to differential privacy initiatives (Dwork and Roth, 2014), which take various forms of adding noise to user inputs. Yang and Dobbie (2020) summarizes the existing law on algorithmic regulation and proposes new legal policies for mitigating algorithmic bias.

There is a separate literature studying the equity-efficiency tradeoffs of affirmative action programs. Lundberg (1991) and Chan and Eyster (2003) model affirmative action as a ban on the use of group identity in admissions decisions, showing that this can lead organizations to condition on proxies in a way that reduces both efficiency and equity.[6] (A similar point is made in Agan and Starr (2018) regarding the use of prior criminal history in hiring decisions in "ban-the-box" policies.) Ellison and Pathak (2021) empirically quantify the equity and efficiency losses of race-neutral affirmative action (based on geographic proxies for race) as compared to plans that explicitly consider race. While these papers relate to our study of the impact of excluding group identity, they focus on comparing a designer's optimal algorithm with and without access to group identity. We instead examine how the frontier of feasible outcomes changes when the designer can design a group-dependent garbling versus when the designer must choose a group-independent garbling. These analyses are not nested; see Section 4.3.2 for more detail.

1.1.3. *Algorithmic Bias.* The recent literature on algorithmic bias has emerged around the concern that algorithms have error rates that differ substantially across social and demographic groups (see Kleinberg et al. (2018) and Cowgill and Tucker (2020) for overviews). In this literature and in the accompanying policy discussion (e.g, Angwin and Larson (2016)), algorithms are often considered to be "less fair" if their harms are more unequally borne across groups, with this comparison formalized as a larger disparity in error rates across groups (Hardt et al., 2016; Kleinberg et al., 2017; Chouldechova, 2017).[7] A growing body of empirical work documents and quantifies these disparate impacts (Obermeyer et al., 2019; Arnold et al., 2021; Fuster et al., 2021).

The tradeoff between accuracy (error rates of the algorithm for each group) and fairness (discrepancy between error rates across groups) is a special kind of equity-efficiency tradeoff. A common approach for resolving this tradeoff is to posit a particular objective criterion (Hardt et al., 2016; Diana et al., 2021). Other papers identify improvements with respect to both objectives simultaneously (Rose, 2021; Feigenberg and Miller, 2021). Our paper is closest to a smaller part of this literature, which engages with the tension between fairness

[6]Another set of papers shows that access to group identity must weakly improve the designer's payoffs when the designer has control of the algorithm (see for example Menon and Williamson (2018), Agarwal et al. (2018), Lipton et al. (2018), Rambachan et al. (2021), and Manski et al. (2023)), as adding group identity is a Blackwell improvement in information. This is no longer guaranteed to be the case when the designer cannot choose the algorithm, as in our model in Section 4.

[7]There are exceptions, for example the concept of individual fairness proposed in Dwork et al. (2012).

and accuracy by quantifying fairness-accuracy tradeoffs for specific loss functions (Menon and Williamson, 2018) or for specific empirical applications (Wei and Niethammer, 2022; Little et al., 2022; Chohlas-Wood et al., 2024). We are interested in how this fairness-accuracy tradeoff is moderated by the inputs to the algorithm in general, and provide simple conditions on the inputs that qualitatively govern this tradeoff independently of other details of the loss function or informational environment.

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 a finite set $\mathcal{X}$, a *type* Y taking values in a finite set $\mathcal{Y}$, and a *group identity* G taking the value r or b.[8] The definition of the relevant groups is a primitive of the setting (for example, determined by sociopolitical precedent) and outside of our framework. We model G, X, Y as random variables with joint distribution $\mathbb{P}$, and use $p_g := \mathbb{P}(G = g) > 0$ to denote the fraction of the population that belongs to group $g \in \{r, b\}$. We impose no additional assumptions on $\mathbb{P}$,[9] permitting all of the following special cases.

Example 1 (X reveals or closely proxies for G). The group identity G is either itself an input in the covariate vector X, or is predictable from inputs in the covariate vector X. For example, Bertrand and Kamenica (2023) 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.

Example 2 (Biased Covariates). The value of an input in X may be systematically biased depending on group identity. For instance, if G denotes an individual's income bracket, Y represents their ability, and X is a test score influenced by access to test preparation resources, then 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

[8] We assume finiteness to simplify various notations in the exposition, but most of our results generalize to infinite covariate values and/or infinite types.

[9] 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 consider training data that differs in distribution from the data on which the algorithm's errors are evaluated. For example, one could study the optimal sampling of data on which to train the algorithm, or to study feedback loops when the algorithm is trained on data determined by previous algorithms (as in Che et al. (2019) and Jung et al. (2020)).

(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. For example, in Obermeyer et al. (2019), a patient's health care costs are more predictive of their health care needs for White patients than for Black patients, and Rothstein (2004) shows that SAT scores are more informative about future college grades for high-income students than low-income students.

There is a set $\mathcal{A}$ of permissible mappings from the set of covariate vectors $\mathcal{X}$ to the set of decisions $\mathcal{D} = \{0, 1\}$. The designer chooses from $\Delta(\mathcal{A})$, the set of all randomizations over $\mathcal{A}$. We refer to any $a \in \Delta(\mathcal{A})$ as an *algorithm*.[10] Most of our subsequent results hold for arbitrary choices of the permissible set $\mathcal{A}$; some of our results hold specifically for the leading special case where $\mathcal{A}$ is unconstrained, denoted by $\overline{\mathcal{A}}$, and we will be explicit when this is the case.

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 will reoffend, 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.

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.

[10]This mapping a is often referred to as a classifier or prediction rule in machine learning (where its objective is typically to predict an unknown outcome), a treatment rule or policy in econometrics (where its objective is typically to assign a treatment) or a decision rule in statistical decision theory (where X is interpreted as a standard Blackwell signal). We will refer to a as an "algorithm" to encompass these possibilities and to evoke our motivating application of algorithmic fairness.

The consequence of choosing decision d for an individual whose true type is y is evaluated using a loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$, which we view as a measure of inaccuracy independent of fairness. We further aggregate these losses across individuals within each group.

Definition 1. For any algorithm $a \in \Delta(\mathcal{A})$ and group $g \in \{r, b\}$, the *group g error* is

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

That is, group g's error is the average loss for members of group g. For example, if the type Y is binary and we consider the misclassification loss function $\ell(d, y) = \mathbb{1}(d \neq y)$, then $e_g(a)$ is the probability of making any classification error—type I or type II—for individuals in group g. Other loss functions may put different weights on different kinds of errors. We view the choice of the right loss function as application-specific, and demonstrate results that hold for arbitrary ℓ.

Our notion of "loss" for a group measures the inaccuracy of the decisions for that group from the perspective of the designer, which may or may not coincide with welfare for members of that group. In applications such as medical treatment (which we consider in our empirical application), we would expect the two notions to coincide. But in the context of, for example, job hiring, it may be that only hiring high quality applicants minimizes the expected loss of the designer, but does not maximize the welfare of the applicants (e.g., if all of them prefer to be hired). In Online Appendix O.1, we consider the case where there are separate loss functions for both accuracy and fairness, allowing for cases where accuracy reflects a social welfare notion while fairness compares welfare across groups.

Each algorithm a implies a pair of group errors $e(a) = (e_r(a), e_b(a))$. We say there is an *improvement in accuracy* if both $e_r(a)$ and $e_b(a)$ decrease, and an *improvement in fairness* if $|e_r(a) - e_b(a)|$ decreases.[11] Evaluating fairness based on $|e_r(a) - e_b(a)|$ nests many popular fairness criteria, such as the ones surveyed in Mehrabi et al. (2022), through an appropriate specification of the loss function. See Online Appendix O.5 for detailed discussion.

Section 6 discusses several extensions of our framework, including generalizing to fairness criteria of the form $|\phi(e_r) - \phi(e_b)|$ (which includes the ratio of error rates as a special case), and allowing for fairness and accuracy to be defined using different loss functions.

[11] This approach aligns with much of the algorithmic fairness literature. However, it does not capture all important fairness considerations. For example, perfect predictive accuracy may still fail to address historical inequities reflected in differing base rates of Y across groups. Furthermore, as Kasy and Abebe (2021) argue, an algorithm considered fair in one decision context may perpetuate broader inequalities. It will be interesting to study how these algorithmic design decisions might impact decisions in a larger dynamic game.

2.2. **Fairness-Accuracy Preferences.** Recall the usual Pareto dominance order.

Definition 2. The *Pareto dominance (PD)* relation $>_{PD}$ is the strict order on $\mathbb{R}^2$ where $e >_{PD} e'$ if $e_r \leq e'_r$ and $e_b \leq e'_b$ with at least one inequality strict.

We now modify this order to include fairness considerations.

Definition 3. The *fairness-accuracy (FA) dominance* relation $>_{FA}$ is the strict order on $\mathbb{R}^2$ where $e >_{FA} e'$ 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 inequality strict.[12]

That is, if one error pair is simultaneously more accurate for both groups and also more fair, then all designers prefer the former error pair. We call any preference over error pairs $e = (e_r, e_b) \in \mathbb{R}^2$ that is consistent with this order a fairness-accuracy preference.

Definition 4. A *fairness-accuracy (FA) preference* $\succeq$ is any preference relation on $\mathbb{R}^2$ such that $e \succ e'$ whenever $e >_{FA} e'$.

The class of FA preferences reflects a broad range of views regarding how to trade off fairness and accuracy, encompassing some of the most popular and widely-used specifications outlined below.[13]

Example 4 (Utilitarian). The designer evaluates errors $e = (e_r, e_b)$ according to the weighted sum in the population. That is, let $w^U(e) = -p_r e_r - p_b e_b$, and let $\succeq^U$ be the ordering represented by w^U, so that $e \succeq^U e'$ 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). The class of FA preferences also nests a generalization of the Utilitarian rule to the class of utility functions $w^\gamma(e) = -\gamma_r e_r - \gamma_b e_b$, where the weights $\gamma_r, \gamma_b \geq 0$ may be different from

[12] Kleinberg and Mullainathan (2019) define an admissions rule to be a strict improvement over another if it improves both efficiency (the average type of an admitted applicant) and equity (the fraction of admitted students who belong to the disadvantaged group), which is similar to our FA dominance relation but non-nested, as it involves two loss functions. Online Appendix O.1 presents a generalization of our FA dominance relation to that case.

[13] Our consideration of the wide class of FA preferences is motivated in part by the experimental literature on social preferences, which documents substantial heterogeneity across individuals' equity-efficiency preferences. For example, when given the choice between different allocations of payoffs across individuals, some experimental subjects choose Pareto-dominated allocations that are more equal (corresponding in our setting to choice of $e = (e_r, e_b)$ over $e' = (e'_r, e'_b)$ where $e' >_{PD} e$ but $|e_r - e_b| < |e'_r - e'_b|$). These are minority preferences in the population (Andreoni and Miller, 2002; Charness and Rabin, 2002), but constitute 31% of subjects in an experiment in Fisman et al. (2007).

the population weights, as considered in Charness and Rabin (2002), Saez and Stantcheva (2016), Dworczak et al. (2021), and Rambachan et al. (2021) among others.

Example 5 (Rawlsian). The designer evaluates errors $e = (e_r, e_b)$ according to the greater error. That is, let $w^R(e) = -\max\{e_r, e_b\}$, and consider a preference order represented by w^R. A designer with such preferences is called *Rawlsian* (Rawls, 1971).[14]

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 consider the lexicographic preference order that first evaluates errors according to w^E and then breaks ties using the Utilitarian utility w^U. A designer with such preferences is called *Egalitarian*.

Example 7 (Constrained Optimization). The designer evaluates errors $e = (e_r, e_b)$ lexicographically, first determining if they satisfy the fairness constraint $|e_r - e_b| \leq c$, and if so using the Utilitarian criterion. If instead $|e_r - e_b| > c$ then the designer uses the Egalitarian criterion. Any error pair that satisfies $|e_r - e_b| \leq c$ is preferred to any pair that does not.[15]

Example 8 (Accuracy then Fairness). The designer evaluates errors $e = (e_r, e_b)$ by first evaluating accuracy and then fairness. That is, $e \succ e'$ if $e >_{PD} e'$, and if not, then they are compared using w^E. This is the approach recently proposed by Viviano and Bradic (2023).

We do not take a normative stance on which FA preferences are more appropriate, instead viewing the class of FA preferences as encompassing a broad range of designer preferences that may be relevant in practice.[16]

2.3. **The Fairness-Accuracy Frontier.** Fixing any covariate vector X, we define the feasible set of group error pairs as all pairs that can be achieved by some algorithm using X as input.

Definition 5. The *feasible set* given covariate vector X is[17]

$$\mathcal{E}_X := \{e(a) : a \in \Delta(\mathcal{A})\}.$$

[14]This approach is also known as *group distributionally robust optimization* (Sagawa et al., 2020; Hansen et al., 2024).

[15]This is a common approach in the algorithmic fairness literature (Hardt et al., 2016; Corbett-Davis et al., 2017; Menon and Williamson, 2018; Agarwal et al., 2018; Ferry et al., 2023).

[16]A preference that falls outside of this class is the one represented by $w(e) = -p_r e_r + p_b e_b$. This corresponds to a designer with taste-based discrimination, who wants to maximize errors for group b. See Appendix O.4 for a partial generalization of our results for such preferences.

[17]It is clear that the feasible set also depends on the set $\mathcal{A}$ of allowed algorithms. We omit this dependence for ease of exposition.

We first define the standard Pareto frontier, which identifies points that cannot be strictly improved in terms of both groups' accuracies.

Definition 6. The *Pareto frontier* given X is

$$\mathcal{P}_X := \{e \in \mathcal{E}_X : \text{there is no } e' \in \mathcal{E}_X \text{ such that } e' >_{PD} e\}.$$

Similarly, we define the fairness-accuracy frontier as the set of all group error pairs in $\mathcal{E}_X$ that are undominated under the FA dominance relation.

Definition 7. The *fairness-accuracy (FA) frontier* given X is

$$\mathcal{F}_X := \{e \in \mathcal{E}_X : \text{there is no } e' \in \mathcal{E}_X \text{ such that } e' >_{FA} e\}.$$

The FA frontier can be viewed as the set of group error pairs that are optimal for some FA preference. In fact, each point on the FA frontier is uniquely optimal for at least one FA preference, ensuring that no points can be removed without disadvantaging some designer's goals. See Proposition C.1 in Appendix C for details.

3. The Fairness-Accuracy Frontier

This section presents our characterizations of the fairness-accuracy frontier. Section 3.1 defines the properties of *group-balance* and *group-skew* that will play a key role in many of our results. Section 3.2 presents our main characterization of the FA frontier, which says that whether the covariate vector is group-balanced or group-skewed determines what kinds of fairness-accuracy tradeoffs are relevant. Section 3.3 presents an additional characterization of the FA frontier by describing a class of algorithms that implements the fairness-accuracy frontier.

3.1. **Key Property: Group-Balance vs Group-Skew.**

For all covariate vectors X, the feasible set $\mathcal{E}_X$ is closed and convex as shown by Lemma A.1 in the appendix. Two special feasible points are the following.

Definition 8 (Group Optimal Points). For any covariate vector X, define

$$r_X := \arg\min_{e \in \mathcal{E}_X} e_r \qquad\qquad b_X := \arg\min_{e \in \mathcal{E}_X} e_b$$

to be the feasible points that minimize group r's error and group b's error respectively, where ties are broken by further minimizing the other group's error.

Group optimal points characterize the Pareto frontier (Definition 6). To ease exposition, throughout the paper we use *lower boundary 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.[18]

Observation 1. *The Pareto frontier $\mathcal{P}_X$ is the lower boundary of $\mathcal{E}_X$ between r_X and b_X.*

We similarly define the fairness optimal point.

Definition 9 (Fairness Optimal Point). For any covariate vector X, define

$$f_X := \arg\min_{e \in \mathcal{E}_X} |e_r - e_b|$$

to be the point that minimizes the absolute difference between group errors, where ties are broken by minimizing either group's error.[19]

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 have the lower error. Whether this is the case turns out to be the key property determining the nature of the fairness-accuracy tradeoff given X.

Definition 10. 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*.[20]

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 (even though the feasible set may still intersect the 45 degree line, as in Figure 4). 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.

[18]In the degenerate case when the two points coincide, we will use lower boundary to refer to the common point purely for expositional convenience.

[19]Since $\mathcal{E}_X$ is convex, the point f_X is the same regardless of which group is used to break ties.

[20]Group-balance/group-skew additionally depends on the joint distribution (X, Y, G) and the loss function. Since we are primarily interested in studying how tradeoffs are affected by changes in the covariate vector X (e.g. adding/banning a covariate), we often treat the other random variables (and loss function) as fixed primitives of our model.

Loosely speaking, a covariate vector is group-balanced if it is possible to separate accurate predictions for one group from accurate predictions for another, and it is group-skewed if this is not feasible. We provide several examples below in which the covariate vector is either group-balanced or group-skewed (all supporting arguments are contained in Appendix A.1).

Example 9 (Strong Independence). Suppose $G \perp\!\!\!\perp (X, Y)$. Then X is group-balanced.

Example 10 (Unequal Means). Suppose there are two functions $a_r, a_b \in \mathcal{A}$ and a mean-zero random variable ε independent of X, such that $Y = a_g(X) + \varepsilon$ for members of group g. If the loss function is $\ell(d, y) = (d - y)^2$, then X is group-balanced.

Example 11 (Asymmetric Predictive Power). Suppose there is a function $a_0 \in \mathcal{A}$ and mean-zero random variables $\varepsilon_r, \varepsilon_b$ independent of X, such that $Y = a_0(X) + \varepsilon_g$ for members of group g. If the loss function is $\ell(d, y) = (d - y)^2$ and ε_b has larger variance than ε_r, then X is r-skewed.

Example 12 (Conditional Independence). Suppose $\mathcal{A} = \overline{\mathcal{A}}$ (all algorithms are feasible) and $G \perp\!\!\!\perp Y \mid X$, i.e., once X is observed there is no additional predictive value to knowing a subject's group identity.[21] Equivalently, $Y = a_0(X) + \varepsilon_X$ for both groups, where the noise term is possibly dependent on X. Then either $r_X = b_X = f_X$ or X is group-skewed.

3.2. **Geometric Characterization of the FA Frontier.** We now provide our main characterization of the FA frontier. Theorem 1 shows that whether the covariate vector X is group-balanced or group-skewed determines the shape of the frontier.

Theorem 1.

(a) *If X is group-balanced, then $\mathcal{F}_X = \mathcal{P}_X$.*
(b) *If X is g-skewed, then $\mathcal{F}_X$ is the boundary of $\mathcal{E}_X$ between g_X and f_X containing $\mathcal{P}_X$.*[22]

These two cases are depicted in Figure 2. 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 the standard Pareto frontier, i.e. the lower boundary of the

[21]This kind of conditional independence appears when the coefficient on group identity is zero in a regression of Y on observables. For example, Ludwig and Mullainathan (2024) find that race (G) is not predictive of a criminal's risk (Y) conditional on arrest (X) in their data.

[22]When $r_X = b_X$ so the Pareto frontier is a singleton, there are two boundaries between g_X and f_X containing $\mathcal{P}_X$, and we mean the lower boundary as defined in Section 3.1 (setting e_g to be on the x-axis).

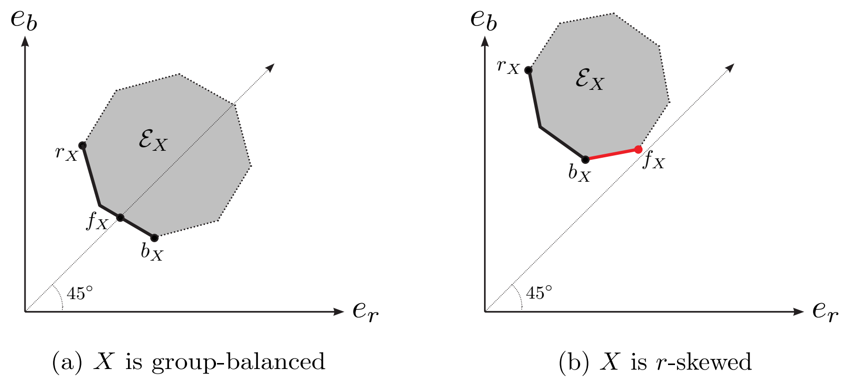


FIGURE 2. Example feasible set and fairness-accuracy frontier for (a) a group-balanced covariate vector X and (b) an r-skewed covariate vector X.

feasible set connecting these two points. Here, the only tradeoffs are between the accuracy of one group and the accuracy of the other group. 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 (depicted in Figure 2 in red). This positively sloped line segment is novel; here, tradeoffs are between fairness and the accuracy of *both* groups, representing a strong conflict between fairness and accuracy. We summarize these observations in the following corollary.

Corollary 1. *Suppose f_X is distinct from r_X and b_X. Then X is group-skewed if and only if there exist points $e, e' \in \mathcal{F}_X$ such that $e >_{PD} e'$ but $|e_r - e_b| > |e'_r - e'_b|$.*

The corollary implies that if the covariate vector is group-skewed (and the fairness optimal point is distinct from the group optimal points), then the FA frontier must consist of a positively-sloped segment along which every pair of points can be Pareto-ranked. In practice, designers may end up at these Pareto-dominated outcomes by choosing to ignore certain available covariates. For example, some medical practitioners recommend removing race-based covariates from healthcare prediction algorithms, even as they acknowledge the predictive value of those covariates for both groups (Vyas et al., 2020; Cerdeña et al., 2020; Delgado et al., 2021). Similarly, some university admissions committees have elected to exclude consideration of test scores from admissions decisions.[23] Corollary 1 says that if

[23]Test scores are predictive of college grades for all of the relevant demographic groups (see Section A.5 of Systemwide Academic Senate (2020)), but are more predictive for applicants in some groups than others (Rothstein, 2004). In Section 4.3.3 we return to this application, interpreting the exclusion of test scores

the covariate vectors in these settings are group-skewed, then disagreements over whether to implement Pareto-dominated errors can be explained by different fairness preferences. On the other hand, if inputs are group-balanced, then a policy proposal that implements Pareto-dominated errors cannot be rationalized by any fairness-accuracy preference, regardless of how strongly the designer weights fairness.

3.3. **Algorithms that Implement the FA Frontier.** In this section we provide a different characterization of the FA frontier, explaining how the design of the optimal algorithm varies along the frontier when the set of algorithms is unconstrained.[24] Below we suppose without loss of generality that the covariate vector is either group-balanced or r-skewed (where the b-skewed case follows identically by flipping the group labels in the subsequent definitions). For each realization x of the covariate vector, define

$$\Delta_g^x := \mathbb{P}(X = x \mid G = g) \cdot \mathbb{E}[\ell(1, Y) - \ell(0, Y) \mid X = x, G = g]$$

This quantity measures how group g's expected error changes if the decision at covariate vector x is switched from 0 to 1. The optimal decision at x for minimizing group g's error is $d = 0$ if $\Delta_g^x > 0$ and $d = 1$ if the opposite inequality holds. Thus an algorithm a_r^* that achieves the r-optimal point r_X can be described as follows:

$$a_r^*(x) = \begin{cases} 0 & \text{if } \Delta_r^x > 0 \text{ or } \Delta_b^x \geq \Delta_r^x = 0 \\ 1 & \text{if } \Delta_r^x < 0 \text{ or } \Delta_b^x < \Delta_r^x = 0 \end{cases}$$

where we tie-break in favor of lower group b error if $\Delta_r^x = 0$.[25]

Define

$$h(x) = \begin{cases} \frac{\Delta_b^x}{\Delta_r^x} & \text{if } \Delta_r^x \neq 0 \\ \infty & \text{if } \Delta_r^x = 0 \end{cases}$$

to be the ratio of the relative impact of the decision at x on group b's error to group r's error (except when there is no impact on group r, in which case $h(x)$ is set to infinity). Then order the realizations $x_1, \ldots, x_n$ of X so that $h(x_1) \leq h(x_2) \leq \cdots \leq h(x_n)$, with ties broken arbitrarily. Recall the r-optimal algorithm a_r^* defined above, and consider the following class

slightly differently—not as a choice made by the agent setting the algorithm, but as an informational regulation imposed by a designer whose preferences are different from those of the agent.

[24]Appendix C offers an additional characterization of the FA frontier as the set of optimal error pairs for a class of "simple" preferences.

[25]The algorithm a_r^* so defined is the unique algorithm that achieves r_X, except when $\Delta_b^x = \Delta_r^x = 0$ for some realization of the covariate vector x. At those covariate vectors x, either decision leads to the same expected error for group r, as well as for group b.

of algorithms $a_0, a_1, \cdots, a_n$ given by

$$a_i(x_j) := \begin{cases} a_r^*(x_j) & \text{if } j > i \\ 1 - a_r^*(x_j) & \text{if } j \leq i \end{cases}$$

In words, $a_0 = a_r^*$ is the r-optimal algorithm that assigns to every covariate vector the decision that is best for group r. Moving along the sequence $a_1, a_2, \ldots$ entails successively switching the assignment at the next covariate vector in the order $x_1, x_2, \ldots$. The following proposition says that the FA frontier is exactly the set of outcomes generated by algorithms of this form.

Proposition 1. *Suppose $\mathcal{A} = \overline{\mathcal{A}}$, and define $n_c = \max\{i : \ h(x_i) < c\}$ for $c \in \{0, 1\}$. Then $\mathcal{F}_X$ is the set of errors generated by the algorithms $(1-\beta)a_{i-1} + \beta a_i$ for all $(\beta, i) \in \Psi$ where*

(a) *if X is group-balanced, then $\Psi = [0,1] \times \{1, \ldots, n_0\}$, and*

(b) *if X is r-skewed, then $\Psi = ([0,1] \times \{1, \ldots, n^* - 1\}) \cup \big([0, \overline{\beta}] \times \{n^*\}\big)$ for some $\overline{\beta} \in [0,1]$ and $n^* \in \{n_0, \ldots, n_1\}$. Moreover, if $e_b > e_r$ at f_X, then $(\overline{\beta}, n^*) = (1, n_1)$.*

The proposition is illustrated in Figure 3. The two groups disagree over the optimal decision at the realizations $x_1, \ldots, x_{n_0}$, since one group prefers $d = 1$ while the other group prefers $d = 0$. (In the figure, $n_0 = 2$.) Here, the tradeoff is between the accuracy of one group and that of the other, and these realizations are ordered so that the relative error impact on group b compared to group r is maximized at x_1, and decreases as we progress along the sequence. When the assignments for all of these disagreement covariate vectors have been swapped from the r-optimal assignment to the b-optimal assignment, we obtain an algorithm that implements the b-optimal point b_X. Part (a) of Proposition 1 says that when X is group-balanced, the algorithms $(a_i)_{0 \leq i \leq n_0}$ implement the extreme points of the FA frontier (which is also the Pareto frontier by Theorem 1), and that mixing between neighboring algorithms of this form yields the entire FA frontier.

When X is group-skewed, the FA frontier continues further by swapping the assignment at covariate vectors $x_{n_0+1}, x_{n_0+2}, \ldots$. The two groups agree on the preferred decision at these covariate vectors, so swapping the assignment hurts both groups. Here, the tradeoff is between the accuracy of both groups and fairness; the realizations are ordered so that the relative impact on group r's error to group b's error is largest at x_{n_0+1}, and decreases as we progress along the sequence. These swaps continue until we reach the fairness-optimal point f_X, which is either the extreme point implemented by the algorithm a_{n^*} as in Figure 3 (in

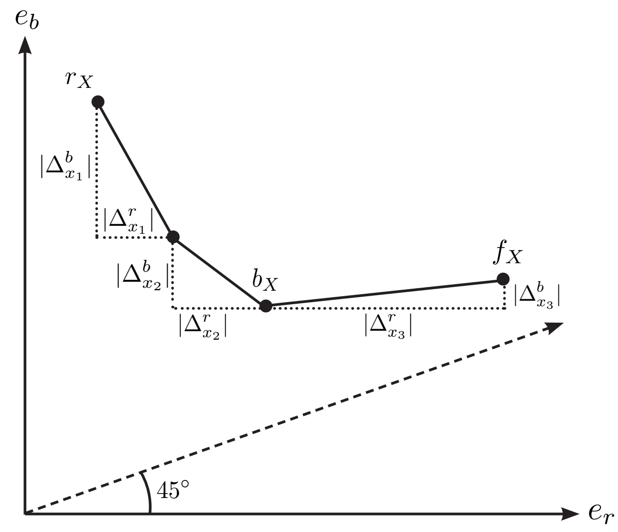


FIGURE 3. In this example, the point r_X is implemented by the algorithm $a_0 = a_r^*$. The next extreme point is implemented by a_1, which satisfies $a_1(x_1) = 1 - a_r^*(x_1)$, and $a_1(x_i) = a_r^*(x_i)$ for all $i > 1$. The point b_X is implemented by the algorithm a_2, which satisfies $a_2(x_i) = 1 - a_r^*(x_i)$ for $i = 1, 2$, and $a_2(x_i) = a_r^*(x_i)$ for all $i > 2$. Finally, f_X is implemented by the algorithm a_3, which satisfies $a_3(x_i) = 1 - a_r^*(x_i)$ for $i = 1, 2, 3$, and $a_3(x_i) = a_r^*(x_i)$ for all $i > 3$. In this example, $n_0 = 2$ (i.e., groups r and b have different optimal decisions for covariate vectors x_1, x_2), while $n^* = n_1 = 3$.

which case $\overline{\beta} = 1$), or an interior point on the line segment that ends at this extreme point (in which case $\overline{\beta} < 1$).

3.4. **Special Cases.** In some applications, the designer may be able to explicitly condition the algorithm on group identity. That is, the set of mappings is all pairs $(a_r, a_b) \in \mathcal{A} = \mathcal{A}_r \times \mathcal{A}_b$, where $\mathcal{A}_g$ is the permissible set of mappings applied for members of group g.[26]

When $\mathcal{A}_r = \mathcal{A}_b = \overline{\mathcal{A}}$, then this setup is equivalent to our main framework with $\mathcal{A} = \overline{\mathcal{A}}$ and group identity G as a covariate. As before, we permit randomizations over pairs of this form so the set of algorithms is $\Delta(\mathcal{A})$. The feasible set and fairness-accuracy frontier simplify as follows.

Proposition 2. *Suppose the algorithm can be conditioned on group identity. Then $\mathcal{E}_X$ is a rectangle whose sides are parallel to the axes, and $\mathcal{F}_X$ is the line from $r_X = b_X$ to f_X.*

[26]To formally nest this within the original framework, consider $\mathcal{A}_r \times \mathcal{A}_b$ where $\mathcal{A}_g$ is the permissible set of mappings $a_g : \mathcal{X} \to \mathcal{D}$ for group g. Expand the covariate space to $\hat{\mathcal{X}} = \mathcal{X} \times \{r, b\}$ and note that any pair (a_r, a_b) can be associated with the mapping $\hat{a} : \hat{\mathcal{X}} \to \mathcal{D}$ where $\hat{a}(x, g) = a_g(x)$.

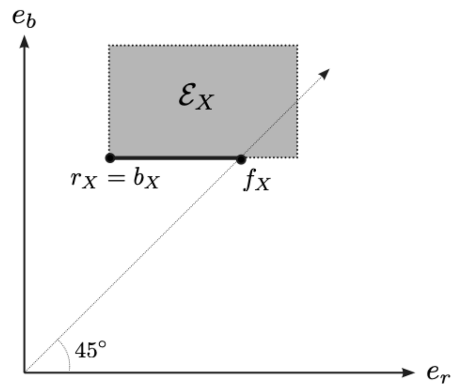


FIGURE 4. When group-specific algorithms are permitted, then the feasible set is a rectangle and the fairness-accuracy frontier is a line segment.

Figure 4 depicts such a feasible set and fairness-accuracy frontier. One endpoint, the Utilitarian-optimal point labeled $r_X = b_X$, gives both groups their minimal feasible error. The other endpoint, the Egalitarian-optimal point f_X, maximizes fairness. Everywhere along the fairness-accuracy frontier $\mathcal{F}_X$, the higher error (worse-off) group receives its minimal feasible error. Thus when group-specific algorithms are permitted, the fairness-accuracy tradeoff simply involves comparing fairness with the welfare of the advantaged group (i.e., the group that the covariate vector is skewed towards).

When the joint distribution $\mathbb{P}$ relating (X, Y, G) satisfies certain independence properties, the FA frontier again simplifies. Specifically, under the assumptions of our previous Examples 9 and 12, we obtain the following result.

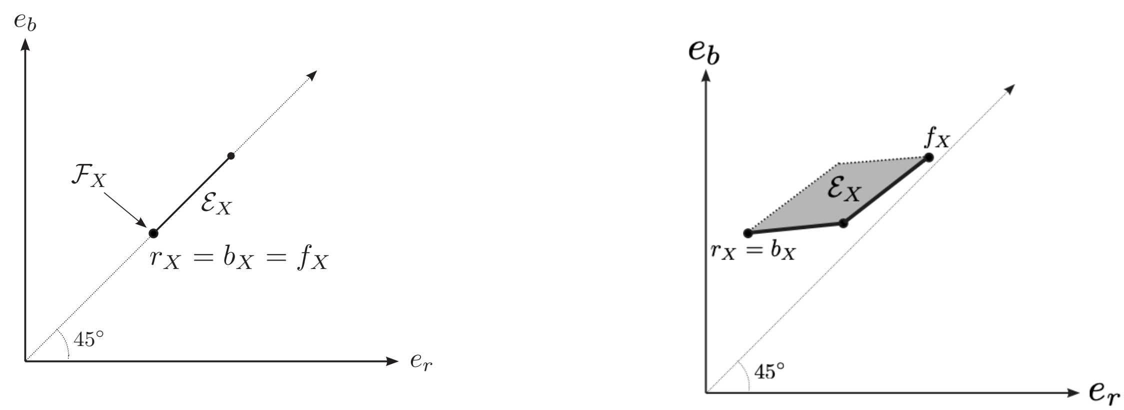
Proposition 3. (a) *Suppose $G \perp\!\!\!\perp (X, Y)$. Then $\mathcal{E}_X$ is a line segment on the 45-degree line, and $\mathcal{F}_X$ is a single point.*

(b) *Suppose $G \perp\!\!\!\perp Y \mid X$ and $\mathcal{A} = \overline{\mathcal{A}}$. Then $r_X = b_X$, and $\mathcal{F}_X$ is that part of the lower boundary of the feasible set $\mathcal{E}_X$ from the point $r_X = b_X$ to the point f_X.*[27]

These cases are depicted in Figure 5. In Panel (a), the FA frontier is a singleton, and fairness-accuracy preferences are irrelevant: All designers who agree on the basic FA-dominance principle outlined in Definition 6 prefer the same algorithm. The statistical

[27]Without loss, let X be either group-balanced or r-skewed. Then by "lower boundary" we mean the definition from Section 3.1, setting e_r to be on the x-axis.

condition in this part of the result, $G \perp\!\!\!\perp (X, Y)$, is a special case of the conditional independence condition, $G \perp\!\!\!\perp Y \mid X$, in the second part of the result. In Panel (b), the leftmost point is the (shared) group optimal point $r_X = b_X$, and the right point is the fairness optimal point f_X. 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 the only relevant difference across designers is how they choose to resolve strong fairness-accuracy conflicts of this form. Although the statistical conditions in Proposition 3 are strong and unlikely to be exactly satisfied, the fairness-accuracy tradeoffs they describe are relevant to real data (see Section 5).



(A) Strong Independence: $G \perp\!\!\!\perp (X, Y)$ (B) Conditional Independence: $G \perp\!\!\!\perp Y \mid X$

Figure 5. Special Cases

Relative to the setting of Proposition 2 (where we assumed it was possible to directly condition on group identity), the conditional independence condition $G \perp\!\!\!\perp Y \mid X$ need not permit a full separation of the two groups' errors.[28] For example, it may not be possible to increase e_r without also increasing e_b. Loosely speaking, the more information X contains about G, the more the frontier "flattens" to the horizontal line depicted in Figure 4.

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

[28]Clearly, the assumption that $G \mid X = x$ is degenerate for every covariate vector x (which is implied if group identity is a covariate) is stronger than $G \perp\!\!\!\perp Y \mid X$.

of some settings, e.g., a company may internalize both fairness and accuracy 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 healthcare provider (agent) determining treatment may seek to maximize the number of correct diagnoses, while a policymaker (designer) may additionally prefer that the accuracy of the provider's treatments be similar across 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 sharply across groups. In these settings, the designer can often influence the algorithm indirectly by implementing policies that constrain the algorithm's inputs. As one illustration of this, Chan and Eyster (2003) report that as part of an effort to influence Berkeley law school's admissions policy in 1997, UC Berkeley administrators coarsened candidates' LSAT scores into intervals and reported this coarsened variable to the law school admissions committee.

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 provide conditions under which the designer is nevertheless able to implement his favorite point on the fairness-accuracy frontier. In Section 4.3, we study whether the designer's optimal regulation of inputs may involve completely banning a covariate such as group identity or a test score.

4.1. **Input Design Model.** A designer chooses a *garbling* of the covariate vector X, which is represented as a mapping $T : \mathcal{X} \to \Delta(\mathcal{T})$ taking realizations of X into distributions over possible realizations of T, which we denote by the set $\mathcal{T}$ (assumed without loss to be finite). Examples of garblings include the following.

Example 13 (Banning an Input). Suppose $X = (X_1, X_2, X_3)$ and the designer wants to ban the last coordinate X_3. The corresponding garbling is $T(x_1, x_2, x_3) = (x_1, x_2)$ with probability 1.

Example 14 (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.

Example 15 (Adding Noise). $T(x) = x + \varepsilon$ where the noise term ε takes value $+1$ or -1 with equal probability.

We view these garblings as information policies that the designer can possibly commit to by law. Real examples of garblings are abundant: The "ban-the-box" campaign (Agan and Starr, 2018) restricted employers from using criminal history as an input into hiring decisions (similar to Example 13); the College Board coarsens a test-taker's answers into an integer-valued score between 400 and 1600 (similar to Example 14); and organizations such as the US Census Bureau, Apple, and Google add noise to users' inputs under differential privacy initiatives (similar to Example 15), see Garfinkel et al. (2018).

The agent chooses an algorithm $a : \mathcal{T} \to \mathcal{D}$ that takes as input the garbled variable chosen by the designer. For simplicity in this section, we assume that the full set of algorithms $\overline{\mathcal{A}}$ is available; as before, we allow for randomizations over algorithms. The agent evaluates errors according to the utility function

$$w(e) = -\alpha_r e_r(a) - \alpha_b e_b(a) \tag{1}$$

for some constants $\alpha_r, \alpha_b > 0$ that are known to the designer.[29] When $\alpha_g = p_g$, the agent is Utilitarian and exclusively cares about aggregate accuracy; otherwise, the agent's preference falls in the broader class of generalized Utilitarian preferences mentioned in Example 4.[30] We can rewrite (1) as

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

where p_t is the probability of $T = t$. Thus the agent's problem of minimizing ex-ante error is equivalent to the following ex-post problem[31]

$$a(t) \in \arg\min_{d \in \mathcal{D}} \sum_{y,g} \frac{\alpha_g}{p_g} \cdot \mathbb{P}\left(Y = y, G = g \mid T = t\right) \cdot \ell\left(d, y\right). \tag{2}$$

[29]In Online Appendix O.4, we prove additional results for the case when a coefficient α_g is negative, so that the agent is adversarial or biased against group g and prefers to *increase* error for that group. This could reflect taste-based discrimination by the agent. Note that this falls outside of our class of FA preferences. The case where the agent has fairness concerns that the policymaker does not share is also interesting. See Section 6 for a brief discussion of some technical complications that arise in this case.

[30]The agent's utility may involve weights different from 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.

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

Definition 11. An error pair $e = (e_r, e_b)$ is *implemented by* T if there exists an algorithm a_T satisfying (2) such that $e = e(a_T)$.

Fixing any covariate vector X, we define the input-design feasible set to be all error pairs that can be implemented by some garbling T, and the input-design fairness-accuracy frontier to be the set of group error pairs that are FA-undominated in the input-design feasible set.

Definition 12. The *input-design feasible set* given covariate vector X is

$$\mathcal{E}_X^* := \{e(a_T) : T \text{ is a garbling of } X\}.$$

Definition 13. The *input-design FA frontier* given X is

$$\mathcal{F}_X^* := \{e \in \mathcal{E}_X^* : \text{there exists no } e' \in \mathcal{E}_X^* \text{ such that } e' >_{FA} e\}.$$

4.2. **Comparing Input Design and Algorithm Design.** The following proposition says that under relatively weak conditions, it is 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., given control of the algorithm) can also be achieved under input design. To state the result, we define

$$e_0 := \min_{d \in \mathcal{D}} \left(\alpha_r \cdot \mathbb{E}[\ell(d, Y) \mid G = r] + \alpha_b \cdot \mathbb{E}[\ell(d, Y) \mid G = b]\right) \tag{3}$$

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

$$H := \{(e_r, e_b) \, : \, \alpha_r e_r + \alpha_b e_b \leq e_0\} \tag{4}$$

to be the halfspace including all error pairs that improve the agent's payoff relative to no information.

Proposition 4 (When Input Design is Without Loss)**.** *The following hold:*

(a) *Suppose X is group-balanced. Then, $\mathcal{F}_X^* = \mathcal{F}_X$ if and only if $r_X, b_X \in H$.*
(b) *Suppose X is g-skewed. Then, $\mathcal{F}_X^* = \mathcal{F}_X$ 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. Related results appear in Alonso and Câmara (2016) and Ichihashi (2019), although we provide an independent argument in Appendix B.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 FA frontier is $\mathcal{F}_X^* = \mathcal{F}_X \cap H$.*

Clearly the designer cannot hold the agent to a payoff lower than what the agent can guarantee with no information, so $\mathcal{E}_X^* \subseteq \mathcal{E}_X \cap H$. In the other direction, we need to show that every point in $\mathcal{E}_X \cap H$ can be implemented by a garbling of X. The proof is by construction: If the designer garbles X into recommendations of the decision,[32] then the obedience constraints reduce precisely to the condition that the agent's payoff is improved relative to no information, i.e., the error pair belongs to H. This yields the lemma, and Figure 6 illustrates how Proposition 4 follows from 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, r_X and f_X in the case of an r-skewed X, or b_X and f_X in the case of a b-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 for some designers.

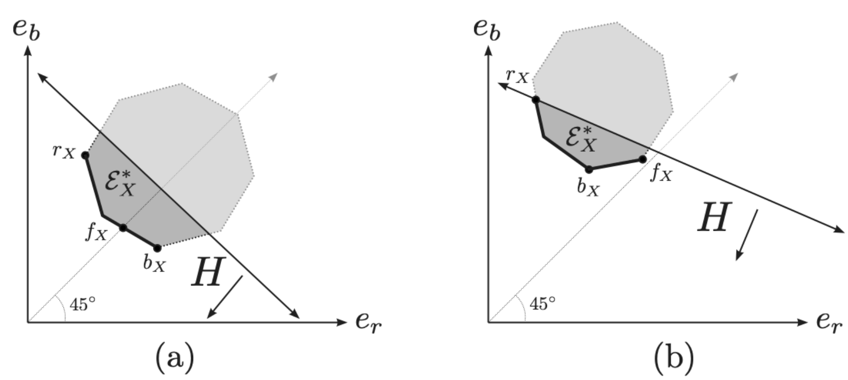


FIGURE 6. Depiction of an example input-design fairness-accuracy frontier for (a) a group-balanced covariate vector X and (b) an r-skewed covariate vector X. In Panel (a), it is sufficient to check $r_X, b_X \in H$ to determine whether the entire unconstrained fairness-accuracy frontier belongs to H. This condition is satisfied in the figure, so every designer can implement his favorite unconstrained outcome using input design. In Panel (b), it is sufficient to check whether $r_X, f_X \in H$. This condition is failed in the figure, so some designer cannot implement his favorite unconstrained outcome using input design.

[32]An interesting direction for future research would be to impose "procedural fairness" requirements on the garbling (such as requiring the garbling to be deterministic, or to be monotone in certain covariates), and ask what is achievable under those constraints. Throughout the present paper, we maintain full flexibility to garble covariates.

4.3. **Excluding a Covariate.** We next turn to the question of whether the optimal garbling may involve completely banning the use of a specific covariate. For example, healthcare providers disagree over whether race should be a permitted input into clinical prediction algorithms (Vyas et al., 2020; Manski, 2022; Manski et al., 2023), and universities differ in whether they choose to exclude consideration of standardized test scores.[33]

Since the designer and agent have (potentially) misaligned preferences, banning an input can be optimal, and we demonstrate this in a simple example in Section 4.3.1. But as we show in Sections 4.3.2 and 4.3.3, there are two important classes of inputs for which bans are strictly worse for all designers. We formalize "strictly worse for all designers" by comparing the fairness-accuracy frontier with and without the input under consideration.

Definition 14. For any pair of non-empty sets $S, S' \subseteq \mathbb{R}^2$, write $S >_{FA} S'$ if every $e' \in S'$ is FA-dominated by some $e \in S$.

That is, $S >_{FA} S'$ if for each point in S', there exists a point in S that is strictly better in both fairness and accuracy.

When $\mathcal{F}^*_{X,X'} >_{FA} \mathcal{F}^*_X$ (where (X, X') denotes the concatenation of two random vectors X and X') then excluding X' is not optimal for any designer with a FA-preference. That is, every designer can achieve a strictly higher payoff by garbling (X, X') rather than by garbling X alone. This comparison does not in general rank the information policy of fully revealing X versus fully revealing (X, X'). That is, it may be that $\mathcal{F}^*_{X,X'} >_{FA} \mathcal{F}^*_X$, but the designer's payoff is higher from revealing X than from revealing (X, X').

4.3.1. *Example.* We start with an example to demonstrate that banning an input can be optimal for the designer. 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; for example, let $X = x_0$ with probability one. Further let X' be a binary signal with the following conditional probabilities $\mathbb{P}(X' \mid Y, G)$:

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

$G = r$

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

$G = b$

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

That is, X' is perfectly informative about the individuals in group r, and imperfectly informative about those in group b. Suppose the loss function is $\ell(d, y) = \mathbb{1}(d \neq y)$, the agent is Utilitarian ($\alpha_r = p_r = 1/2$ and $\alpha_b = p_b = 1/2$), and the designer is Egalitarian.

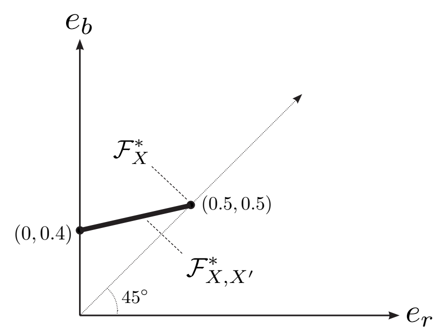


FIGURE 7. The fairness-accuracy frontier given (X, X') is the line segment connecting $(0, 0.4)$ with $(0.5, 0.5)$. Every nontrivial garbling of (X, X') leads to a point on this frontier that differs from $(0.5, 0.5)$, and hence yields a strictly negative payoff for the designer.

The input-design feasible set given X only is the singleton $\{(0.5, 0.5)\}$, and the Egalitarian designer's payoff at this point is zero. But if the designer chooses any garbling of (X, X') that provides information about X', his payoff will be strictly negative. Intuitively, the Utilitarian agent uses what he learns about X' to maximize aggregate accuracy. Since this information is necessarily more informative about group r than about group b, decisions based on this information increase the gap between the two group errors, reducing the designer's payoff. Thus, it is strictly optimal for the designer to exclude all information about X'.

4.3.2. *Excluding Group Identity.* We next consider the special case in which $X' = G$. The property of group balance (suitably strengthened) turns out to imply that banning group identity is never optimal.

Definition 15. 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 5. *Suppose $r_X, b_X \in H$. Then $\mathcal{F}^*_{X,G} >_{FA} \mathcal{F}^*_X$ if and only if X is strictly group-balanced.*[34]

[34]The assumption that $r_X, b_X \in H$ makes the above result easier to state as an if-and-only-if condition. But it follows from our proof of Proposition 5 that even when this assumption fails, strict group-balance is a sufficient condition for $\mathcal{F}^*_{X,G} >_{FA} \mathcal{F}^*_X$.

That is, if (and only if) X is strictly group-balanced, every error pair on the fairness-accuracy frontier given X is FA-dominated by an error pair on the fairness-accuracy frontier given (X, G). This result builds on previous findings that *disparate treatment* (using different rules for individuals in different groups) may be necessary to preclude *disparate impact* (effecting disparate harms across groups).[35] Specifically, Proposition 5 implies that to reduce disparate impact, it may be necessary to impose information policies that are asymmetric across groups. Interestingly, this may not involve fully revealing G, so the algorithm may be formally group-blind (thus not exhibiting disparate treatment). Nevertheless, if we consider the total procedure—taking into account both information design and algorithm design—then two individuals who are otherwise identical but belong to different groups may receive different distributions of outcomes. This distinction raises an interesting question for future work regarding how disparate treatment should be conceptualized in settings where both information design and algorithm design are present.

For restricted classes of preferences, we can further characterize how the variable G is used in the designer's optimal garbling. Consider the following class of utility functions that linearly trade off accuracy and fairness:

$$-\gamma_r e_r - \gamma_b e_b - \gamma_f |e_r - e_b|$$

where the parameters $(\gamma_r, \gamma_b, \gamma_f)$ satisfy $\gamma_r, \gamma_b > 0$ and $\gamma_f \geq 0$. We call such preferences *simple*. Every designer with a preference of this form can achieve their optimal payoff using either of two garblings of (X, G) that we now define.

First, for any (X, G), let *the fully revealing garbling* $T_{X,G}$ be the garbling that directly reveals (X, G), i.e., every (x, g) is mapped to itself with probability 1.

Second, recall that a_g^* denotes the g-optimal algorithm (which assigns to each covariate vector x the decision that minimizes group g's error), and let $\underline{e}_g$ denote the group g error under a_g^*. Further let $\overline{a}_r$ denote the algorithm that instead assigns to each covariate vector x the action $1 - a_r^*(x)$ (i.e., the action which is not optimal for group r), and let $\overline{e}_r$ denote the group r error under $\overline{a}_r$. The *r-shaded garbling*, denoted $T_{X,G}^r$, maps each (x, b) to the

[35]See `https://www.justice.gov/crt/book/file/1364106/download` for definitions of disparate treatment and disparate impact. The tension between them is noted explicitly in works such as Chouldechova (2017) and Rambachan et al. (2021), and is implied by results in Chan and Eyster (2003).

message $a_b^*(x)$ with probability 1, and maps each (x, r) to the message $a_r^*(x)$ with probability

$$\beta = \max\left\{\frac{\overline{e}_r - \underline{e}_b}{\overline{e}_r - \underline{e}_r},\ 0\right\}$$

and to the message $1 - a_r^*(x)$ otherwise. Intuitively, the r-shaded garbling preserves all the information in X for members of group b, but adds noise to this information for members of group r. Fixing any value of $\overline{e}_r$, the amount of weight on the message $a_r^*(x)$ (i.e., the "right" action) is increasing in $\underline{e}_r$ and decreasing in $\underline{e}_b$. That is, the better off group r is at group r's optimal point (and the worse off group b is at its optimal point), the more noise the r-shaded garbling adds for group r.

Proposition 6. *Suppose the designer has a simple preference with parameters $(\gamma_r, \gamma_b, \gamma_f)$. Fix any X for which $f_{X,G} \in H$. Suppose (without loss) that (X, G) is either group-balanced or r-skewed. Then:*

(a) *The fully revealing garbling $T_{X,G}$ is optimal if $\gamma_r \geq \gamma_f$.*
(b) *The r-shaded garbling $T^r_{X,G}$ is optimal if $\gamma_r \leq \gamma_f$.*

That is, if the designer's weight on the inequity term $|e_r - e_b|$ is sufficiently strong, he will prefer the r-shaded garbling. Otherwise, he achieves his optimal point by revealing all of the information in (X, G) un-garbled. In the knife-edge case $\gamma_r = \gamma_f$, the designer is indifferent between these garblings (both are optimal).

This result complements papers that show that disclosing (X, G) can be superior to disclosing X alone (Chan and Eyster, 2003) and to disclosing any strict subset of covariates (Rambachan et al., 2021).[36] Part (a) of Proposition 6 says that under certain conditions, not only is disclosing (X, G) superior to those information policies, but it is optimal over the class of all possible garblings of (X, G). Our parameterized model moreover allows us to provide a specific cutoff of the designer's weight on fairness that determines which group-dependent information disclosure achieves the optimal outcome.

4.3.3. *Excluding a Covariate When Group Identity is Known.* Next compare the frontier implemented by garblings of (X, G) with the frontier implemented by garblings of (X, G, X'), where X and X' are arbitrary covariate vectors. This allows us to investigate whether

[36]Our frameworks are non-nested; for example, the agent in Chan and Eyster (2003) and Rambachan et al. (2021) choose algorithms subject to capacity constraints, and the agent in Chan and Eyster (2003) has a preference different from the utility form given in (1).

additional covariates can strictly improve outcomes, when group identity is available as a covariate.

Definition 16. Say that X' is *decision-relevant over* X if for each group g, there are realizations (x, x') and $(x, \tilde{x}')$ of (X, X') such that

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

while

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

where each of (x, x', g) and $(x, \tilde{x}', g)$ has strictly positive probability.

That is, X' influences the optimal decision for at least one individual in each group, conditional on X. For example, if X' is a test score and X is high school GPA, then X' is decision-relevant if taking the test score into consideration changes the admission decision for at least one individual in each group g relative to the decision based on GPA alone.

Proposition 7. *Suppose $r_X, b_X \in H$. For any covariate vector X and any covariate vector X' that is decision-relevant over X, $\mathcal{F}^*_{X,X',G} >_{FA} \mathcal{F}^*_{X,G}$.*

This result says that every error pair in $\mathcal{F}^*_{X,G}$ can be improved on in both fairness and accuracy by some error pair in $\mathcal{F}^*_{X,X',G}$. Thus so long as the designer has access to group identity, banning a minimally informative covariate cannot be justified by any fairness-accuracy preference in our class. If the designer moreover has a simple preference (as defined in Section 4.3.2), then Proposition 6 implies the following corollary about the designer's optimal garbling.

Corollary 2. *Suppose the designer has a simple preference with parameters $(\gamma_r, \gamma_b, \gamma_f)$. Fix any (X, X'), and suppose without loss that (X, X', G) is either group-balanced or r-skewed. Then:*

(a) *The fully revealing garbling $T_{X,X',G}$ is optimal when $\gamma_r \geq \gamma_f$.*
(b) *The r-shaded garbling $T^r_{X,X',G}$ is optimal when $\gamma_r \leq \gamma_f$.*

We can apply the two results above to the question of whether to ban test scores in college admissions decisions. College entrance exams are decision-relevant for admissions, even given the rest of the application (Systemwide Academic Senate, 2020).[37] Thus Proposition 7 implies that so long as group identities are permissible inputs for college admission decisions, then excluding test scores reduces welfare for any designer with the ability to garble available covariates. On the other hand, if group identity is not permitted as an input into college admissions decisions, then it may be optimal to completely exclude test scores. With regards to the recent Supreme Court case *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College*, our result suggests that banning affirmative action nationwide may give universities with certain FA preferences reason to ban the use of test scores in admissions decisions (for reasons similar to our example in Section 4.3.1).[38] Corollary 2 further says that among designers with simple preferences, those who sufficiently value accuracy will prefer to reveal test scores for all students, while those who sufficiently value fairness will prefer to use the full informational content of test scores for students in the disadvantaged group, but to add noise to this information for students in the advantaged group.[39]

While our framework abstracts away from many important features of the college admissions process—including capacity constraints (see Subsection 6.4), access to testing (Garg et al., 2021) and test-optional admissions (Dessein et al., 2023))—the link between the availability of group identity and the value of additional information, such as test scores, is one that we believe holds more generally. The crucial point is that when group identity is available, the designer can tailor how the additional information is used for each group separately. In this sense, access to group identity has a positive spillover effect for the value of other covariates, guaranteeing that there is some (possibly group-dependent) garbling of the other information that aligns the agent's and designer's incentives.

We conclude by returning to our example in Section 4.3.1, and illustrating how our previous observations change when group identity is available as a covariate.

[37]Specifically, Section A of Systemwide Academic Senate (2020) finds that test scores are predictive of college success, predictive above other covariates (such as GPA), and predictive for all demographic groups that they consider (with individuals disaggregated by factors such as parental education, family income, and racial/ethnic identity).

[38]Dessein et al. (2023) demonstrate a similar finding in a model in which universities experience costs when making decisions that differ from the preferences of a broader society.

[39]Systemwide Academic Senate (2020) report that one use of test scores at UC Berkeley (prior to the university's move to test-blind admissions in 2021) was to identify applicants with academic promise from relatively disadvantaged backgrounds, who would not otherwise have been accepted (see e.g., the discussion in Section III, 2) Part B-3). Use of an r-shaded garbling leads to a similar policy, which prioritizes use of the available information (in this case, test scores) to make better decisions for the disadvantaged group.

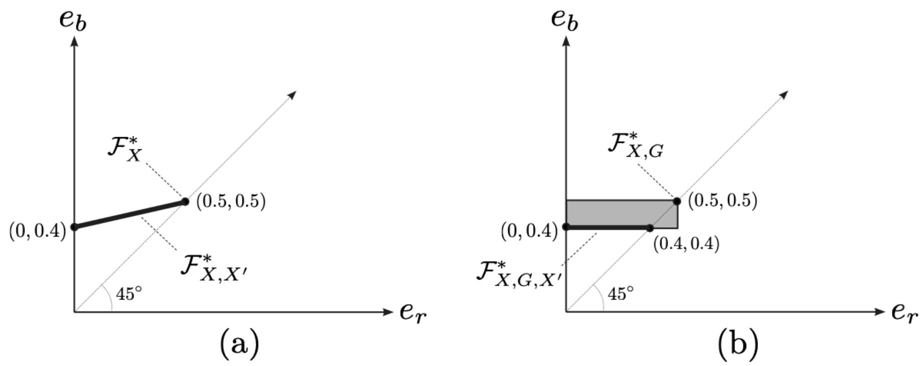


FIGURE 8. Panel (a) reproduces the original figure comparing $\mathcal{F}^*_X$ and $\mathcal{F}^*_{X,X'}$, and Panel (b) compares $\mathcal{F}^*_{X,G}$ and $\mathcal{F}^*_{X,G,X'}$.

Panel (a) reproduces the previous Figure 7 (which compares the frontiers $\mathcal{F}^*_X$ and $\mathcal{F}^*_{X,X'}$), while Panel (b) compares the fairness-accuracy frontiers $\mathcal{F}^*_{X,G}$ and $\mathcal{F}^*_{X,G,X'}$. Without access to group identity, we argued that the Egalitarian designer preferred to ban the covariate X'. With access to group identity, we see that the Egalitarian designer is able to achieve the superior outcome $(0.4, 0.4)$ by garbling (X, G, X'). Thus, while making information about X' available to the agent is strictly harmful for the designer when group identity is not available, this ceases to be true once the designer can garble X' in different ways depending on G.

5. Empirical Application

We have so far emphasized qualitative findings that hold across settings. However, our framework can also be used to better understand quantitative fairness-accuracy tradeoffs in specific datasets. In this final section, we empirically illustrate some of our key definitions using two popular healthcare datasets from the algorithmic fairness literature. This analysis demonstrates that our frontier is computable in practice, and demonstrates the practical relevance of our framework for empirical evaluations of algorithms.

Section 5.1 describes the two datasets. Section 5.2 evaluates group-balance and group-skew, finding that the first satisfies group-skew, while the second corresponds to an interesting case of group-balance in which the group-optimal points approximately lie on the 45 degree line. Section 5.3 depicts the fairness-accuracy frontier for both datasets, setting $\mathcal{A}$ to be the set of linear algorithms. We find that in one dataset, the fairness-accuracy frontier consists primarily of strong fairness-accuracy conflicts (where the designer can only increase fairness by decreasing accuracy for both groups), while in the other there is nearly no conflict between fairness and accuracy.

Having estimated these frontiers, we apply Proposition 4 to study the effectiveness of input design for each of the datasets. We find that for one of the datasets, input design is sufficient to recover the entire frontier, while in the other there are designers who cannot implement their favorite point using input design. Finally, we apply Proposition 2 to study how the fairness-accuracy frontier changes when the designer is permitted group-specific algorithms. We find that for both datasets, neither Utilitarian nor Egalitarian designers have substantially improved payoffs, but designers with moderate fairness and accuracy preferences do benefit from group-specific algorithms.

5.1. **Data.** Our first healthcare dataset consists of the 48,784 patient observations reported in Obermeyer et al. (2019). The covariate vector X includes 8 demographic variables, 34 indicators of specific chronic illnesses, 13 healthcare cost variables, and 94 biomarker and medication variables. We take as the two group identities whether the patient self-reported as Black or White, denoted $g \in \{b, w\}$.

As described in Obermeyer et al. (2019), a large academic hospital used the covariates in X to identify high-risk patients to enroll in an intensive health care program, automatically enrolling those top 3% "highest risk" patients into this program. In the data, "true health needs" are reported as each patient's total number of active chronic illnesses in the subsequent year, where the 97% percentile is 6 health conditions. We thus define the patient's type Y to be an indicator for whether their true health needs are strictly larger than the 97% percentile, and consider algorithms $a : \mathcal{X} \to \{0, 1\}$ that predict Y. We use $\ell(d, y) = \mathbb{1}(d \neq y)$ as our loss function, implying that algorithms are more accurate if they have a lower misclassification rate for each group, and more fair if they have a smaller disparity between the misclassification rates for the two groups.

Our second dataset is from Strack et al. (2014) and contains 101,766 clinical care observations for patients with diabetes diagnoses. The covariate vector X includes 25 variables, including demographic data (e.g., age) and medical information (e.g., diabetic medications and number of inpatient days). We take gender (woman or man) to be the group identities, denoted $g \in \{w, m\}$. The patient's type Y is whether the patient was readmitted to the hospital after release. This variable is reported in the data. We consider algorithms

$a : \mathcal{X} \to \{0, 1\}$ that predict whether patients will be readmitted upon release, and again use $\ell(d, y) = \mathbb{1}(d \neq y)$ as our loss function.

For both datasets, we suppose that the designer chooses a single algorithm for both groups and does not have access to group identity. (This is consistent with the analysis of Obermeyer et al. (2019).) We subsequently consider how the fairness-accuracy frontier changes when group-specific algorithms are permitted.

5.2. **Estimating Group-Balance versus Group-Skew.** For each group g, let (e_r^g, e_b^g) denote the group-g optimal point. We first provide point estimates of these group-optimal points, given which we form conjectures regarding whether the data are group-balanced or group-skewed. Subsequently we formulate suitable statistical tests of those conjectures.

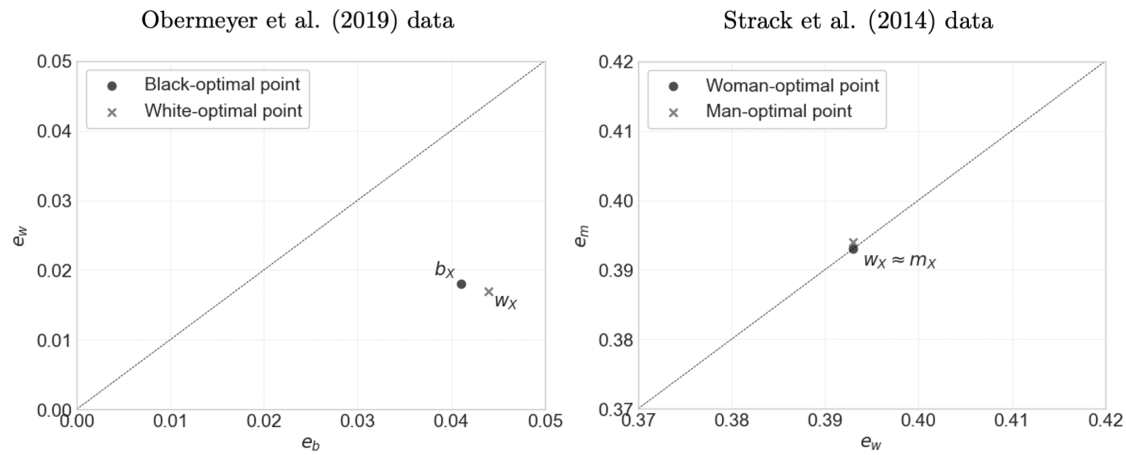
Figure 9 reports five-fold cross-validated estimates of the group-optimal points in each dataset.[40] In the Obermeyer et al. (2019) data, the classification error for Black patients is higher than for White patients at both groups' optimal points, suggesting that the covariate vector X is w-skewed. In contrast, for the Strack et al. (2014) data, the two group-optimal points are nearly on the 45 degree line—that is, the consequences for the two groups are nearly the same regardless of whether the algorithm designer minimizes error for female patients or male patients. This suggests that that the covariate vector satisfies a special case of group-balance.[41]

We now evaluate these conjectures with suitable statistical tests. To assess whether the Obermeyer et al. (2019) dataset is w-skewed, we test the null hypothesis

$$H_0 : e_w^b \geq e_b^b \tag{5}$$

[40]We randomly split the data into five equally sized subsets. In each step of the procedure, we designate one of these subsets to be the test set. We then split the remaining data (the training set) into group-r and group-b observations, and train a random forest algorithm on each part separately, thus obtaining a classifier for each group. We apply group g's classifier to the test set and evaluate each group's error under this classifier, obtaining an estimate of the group-g optimal point $g_X = (e_r^G, e_b^G)$. Finally, we average these estimates across the five choices of which subset to use as the training data. We also tested other algorithms, including neural nets and support vector machines. We find that random forests perform well across both datasets, and all algorithms provide similar results regarding group-balance (see Online Appendix O.6 for details.)

[41]The substantially larger errors for the latter dataset are due to a more difficult prediction problem: In the Strack et al. (2014) data, the outcome variable is roughly equally likely to be 1 or 0 (respectively, 0.46 versus 0.54), whereas in Obermeyer et al. (2019) data, the outcome variable takes the value 0 for almost all patients. Thus, for example, the naive rule that classifies all patients as 0 in the first data achieves a misclassification rate of 0.06 for Black patients and 0.02 for White patients, but these misclassification rates are not feasible for the second dataset.



(a) (b)

FIGURE 9. The group-optimal points for the Obermeyer et al. (2019) data are estimated to be $b_X = (\hat{e}_b^b, \hat{e}_w^b) = (0.041, 0.018)$ and $w_X = (\hat{e}_b^w, \hat{e}_w^w) = (0.044, 0.017)$, suggesting group-skew. The group-optimal points for the Strack et al. (2014) data are estimated to be $w_X = (\hat{e}_w^w, \hat{e}_m^w) = (0.393, 0.393)$ and $m_X = (\hat{e}_w^m, \hat{e}_m^m) = (0.393, 0.394)$, suggesting group-balance.

against the alternative $H_1 : e_w^b < e_b^b$. (Since $e_w^w \leq e_w^b$ and $e_b^b \leq e_b^w$, the inequality $e_w^b < e_b^b$ implies $e_w^w \leq e_b^w$; thus, we only need to verify this part of the definition of group-skew.) We will conclude that the data is w-skewed if we can reject this null hypothesis at the desired significance level.

For the Strack et al. (2014) data, because Figure 9 suggests that the knife-edge condition $e_w^w = e_m^w = e_w^m = e_m^m$ is approximately met, we need to be more careful with our formulation of the null and alternative.[42] We test a relaxed notion of group-balance corresponding to whether e_w^w, e_m^w, e_w^m, and e_m^m are located within a small neighborhood of one another. Specifically, we test

$$H_0\colon |e_w^w - e_m^w| \geq \delta \text{ OR } |e_w^m - e_m^m| \geq \delta \tag{6}$$

against the alternative $H_1\colon |e_f^w - e_m^w| < \delta$ AND $|e_f^m - e_m^m| < \delta$ for $\delta = 0.01$. Under the alternative, both group-optimal points have the property that each group's misclassification rate is within 1 percentage point of the other. (For a sense of whether 0.01 is "small," recall that our estimate for the misclassification rates at the group-optimal points are about

[42]For example, a natural test for strict group-balance (Definition 15) would be to formulate the null H_0 : $e_w^b \leq e_b^b$ OR $e_w^w \geq e_b^w$ against the alternative $H_1 : e_w^b > e_b^b$ AND $e_w^w < e_b^b$, but Figure 9 does not suggest the data is strictly group-balanced. Thus we do not expect to reject this null.

0.40.) When the two group-optimal points are nearly on the 45-degree line, they must nearly coincide. We will thus conclude (approximate) group-balance if (6) is rejected at the desired significance level.

Following Auerbach et al. (2024), we implement the following procedure to test these null hypotheses: In each of $K = 5$ iterations, we randomly split the data into a training set (consisting of two-thirds of the observations) and a test set (consisting of the remaining observations). On the training subset, we search for the algorithm that minimizes group g's error on the group g observations. Then we evaluate each group's error under this algorithm on the test data, and compute a p-value for the suitable null hypothesis via bootstrap (see Appendix D.1 for details). A valid test is obtained by rejecting the null hypothesis whenever the median p-value across these iterations falls below half of the desired significance level (Auerbach et al., 2024). For a 5% significance level, this corresponds to rejecting whenever the median p-value falls below 0.025.

We report the outcome of this test for each dataset under two possible specifications of the set of algorithms $\mathcal{A}$: the set of unconstrained algorithms $\overline{\mathcal{A}}$ and the set of all linear algorithms. For the former specification, we employ a random forest algorithm to search for each group's optimal algorithm among the class of unconstrained algorithms.[43] For the latter, we use logistic regression to look for the linear classifier that minimizes each group's error on the training data. Table 1 reports the output of this analysis. We reject the null in (5) at a 5% significance level for both specifications of $\mathcal{A}$, suggesting that the covariates in the Obermeyer et al. (2019) data are group-skewed. (Recall that we reject at a 5% significance level if the median p-value is less than 0.025.) We reject the null in (6) at a 10% significance level for the unconstrained set of algorithms and at a 5% significance level for the set of linear algorithms, suggesting that the covariates in the Strack et al. (2014) dataset are (approximately) group-balanced with group-optimal points on the 45 degree line.

Thus, by Corollary 1, the first dataset involves a potentially strong conflict between fairness and accuracy, where designers who put sufficient weight on reducing disparities across Black and White patients may prefer to increase errors for both groups. In contrast, fairness considerations across male and female patients do not rationalize implementing Pareto-dominated error rates for the second dataset, regardless of how much weight the designer puts on fairness.

[43] As a robustness check, we verify in Online Appendix O.6 that the performance of three other highly flexible machine learning algorithms yields similar errors to that of the random forest algorithm.

	Dataset 1		**Dataset 2**	
	Group-Balance/Skew	median p-value	Group-Balance/Skew	median p-value
Unconstrained	w-skewed	< 0.0001	group-balanced	0.0437
Linear	w-skewed	< 0.0001	group-balanced	< 0.0001

TABLE 1. We report whether each dataset is group-balanced or group-skewed for the two specifications of $\mathcal{A}$. Reported p-values are computed via bootstrap with size 10,000.

5.3. **The Fairness-Accuracy Frontier.** We next depict the feasible set $\mathcal{E}_X$ and FA frontier $\mathcal{F}_X$ for each of these datasets when $\mathcal{A}$ is the set of linear algorithms. First observe that the extreme points of the feasible set can be found by solving the optimization problem $\min\{\alpha_r e_r(a) + \alpha_b e_b(a) \colon a \in \mathcal{A}\}$ for different choices of $(\alpha_r, \alpha_b) \in \mathbb{R}^2$. We consider the empirical analogue of this optimization problem, in which $e_g(a)$ is replaced by its sample analogue:

$$\min\left\{\alpha_r \hat{e}_r(a) + \alpha_b \hat{e}_b(a) \colon \forall g,\ \hat{e}_g(a) = \frac{1}{n_g}\sum_{i\colon G_i = g} \ell\left(a(X_i), Y_i\right),\ a \in \mathcal{A}\right\}, \tag{7}$$

with n_g denoting the number of group-g observations in the dataset.

Any linear classifier $a \in \mathcal{A}$ can be represented as $a(x) = \mathbb{1}\left\{x^\top \beta \geq 0\right\}$ for some $\beta \in \mathbb{R}^{\dim(\mathcal{X})}$.[44] Thus, we can recast the optimization problem in (7) as the following mixed-integer linear program (MILP):

$$\text{(MILP)}\quad \left[\begin{array}{ll} \min\limits_{\beta, \hat{e}_r \geq 0, \hat{e}_b \geq 0, \hat{Y}} & \alpha_r \hat{e}_r + \alpha_b \hat{e}_b \\ \text{s.t.} & \hat{e}_r = \dfrac{1}{n_r}\sum\limits_{i\colon G_i = r} \mathbb{1}\left\{\hat{Y}_i \neq Y_i\right\} \\ & \hat{e}_b = \dfrac{1}{n_b}\sum\limits_{i\colon G_i = b} \mathbb{1}\left\{\hat{Y}_i \neq Y_i\right\} \\ & \dfrac{X_i^\top \beta}{C_i} < \hat{Y}_i \leq 1 + \dfrac{X_i^\top \beta}{C_i} \text{ and } \hat{Y}_i \in \{0,1\} \quad \text{for all } i \end{array}\right.$$

where C_i is chosen to satisfy $C_i > \sup_{\beta \in \mathcal{B}} |X_i^\top \beta|$, with $\mathcal{B}$ some compact set such that we restrict $\beta \in \mathcal{B}$.[45] In principle, we could use this MILP to estimate all the extreme points of the feasible set by varying the normal vector (α_r, α_b) within $[0, 2\pi) \times [0, 2\pi)$. Since this program is

[44] We assume that X includes a constant term, so it is without loss to set the threshold to zero.

[45] The third constraint is equivalent to $\hat{Y}_i = \mathbb{1}\left\{X_i^\top \beta \geq 0\right\}$ for all i since if $X_i^\top \beta$ is weakly positive then the constraint $\hat{Y}_i \in \{0,1\}$ implies $\hat{Y}_i = 1$, while if $X_i^\top \beta$ is strictly negative then the constraint $\hat{Y}_i \in \{0,1\}$ implies $\hat{Y}_i = 0$.

computationally burdensome, we instead employ a standard relaxation of the MILP to a linear program (LP) by eliminating the integer constraints using a hinge surrogate loss function (see Appendix D.2 for details).[46] We solve this LP for $(\alpha_r, \alpha_b) \in [0, \pi/72, 2\pi/72, \ldots, 2\pi)^2$, estimate $(\hat{e}_r, \hat{e}_b)$ via five-fold cross-validation for each (α_r, α_b), and finally take the convex hull of the estimated error pairs.[47]

Panels (A) and (B) of Figure 10 depict the estimated feasible set and FA frontier for the Obermeyer et al. (2019) data. Panel (A) shows the entire feasible set, while Panel (B) zooms in on the FA frontier. Consistent with our previous hypothesis test, the group-optimal points b_X and w_X lie on the same side of the 45 degree line. Additionally, our estimates of b_X and w_X coincide, so we cannot reject that b_X and w_X are different points at any reasonable significance level.[48] Thus the main tradeoff between fairness and accuracy is whether the designer is willing to increase errors for both groups in return for a decrease in the disparity between group errors. The depicted FA frontier qualitatively resembles Panel (B) of Figure 5 (Conditional Independence), and is consistent with a setting in which the optimal algorithm is the same for both groups, but the measured covariates are more predictive of the outcome for one group (White patients) than the other (Black patients).[49]

Panels (C) and (D) depict the analogous figures for the Strack et al. (2014) data. Consistent with the result of our earlier hypothesis test, the estimated group-optimal points are very close to the 45-degree line (and consequently are also close to the estimated fairness-optimal point f_X). For this dataset, there is almost no tradeoff between fairness and accuracy. In fact, the depicted feasible set and FA frontier qualitatively resemble Panel (A) of Figure 5 (Strong Independence), and are consistent with a setting in which the joint distributions of covariates and outcomes are nearly identical for the two groups.

---

[46]The same reduction underlies the construction of support vector machines for linear classification. See Hastie et al. (2009) for a textbook reference.

[47]For each pair (α_r, α_b), execute the following steps: Partition the dataset into five subsets of equal size. In each of the five iterations, use one subset for testing and the remaining four for training. Determine $\hat{\beta}^{(k)}$ by solving the LP with the training data, which establishes a linear classifier $a^{(k)}$. Compute the estimates of group errors $(\hat{e}_r^{(k)}, \hat{e}_b^{(k)})$ for classifier $a^{(k)}$ by sample analogues using test data. Finally, calculate the average group error $\hat{e}_g$ for each group g across all five folds: $\hat{e}_g = \frac{1}{5}\sum_{k=1}^{5} \hat{e}_g^{(k)}$.

[48]Specifically, we test the null hypothesis

$$H_0 : |e_b^b - e_b^w| \geq \delta \quad \text{OR} \quad |e_w^b - e_w^w| \geq \delta$$

against the alternative $H_1 : |e_b^b - e_b^w| < \delta$ and $|e_r^b - e_r^w| < \delta$, with $\delta = 0.01$. The resulting median p-value is less than 0.0001, indicating insufficient evidence to assert a significant difference in the group-optimal points.

[49]We cannot directly test the assumptions of Example 12 in our data, due to an insufficient number of observations per covariate vector.

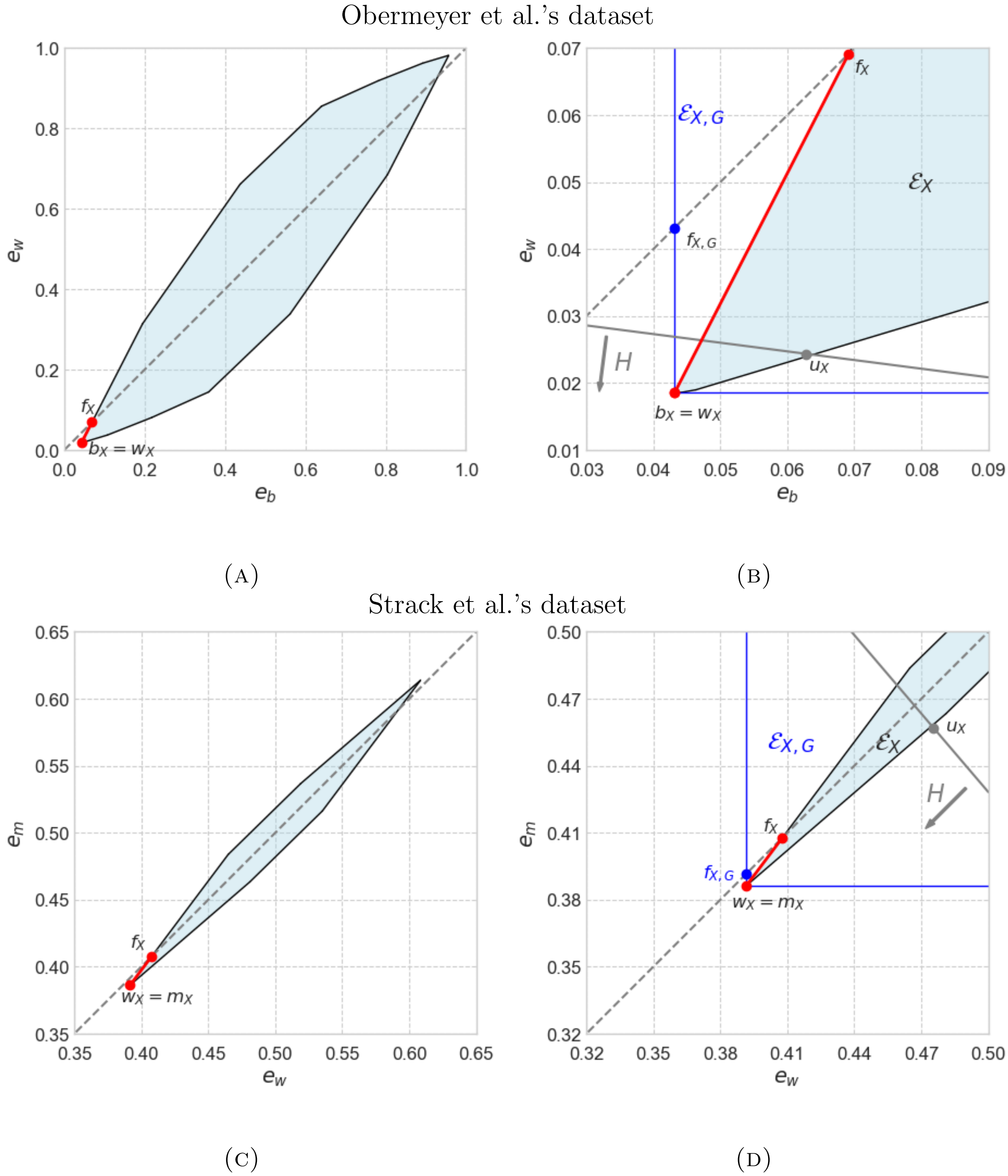


FIGURE 10. Estimated feasible set for the class of linear algorithms

Panels (A) and (C) display the estimated feasible sets for the datasets from Obermeyer et al. and Strack et al., respectively. In these panels, the black solid lines define the boundaries of the feasible sets $\mathcal{E}_X$, and the gray dotted lines serve as 45-degree reference lines. Panel (A) identifies b_X and w_X as the optimal points for Black and White patients, while Panel (C) marks w_X and m_X as the optimal points for women and men. Both panels highlight f_X as the fairness-optimal point. Panels (B) and (D) examine the augmented feasible sets $\mathcal{E}_{X,G}$ for the same datasets with covariate (X, G), represented by blue solid lines. In these panels, $f_{X,G}$ indicates the fairness-optimal point for each dataset with (X, G). The point u_X indicates the error pair chosen by the Utilitarian agent given no information, and the gray solid line in Panel (B) and (D) represents the Utilitarian's indifference curve through u_X.

The two datasets also differ in their implications for the limits of input design. In Panels (B) and (D), we plot the error pair corresponding to the best payoff that a utilitarian agent can achieve with no information (labeled u_X), as well as the agent's indifference curve at this point.[50] In Panel (B), the fairness-optimal point f_X falls outside the halfspace H of error pairs that the utilitarian agent prefers over u_X. Thus, by Proposition 4, input design is with loss for some designers. Specifically, sufficiently fairness-motivated designers cannot induce a utilitarian agent to implement their favorite error outcome, regardless of which garbling of the available covariates they choose. In contrast, in Panel (D) the points w_X, m_X, and f_X all belong to the corresponding halfspace H, implying that every designer with a FA preference can achieve his most preferred outcome.

Finally, we apply Proposition 2 to evaluate the possible fairness-accuracy improvements when different algorithms are used to make predictions for each group. Panels (B) and (D) of Figure 10 illustrate the change in the fairness-accuracy frontier when group-specific linear algorithms are permitted. In both datasets, neither the Utilitarian designer, nor the designer whose payoff is $-|e_r - e_b|$, benefit from using group-specific algorithms: The Utilitarian designer's payoff remains approximately the same, since the group-optimal points change very little when group-specific algorithms are permitted. The designer whose payoff is $-|e_r - e_b|$ does not benefit either, as both f_X and $f_{X,G}$ lie on the 45 degree line (and thus yield an identical payoff of zero for this designer).

This finding is especially interesting when juxtaposed with a recent debate over whether to use race-blind healthcare algorithms. Advocates for race-aware algorithms often adopt a Utilitarian perspective (e.g., Manski et al. (2023)), while proponents of race-blind healthcare algorithms typically argue from the perspective of minimizing healthcare inequalities across groups (e.g., Vyas et al. (2020)). But for these two datasets, it is intermediate designers who value both fairness and accuracy, rather than these two extremes, that benefit the most from use of group-specific algorithms.[51] Whether this particular finding extends to other datasets will depend on specific details of their FA frontiers, but our framework provides a general methodology that can be applied case-by-case to study the particular fairness-accuracy implications of different datasets and algorithmic constraints.

[50]The payoff of the Utilitarian on the indifference curve is e_0 defined in equation (3), setting $(\alpha_r, \alpha_b) := (p_r, p_b)$. Point $u_X = (e_r^U, e_b^U)$ satisfies $p_r e_r^U + p_b e_b^U = e_0$.

[51]If we consider the class of simple preferences for these two datasets, then for any fixed values of α_r and α_b, the increase in the designer's payoff from use of group-specific algorithms is concave in α_f with an interior maximum.

6. Extensions

6.1. **Different loss functions for evaluating fairness and accuracy.** When defining the strict order $>_{FA}$, we used the same loss function to evaluate both accuracy and fairness. This is suitable for healthcare decisions where both the healthcare provider (designer) and patients agree that more accurate decisions are better, and so fairness can be reasonably evaluated as the disparity in accuracy across groups. In other cases where the subjects' utility function is different from the designer's, policymakers sometimes evaluate accuracy using one loss function and fairness using another. For example, a policymaker might evaluate the accuracy of a hiring algorithm based on whether suitable candidates are hired, and the fairness of the algorithm based on the difference in hiring rates across groups. In Appendix O.1 we develop a more general version of our framework that allows for different loss functions, and extends Theorem 1 and Corollary 1 under a generalization of group-balance.

6.2. **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. Theorem 1 holds for any such ϕ with the fairness optimal point f_X suitably defined.[52] We further demonstrate 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 remain unchanged (see Appendix O.3).

6.3. **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 lower errors for both groups. In Appendix O.4, 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 that group's error to be higher. We generalize several results from Section 4 and show that 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).

[52]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 $(e_r - \epsilon_1, e_b - \epsilon_2) >_{FA} (e_r, e_b)$ yielding a contradiction. The rest of the proof follows as in Theorem 1.

Two other potential generalizations would permit the agent and designer to have different loss functions, or permit the agent to also care about fairness. 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, making it 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.

6.4. **Capacity constraints.** In our main model, we allow the designer unconstrained choice of any algorithm. In some applications of interest, there may be an additional capacity constraints on the algorithm, e.g., if only a fixed number of students can be admitted in admissions decisions. One way to formulate a capacity constraint is as 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 2 may no longer be a horizontal line.

6.5. **More than two groups or two decisions.** We have assumed that there are two groups $\mathcal{G} = \{r, b\}$. Some of our results, such as Proposition 4, 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.

We have also 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 a binary decision. We leave a characterization of the input design frontier for this more general case to future work.

6.6. **Group-dependent loss functions.** The main text assumes the same loss function for each group. Although we consider group-dependent loss functions of the form $\ell : \mathcal{D} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}_+$ more difficult to interpret,[53] certain group-dependent loss functions allow us to accommodate additional fairness metrics from the literature such as equalized odds (see Online Appendix O.5.3 for details). All of our results extend directly to this broader class of loss functions (replacing $\ell(y, d)$ with $\ell(y, d, g)$ in all the relevant definitions), with the exception of Part (b) of Proposition 3, which extends if both groups agree on the ordinal ranking of decisions based on expected losses conditional on X.[54]

7. Conclusion

We conclude with possible directions for future work, and some mention of recent work in these directions.

First, our proposed FA frontier is defined with respect to the true underlying population distribution $\mathbb{P}$. In practice, analysts may instead have access to a sample of observations from this distribution. A natural question is whether, and how, the analyst can estimate the FA frontier (or deduce its properties) from such a sample. Liu and Molinari (2024) provide a general nonparametric methodology for estimation and inference for our FA frontier. They leverage the convexity of our feasible set to construct estimators using the support functions that characterize the FA frontier. They then characterize the estimator's asymptotic distribution and show that it converges to a tight Gaussian process as the sample size grows large. Using this framework, they propose simple hypothesis tests for answering important policy questions, such as whether group identity should be banned or whether there are less discriminatory alternatives to a given algorithm. Auerbach et al. (2024) focus on the related question of how we can tell whether a given algorithm produces group errors that are on the FA frontier. This question is especially important in light of disparate impact claims, since algorithms that are shown to have disparate impact can sometimes be justified if that disparate impact is shown to be necessary to achieve other business objectives, such as accuracy. Auerbach et al. (2024) propose a sample-splitting approach, which among other things can be used to discern from a sample of observations whether a given algorithm is simultaneously fairness- and accuracy-improvable.

[53] In particular, a group-dependent loss function would blur the distinction between accuracy and fairness by explicitly incorporating fairness considerations through differentiated weighting of losses for the two groups.
[54] That is, the ranking of decisions based on $\mathbb{E}[\ell(Y, D = d, G = g) \mid X = x]$ is the same for both groups.

Second, our paper takes the population distribution $\mathbb{P}$ as exogenously given. An interesting direction for follow-up work is the question of how to optimally acquire information with fairness-accuracy objectives in mind. In such a framework, $\mathbb{P}$ would be endogenous to the information acquisition choices of the designer. Our results shed some light on certain aspects of this design. For example, suppose it were possible to acquire new covariates that turned a group-skewed covariate vector into a group-balanced covariate vector. Corollary 1 implies that such a change would change the nature of the fairness-accuracy conflict, eliminating the need to consider Pareto-dominated outcomes as a means to improve fairness. We leave to future work a more detailed exploration of endogenously chosen covariates and their fairness-accuracy consequences.

Third, in our input design problem we have assumed that the designer has knowledge of the agent's preferences. In practice, the designer may not know the agent's preferences precisely, or may face a problem of designing regulation for many agents simultaneously who hold different preferences. An interesting direction would thus be to consider optimal garblings given uncertainty over the agent's preferences, or garblings that optimize a worst-case criterion over a set of agent preferences.

Finally, our notion of equity focuses on inequality across groups rather than inequality within groups. In doing this, we are motivated by the widespread interest in algorithms' implications for between-group inequality, and the use of group-based fairness notions as justifications for various policies and laws. We thus consider it an important first step to understand what exactly these group-based fairness notions do and do not rationalize, and what tradeoffs are entailed between these fairness notions and accuracy. Inequality within groups is also an important ethical consideration, and it would be a valuable extension of our framework (and contribution to the algorithmic fairness literature more generally) to consider the more general class of designer preferences that flexibly trade off between within-group and between-group notions of inequality. We leave these questions for future work.

Appendix A. Proofs and Supporting Materials for Section 3

A.1. **Supplementary Material to Section 3.1.** Examples 9 and 12 respectively follow from Parts (a) and (b) of Proposition 3. Suppose the assumptions in Example 10 are satisfied. Given this loss function, the group-optimal point g_X is implemented by the algorithm a_g.

Moreover, we have

$$\mathbb{E}[\ell(a_b(X),Y) \mid G=b] = \mathrm{Var}(\varepsilon) = \mathbb{E}[\ell(a_r(X),Y) \mid G=r] \leq \mathbb{E}[\ell(a_b(X),Y) \mid G=r].$$

Symmetrically $\mathbb{E}[\ell(a_r(X),Y) \mid G=r] \leq \mathbb{E}[\ell(a_r(X),Y) \mid G=b]$. Note that if $a_r \neq a_b$ then both inequalities are strict and X is strictly group-balanced. Otherwise, r_X and b_X are on the 45 degree line and X is also group-balanced.

Finally, suppose the assumptions in Example 11 are satisfied. Given this loss function, the group-optimal points g_X coincide, and they are implemented by the algorithm a_0. We have

$$\begin{aligned}
\mathbb{E}[\ell(a_0(X),Y) \mid G=b] &= \sum_x \mathrm{Var}(Y \mid X=x, G=b)\mathbb{P}(x \mid G=b) \\
&= \mathrm{Var}(\varepsilon_b) > \mathrm{Var}(\varepsilon_r) \\
&= \sum_x \mathrm{Var}(Y \mid X=x, G=r)\mathbb{P}(x \mid G=r) = \mathbb{E}[\ell(a_0(X),Y) \mid G=r]
\end{aligned}$$

so the covariate vector is r-skewed.

A.2. **Characterization of the Feasible Set.**

Lemma A.1. *$\mathcal{E}_X$ is a closed and convex polygon.*

Proof. Given an algorithm $a \in \Delta(\mathcal{A})$, we slightly abuse notation and 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 $p_x \equiv \mathbb{P}(X=x)$. Then the group errors can be written as

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

where p_g is the prior probability that $G=g$. Note that this is linear in the algorithm $a \in [0,1]^{\mathcal{X}}$. Since $\Delta(\mathcal{A})$ is closed and convex with a finite number of extreme points,

$$\mathcal{E}_X = \{e(a) : a \in \Delta(\mathcal{A})\}$$

is also closed and convex. Moreover, it is a polygon. □

A.3. **Proof of Theorem 1.** Define $\overline{d} = \max_{e \in \mathcal{E}_X}(e_b - e_r)$ and $\underline{d} = \min_{e \in \mathcal{E}_X}(e_b - e_r)$ to be the maximum and minimum feasible difference between the two group errors. For every real number $d \in [\underline{d}, \overline{d}]$, consider the following feasible group error pair

$$k_d = \underset{e \in \mathcal{E}_X:\ e_b - e_r = d}{\arg\min}\ e_r.$$

These are well defined since the feasible set $\mathcal{E}_X$ is compact. Moreover, since $\mathcal{E}_X$ is convex, the points k_d vary continuously with d and trace out a part of the boundary of $\mathcal{E}_X$, which we denote by $\mathcal{K}_X$ (see Figure 11).

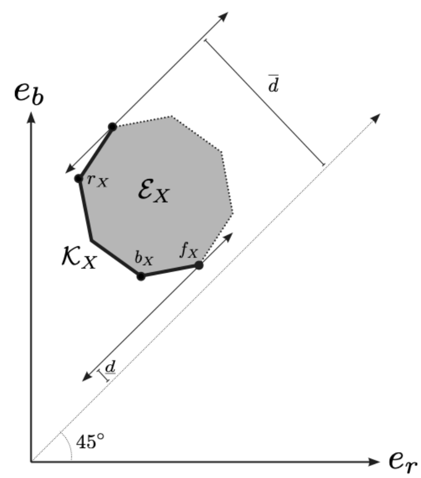


FIGURE 11. The boundary $\mathcal{K}_X$.

Note that we always have $\mathcal{P}_X \subset \mathcal{F}_X \subset \mathcal{K}_X$. $\mathcal{P}_X \subset \mathcal{F}_X$ holds since $e >_{FA} e'$ implies $e >_{PD} e'$. $\mathcal{F}_X \subset \mathcal{K}_X$ holds because if $(e_r, e_b) \in \mathcal{F}_X$ then $(e_r - \delta, e_b - \delta) \notin \mathcal{E}_X$ for any $\delta > 0$, and so $(e_r, e_b) \in \mathcal{K}_X$ by definition. Since $\mathcal{P}_X$ is the lower boundary of $\mathcal{E}_X$ between r_X and b_X, and $\mathcal{P}_X \subset \mathcal{K}_X$, the group optimal points b_X and r_X must correspond to k_{d_b} and k_{d_r} for some $d_b, d_r \in [\underline{d}, \overline{d}]$. Then $\mathcal{P}_X$ simply consists of the points k_d with $d \in [d_b, d_r]$.

Suppose X is group-balanced, in which case we need to show $\mathcal{F}_X = \mathcal{P}_X$. Given $\mathcal{P}_X \subset \mathcal{F}_X \subset \mathcal{K}_X$, it suffices to show that if $e \in \mathcal{K}_X$ belongs to $\mathcal{F}_X$, then it belongs to $\mathcal{P}_X$ as well. That is, we need to show $e = k_d$ for some $d \in [d_b, d_r]$, where $d_r \geq 0 \geq d_b$ due to group-balancedness. Assume for the sake of contradiction that $e = k_d$ with $d > d_r \geq 0$ (the case where $d < d_b \leq 0$ is symmetric). Comparing the group error pair e with r_X, we see that e has weakly higher group r error (by definition of r_X), and strictly higher difference between group b and group r errors (since $e = k_d$, $r_X = k_{d_r}$ and $d > d_r$). Thus e also has strictly higher group b error than r_X, which implies $r_X >_{FA} e$, contradicting $e \in \mathcal{F}_X$. This contradiction proves the result in the group-balanced case.

Next suppose X is r-skewed (the b-skewed case is symmetric). In this case the fairness-optimal point f_X corresponds to k_{d_f} for some $d_f \in [\underline{d}, \overline{d}]$, where $d_f = \underline{d}$ if $\underline{d} \geq 0$ and $d_f = 0$ if $\underline{d} < 0$. We need to show that $\mathcal{F}_X$ consists of the points k_d with $d \in [d_f, d_r]$ (by r-skewness, $d_r \geq d_b \geq d_f \geq 0$). To that end, we first show $\mathcal{F}_X$ is a subset of these points k_d. It suffices to show every k_d with $d \notin [d_f, d_r]$ is FA-dominated by some $k_{d'}$ with $d' \in [d_f, d_r]$. When $d > d_r \geq 0$, this can be proved in exactly the same way as in the group-balanced case. Now if $\underline{d} \geq 0$, then $d_f = \underline{d}$ and there is no point k_d with $d < d_f$; in this case we are done. If instead $\underline{d} < 0$, then $d_f = 0$ and we must additionally rule out those points k_d with $d < 0$. Let $e = k_d$ with $e_r > e_b$, and compare e with b_X. Note that b_X lies strictly above the 45-degree line (by r-skewness), whereas e lies strictly below. Moreover, the higher group error in b_X is weakly less than the lower group error in e (by definition of b_X). Thus, for any point e' on the line segment between b_X and e, it also holds that both group errors in e' are less than both group errors in e. If we choose e' to be on the 45-degree line, then $e' >_{FA} e$ and so $k_0 \geq_{FA} e' >_{FA} e$. This concludes the proof that in the r-skewed case, the only possible FA-undominated points are those k_d with $d \in [d_f, d_r]$.

It remains to show that in the r-skewed case, every point k_d with $d \in [d_f, d_r]$ is indeed FA-undominated. For $d \in [d_b, d_r]$, we have $k_d \in \mathcal{P}_X \subset \mathcal{F}_X$. For $d \in [d_f, d_b)$, if k_d is FA-dominated, then it is FA-dominated by some other $k_{d'}$. Since $d \geq d_f \geq 0$, greater fairness implies $d' < d$. As d lies between d' and d_b, we can write it as $\lambda d' + (1 - \lambda)d_b$ for some $\lambda \in (0, 1)$. Thus the feasible point $\lambda k_{d'} + (1 - \lambda)b_X$ induces the same difference of group errors as k_d. Moreover, by definition and FA-dominance, b_X and $k_{d'}$ both have weakly lower group b error than k_d, so the same is true for $\lambda k_{d'} + (1 - \lambda)b_X$. But then the definition of k_d forces these inequalities to be equal, as k_d also minimizes e_b among those points e with $e_b - e_r = d$. In particular, $k_{d'}$ must have the same group b error as k_d. Again by FA-dominance of $k_{d'}$, it must have weakly lower group r error than k_d. This implies $d' \geq d$ and leads to a contradiction, completing the proof of the theorem.

A.4. **Proof of Corollary 1.** If X is group-balanced, then $\mathcal{F}_X = \mathcal{P}_X$ by Theorem 1, so there cannot exist two points $e, e' \in \mathcal{F}_X$ such that $e >_{PD} e'$. Now suppose X is r-skewed, which means that b_X lies strictly above the 45-degree line. In this case, f_X must be weakly above the 45-degree line as well, since otherwise the intersection of the line segment $b_X f_X$ with the

45-degree line would be more fair than f_X. Let $e = b_X$ and $e' = f_X$. By definition, we have $e_b \leq e'_b$ and $e_b - e_r > e'_b - e'_r \geq 0$ (where the strict inequality follows from the assumption that $b_X \neq f_X$). This implies that

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

so $e_r < e'_r$. Thus, $e >_{PD} e'$ but $|e_b - e_r| > |e'_b - e'_r|$ as desired.

A.5. **Proof of Proposition 1.** Because $\mathcal{E}_X$ is a convex polygon, each edge can be identified by minimizing a linear functional of the form $\alpha_r e_r + \alpha_b e_b$. Specializing to $\alpha_b = 1$ and $\alpha_r = -\lambda$ reveals how these edges relate to Δ_r^x and Δ_b^x. Specifically, given any real number λ, consider minimizing $e_b - \lambda e_r$ among all feasible points $e \in \mathcal{E}_X$. Note that for any algorithm a, the resulting group errors can be written as

$$e_g(a) := \sum_x \mathbb{P}(X = x \mid G = g) \cdot \mathbb{E}[\ell(a(x), Y) \mid X = x, G = g].$$

It follows that

$$e_b(a) - \lambda e_r(a) = \sum_x \left\{\mathbb{P}(x \mid b) \cdot \mathbb{E}[\ell(a(x), Y) \mid x, b] - \lambda \mathbb{P}(x \mid r) \cdot \mathbb{E}[\ell(a(x), Y) \mid x, r]\right\},$$

where we abbreviated $\mathbb{P}(X = x \mid G = g)$ to $\mathbb{P}(x \mid g)$ and $\mathbb{E}[\ell(a(x), Y) \mid X = x, G = g]$ to $\mathbb{E}[\ell(a(x), Y) \mid x, g]$.

Thus, choosing an entire algorithm a to minimize $e_b(a) - \lambda e_r(a)$ is the same as choosing, *for each covariate realization* x, a decision $a(x) \in \Delta(\{0, 1\})$ to minimize the linear objective $\mathbb{P}(x \mid b) \cdot \mathbb{E}[\ell(a(x), Y) \mid x, b] - \lambda \mathbb{P}(x \mid r) \cdot \mathbb{E}[\ell(a(x), Y) \mid x, r]$. If we recall the definition of Δ_b^x and Δ_r^x from Section 3.3, then we immediately see that $a(x) = 0$ minimizes this objective when $\Delta_b^x - \lambda \Delta_r^x > 0$, while $a(x) = 1$ is optimal when the opposite inequality holds. In the edge case where $\Delta_b^x - \lambda \Delta_r^x = 0$, any randomized decision $a(x) \in \Delta(\{0, 1\})$ would be optimal.

We can rewrite the condition $\Delta_b^x - \lambda \Delta_r^x > 0$ as $\Delta_r^x \cdot (h(x) - \lambda) > 0$ whenever $\Delta_r^x \neq 0$, where we recall that $h(x) = \Delta_b^x / \Delta_r^x$. Thus, if the covariate x satisfies $h(x) > \lambda$, then to minimize $e_b - \lambda e_r$ we must set $a(x)$ to be 0 or 1 depending on whether $\Delta_r^x > 0$ or $\Delta_r^x < 0$. This decision $a(x)$ coincides with the r-optimal decision $a_r^*(x)$ in Section 3.3. In fact, this coincidence remains true when $h(x) = \infty$ which arises if $\Delta_r^x = 0$. In that case, the current condition for $a(x) = 1$ to be optimal, $\Delta_b^x - \lambda \Delta_r^x > 0$, reduces to $\Delta_b^x > 0$ as in our previous definition of $a_r^*(x)$. On the other hand, if the covariate x satisfies $h(x) < \lambda$, then to minimize

$e_b - \lambda e_r$ we should set $a(x)$ to be 0 or 1 depending on whether $\Delta_r^x < 0$ or $\Delta_r^x > 0$. This decision $a(x)$ is thus the opposite of $a_r^*(x)$.

In summary, we have shown that to minimize $e_b - \lambda e_r$, the optimal $a(x)$ coincides with $a_r^*(x)$ when $h(x) > \lambda$ and differs from $a_r^*(x)$ when $h(x) < \lambda$; the converse also holds. We now use this observation to prove Proposition 1. As in Section 3.3, suppose the covariate realizations are ordered so that $h(x_1) \leq h(x_2) \leq \cdots \leq h(x_n)$. For any positive integer i such that $h(x_i) < \infty$, we can choose $\lambda = h(x_i)$ and consider minimizing $e_b - \lambda e_r$. With $\lambda = h(x_i)$, we see that every x with $h(x) > \lambda$ must belong to $x_{i+1}, \ldots, x_n$, and every x with $h(x) < \lambda$ must belong to $x_1, \ldots, x_{i-1}$. If we recall the definition of the algorithms a_{i-1} and a_i from Section 3.3, then we see that both of them coincide with $a_r^*(x)$ when $h(x) > \lambda$, and differ from $a_r^*(x)$ under the opposite inequality. This proves that both algorithms a_{i-1} and a_i lead to error pairs that minimize $e_b - \lambda e_r$ for $\lambda = h(x_i)$. By linearity, any mixture $(1-\beta)a_{i-1} + \beta a_i$ with $\beta \in [0, 1]$ also minimizes this objective.

Let $n_0 = \max\{i : \ h(x_i) < 0\}$. For each $i \leq n_0$, the algorithms $\{(1-\beta)a_{i-1} + \beta a_i : \ \beta \in [0, 1]\}$ lead to error pairs that minimize $e_b - h(x_i)e_r$ across the entire feasible set $\mathcal{E}_X$, which is a convex polygon. So these error pairs must be (part of) an edge of $\mathcal{E}_X$. If we combine *all* $i \leq n_0$ and consider the resulting error pairs implemented by such algorithms, then these error pairs trace out a boundary of $\mathcal{E}_X$ that connects the error pair of a_0 and the error pair of a_{n_0}. This boundary must be part of the Pareto frontier, since any point on it minimizes some linear objective $e_b - h(x_i)e_r$ with $-h(x_i) > 0$.

By definition, the error pair implemented by $a_0 = a_r^*$ is the r-optimal point r_X. On the other hand, the algorithm a_{n_0} flips the decision of $a_r^*(x)$ precisely when $h(x) < 0$, so it satisfies the above criterion for minimizing $e_b - \lambda e_r$ with $\lambda = 0$. As the error pair of a_{n_0} minimizes e_b and belongs to the Pareto frontier, it must in fact be b_X. We have thus shown that the algorithms $\{(1-\beta)a_{i-1} + \beta a_i : \ 1 \leq i \leq n_0, \ \beta \in [0, 1]\}$ implement error pairs on the Pareto frontier that connect r_X to b_X. That is, these algorithms implement the entire Pareto frontier from r_X to b_X. This proves Proposition 1 in the group-balanced case, where $\mathcal{F}_X$ coincides with $\mathcal{P}_X$ by Theorem 1.

Now consider the r-skewed case, and let $n_1 = \max\{i : \ h(x_i) < 1\}$. By the same argument as above, the error pairs of the algorithms $\{(1-\beta)a_{i-1} + \beta a_i : \ 1 \leq i \leq n_1, \ \beta \in [0, 1]\}$ (where we now allow $i \leq n_1$) trace out a boundary of the feasible set that connects r_X to the error pair of a_{n_1}. Note that similar to a_{n_0} implementing b_X, the algorithm a_{n_1} leads to the error pair that minimizes $e_b - e_r$, breaking ties in favor of smaller e_b and e_r.

When $e_b > e_r$ at f_X (and thus at every feasible point), this error pair of a_{n_1} is exactly f_X since minimizing the difference $e_b - e_r$ would be the same as minimizing the absolute difference. In this case, the set of algorithms $\{(1-\beta)a_{i-1} + \beta a_i : \ 1 \leq i \leq n_1, \ \beta \in [0,1]\}$ implement a boundary of the feasible set that connects r_X to f_X. To ensure this boundary is $\mathcal{F}_X$ and not the other boundary connecting r_X to f_X, note that for $1 \leq i \leq n_0$, such an algorithm implements a point on the Pareto frontier as we already showed. Whereas for $n_0+1 \leq i \leq n_1$, the algorithm $(1-\beta)a_{i-1}+\beta a_i$ minimizes $e_b - \lambda e_r$ for some $\lambda = h(x_i) \in [0,1)$. Since $e_b - \lambda e_r = (e_b - e_r) + (1-\lambda)e_r$, any other feasible point necessarily has strictly larger $e_b - e_r$ (thus strictly larger $|e_b - e_r|$) or strictly larger e_r. This shows that every feasible point minimizing $e_b - \lambda e_r$ belongs to the FA frontier. It follows that the above set of algorithms implement the entire FA frontier from r_X to f_X, as stated in part (b) of Proposition 1.

Finally, suppose X is r-skewed and $e_b = e_r$ at f_X. To prove part (b) in this case, we only need to show that f_X is implemented by *some* algorithm in the above set, and then we can define n^* and $\overline{\beta}$ accordingly. Without loss assume b_X is strictly above the 45-degree line (that is, $e_b > e_r$ at b_X), since otherwise $f_X = b_X$ and we are done. Moreover, as we currently assume f_X is on the 45-degree line, the error pair of a_{n_1} (which minimizes $e_b - e_r$) must be weakly below the 45-degree line. So the error pairs implemented by the algorithms $\{(1-\beta)a_{i-1} + \beta a_i : \ n_0 + 1 \leq i \leq n_1, \ \beta \in [0,1]\}$ connect b_X to the error pair of a_{n_1}, and must cross the 45-degree line at some point f which we will show to be the fairness-optimal point f_X. Indeed, since f minimizes $e_b - \lambda e_r$ for some $\lambda = h(x_i) \in [0,1)$, it clearly minimizes $(1-\lambda)e_b$ subject to $e_b = e_r$. This shows $f = f_X$ and completes the proof.

A.6. **Proof of Proposition 2.** Following the formulation in footnote 26, let $\hat{\mathcal{X}} = \mathcal{X} \times \{r, b\}$, where the final covariate is the group identity, and $\hat{a}(x,g) = a_g(x)$ for every $x \in \mathcal{X}$. As in the proof of Lemma A.1, we can rewrite the group error pair as

$$e(\hat{a}) = \sum_{\hat{x} \in \hat{\mathcal{X}}} p_{\hat{x}} \left(\hat{a}(\hat{x}) c_1(\hat{x}) + (1 - \hat{a}(\hat{x})) c_0(\hat{x}) \right)$$

where $c_i(\hat{x}) := \left(\sum_y \frac{\hat{x}_{y,r}}{p_r} \ell(i,y), \sum_y \frac{\hat{x}_{y,b}}{p_b} \ell(i,y) \right) \in \mathbb{R}^2$. Further let $\mathcal{X}_g \equiv \mathcal{X} \times \{g\}$ be the realizations of the covariate vector for members of group g. We then have

$$\mathcal{E}_X = \{ e(\hat{a}) \ : \ \hat{a} \in \Delta(\mathcal{A}) \}$$

$$= \left\{ \sum_{x\in\mathcal{X}} p_{(x,r)} \left(a_r(x)\, c_1(x,r) + (1-a_r(x))\, c_0(x,r)\right) \right.$$
$$\left. + \sum_{x\in\mathcal{X}} p_{(x,b)} \left(a_b(x)\, c_1(x,b) + (1-a_b(x))\, c_0(x,b)\right) \;:\; a_r \in \Delta(\mathcal{A}_r), a_b \in \Delta(\mathcal{A}_b) \right\}$$

Note that this is the Minkowski addition of two sets. Since $\hat{x}_{y,b} = 0$ for all $\hat{x} \in \mathcal{X}_r$, $c_1(\hat{x})$ and $c_0(\hat{x})$ are all points on the vertical axis for all $\hat{x} \in \mathcal{X}_r$. By symmetric reasoning, $c_1(\hat{x})$ and $c_0(\hat{x})$ are all points on the horizontal axis for all $\hat{x} \in \mathcal{X}_b$. Since $\mathcal{E}_X$ is the Minkowski addition of a set on the vertical axis and a set on the horizontal axis, it must be a rectangle. Moreover, $r_X = b_X$ must be its bottom-left vertex.

Finally, 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.

A.7. **Proof of Proposition 3.** Part (a): We continue to follow the notation laid out in the proof of Lemma A.1. Note that under strong independence,

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

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.

Part (b): We will show that $b_X = r_X$ under conditional independence, and the result then follows from Theorem 1. Recall from the proof of Lemma A.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, $x_{y,g} = x_y x_g$ (with $x_y := \mathbb{P}(Y = y \mid X = x)$ and $x_g := \mathbb{P}(G = g \mid X = x)$) 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 the decision $d = 1$ is optimal for both groups (and vice-versa if the inequality is reversed). Consider the following algorithm:

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

This algorithm will deliver the lowest error for both groups and

$$(e_r(a^*), e_b(a^*)) = r_X = b_X$$

as desired.

Appendix B. Proofs and Supporting Materials to Section 4

B.1. **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 chooses, the agent's payoff will be weakly higher than if given 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 direction $\mathcal{E}_X^* \subseteq \mathcal{E}_X \cap H$.

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 $d = 1$:[55]

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

where $p_g \equiv \mathbb{P}(G = g)$ and $p_{x,y,g} \equiv \mathbb{P}(X = x, Y = y, G = g)$. The above is just equation (2) 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 $\sum_x p_{x,y,g} \cdot a(x)$.

Let us rewrite the above displayed inequality as

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

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

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

Now, the LHS above can be rewritten as $\sum_{x,y,g} p_{x,y,g} \frac{\alpha_g}{p_g} \cdot \mathbb{E}_{D \sim a(x)}[\ell(D, y) \mid X = x, Y = y, G = g]$, which is also equal to $\sum_g \alpha_g \cdot \mathbb{E}_{D \sim a(x)}[\ell(D, Y) \mid G = g]$. 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.

[55] 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 $\mathcal{F}^*_X = \mathcal{F}_X \cap H$. 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 $\mathcal{F}_X \cap H \subseteq \mathcal{F}^*_X$. In the opposite direction, suppose for contradiction that a certain point $(e_r, e_b) \in \mathcal{F}^*_X$ does not belong to $\mathcal{F}_X \cap H$. Since $\mathcal{F}^*_X \subseteq \mathcal{E}^*_X \subseteq H$, we know that (e_r, e_b) must not belong to $\mathcal{F}_X$. Thus by definition of $\mathcal{F}_X$, (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 $\alpha_r, \alpha_b > 0$ and the second uses $(e_r, e_b) \in \mathcal{F}^*_X \subseteq \mathcal{E}^*_X$). 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.2. **Supplementary Material to Section 4.3.1.** In this section, we compute the input-design feasible set and fairness-accuracy frontier for the example in Section 4.3.1. 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 $\mathcal{F}^*_{X,X'}$, as illustrated in Figure 7 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, 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.

B.3. **Proof of Proposition 4.** We now deduce Proposition 4 from Lemma 1. If X is group-balanced, then by Theorem 1 we know that $\mathcal{F}_X$ is the part of the boundary of $\mathcal{E}_X$ that connects r_X to b_X from below. Clearly, $\mathcal{F}^*_X = \mathcal{F}_X$ can only hold if $r_X, b_X \in \mathcal{F}^*_X \subseteq H$, 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 m_X be the point that has the same e_r as r_X and the same e_b as b_X. Geometrically, m_X is such that the line segments $m_X r_X$ and $m_X b_X$ are parallel to the axes. Because r_X, b_X have respectively minimal group errors in the feasible set $\mathcal{E}_X$, and because we are considering the lower boundary, any point on this lower boundary $\mathcal{F}_X$ must belong to the triangle with vertices r_X, b_X and m_X. Since r_X, b_X, m_X all belong to the halfspace H ($m_X \in H$ because the agent's payoff weights α_r, α_b are non-negative), we deduce that $\mathcal{F}_X \subseteq H$. Hence whenever $r_X, b_X \in H$, we have by Lemma 1 that $\mathcal{F}^*_X = \mathcal{F}_X \cap H = \mathcal{F}_X$. This argument proves Proposition 4 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. (Throughout we assume e_r to be on the x-axis.) 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.[56] So across all points on this part of the lower boundary, f_X maximizes $\alpha_r e_r + \alpha_b e_b$. Thus the assumption $f_X \in H$ implies that the lower boundary from b_X to f_X belongs to H. In particular $b_X \in H$, which together with $r_X \in H$ implies that the lower boundary from r_X to b_X also belongs to H (by the same argument as in the group-balanced case before). Hence the entire lower boundary from r_X to f_X belongs to H, as we desire to show.

B.4. **Proof of Proposition 5.** We first present a simple lemma which conveniently restates the property of FA-dominance for frontiers:

[56] 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.

Lemma B.1. *$\mathcal{F}^*_{X,X'} >_{FA} \mathcal{F}^*_X$ if and only if $\mathcal{F}^*_X$ does not intersect with $\mathcal{F}^*_{X,X'}$.*

The proof of this lemma is straightforward: If there exists a point in $\mathcal{F}^*_X$ that also belongs to $\mathcal{F}^*_{X,X'}$, then this point is not FA-dominated by any point in $\mathcal{F}^*_{X,X'}$, so $\mathcal{F}^*_{X,X'} \not>_{FA} \mathcal{F}^*_X$. Suppose instead that no point in $\mathcal{F}^*_X$ belongs to $\mathcal{F}^*_{X,X'}$. Since every point in $\mathcal{F}^*_X$ is implementable via a garbling of X, and thus also implementable via a garbling of X, X', we have $\mathcal{F}^*_X \subseteq \mathcal{E}^*_{X,X'}$. But since no point in $\mathcal{F}^*_X$ belongs to $\mathcal{F}^*_{X,X'}$ (by assumption), it must be that every point in $\mathcal{F}^*_X$ is FA-dominated by another point in the (compact) set $\mathcal{E}^*_{X,X'}$. Thus $\mathcal{F}^*_{X,X'} >_{FA} \mathcal{F}^*_X$ as desired.

Below we use Lemma B.1 to deduce Proposition 5. 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 4 tells us that the input-design fairness-accuracy frontier $\mathcal{F}^*_X$ is the same as the unconstrained fairness-accuracy frontier $\mathcal{F}_X$, 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.1, 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 the input-design fairness-accuracy frontier $\mathcal{F}^*_{X,G}$ given (X, G). To characterize the latter frontier, let $m_X = r_{X,G} = b_{X,G}$ 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 m_X lies weakly above the 45-degree line. Then from Proposition 2 we know that the unconstrained fairness-accuracy frontier $\mathcal{F}_{X,G}$ is the horizontal line segment from m_X to $f_{X,G}$. This point $f_{X,G}$ is the intersection between the line segment $m_X b_X$ and the 45-degree line (here we use the fact that m_X lies above the 45-degree line and b_X lies below). As $b_X \in H$, the points m_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 $\mathcal{F}^*_{X,G}$ is also the line segment from m_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 $\mathcal{F}_{X,G}$ is the horizontal line segment from m_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 m_X and $f_{X,G}$. In other words, $b_X \in \mathcal{F}_X$ and $b_X \in \mathcal{F}_{X,G}$. 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 $\mathcal{F}^*_X$ and $\mathcal{F}^*_{X,G}$. This shows that the two frontiers $\mathcal{F}^*_X$ and $\mathcal{F}^*_{X,G}$ intersect, which completes the proof by Lemma B.1.

B.5. **Proof of Proposition 6.** By Proposition 2, the fairness-accuracy frontier $\mathcal{F}_{X,G}$ is a line segment whose left endpoint is $r_{X,G} = b_{X,G}$ and right endpoint is $f_{X,G}$, with e_b being constant on this frontier. We will first identify the optimal points for the designer in the unconstrained problem (i.e., if the designer were given full control over the algorithm, as in our Section 2.1), and then show that the garblings indicated in the result implement these points.

Suppose first that (X, G) is group-balanced. Then $\mathcal{F}_{X,G}$ is the singleton $r_{X,G} = b_{X,G} = f_{X,G}$, which is optimal for all FA preferences. So clearly this point is the designer's preferred point given any simple preference.

Now suppose (X, G) is r-skewed. Then the frontier $\mathcal{F}_{X,G}$ consists entirely of points (e_r, e_b) satisfying $e_b \geq e_r$. So maximizing the designer's payoff $-\gamma_r e_r - \gamma_b e_b - \gamma_f |e_r - e_b|$ over error pairs on the FA frontier is equivalent to solving

$$\max_{(e_r,e_b)\in\mathcal{F}_{X,G}} (\gamma_f - \gamma_r)e_r - (\gamma_b + \gamma_f)e_b.$$

Since e_b is constant on $\mathcal{F}_{X,G}$, the designer simply maximizes $(\gamma_f - \gamma_r)e_r$. If $\gamma_f < \gamma_r$, then the designer minimizes e_r and thus $r_{X,G} = b_{X,G}$ is the uniquely optimal point for the designer. If $\gamma_f = \gamma_r$, then all points on $\mathcal{F}_{X,G}$ are optimal. Finally if $\gamma_f > \gamma_r$, then the designer's weight on e_r is strictly positive, so that $f_{X,G}$ is the uniquely optimal point for the designer.

It remains to determine how the designer can implement these points with garblings. Again start with the group-balanced case. If the designer sends the fully revealing garbling, then the agent's feasible set is $\mathcal{E}_{X,G}$. Since by assumption the agent's preferences can be written as $-\alpha_r e_r - \alpha_b e_b$ for $\alpha_r, \alpha_b > 0$, the agent optimally chooses $r_{X,G} = b_{X,G}$ as desired by the designer. Moreover, when (X, G) is group-balanced,

$$\beta = \max\left\{\frac{\overline{e}_r - \underline{e}_b}{\overline{e}_r - \underline{e}_r},\ 0\right\} = 1$$

since $\underline{e}_r = \underline{e}_b$ at $r_{X,G} = b_{X,G} = f_{X,G}$. Thus the r-shaded garbling maps each (x, g) to the message $a_g^*(x)$ with probability 1, which the agent optimally obeys. So again the agent optimally chooses $r_{X,G}$. This means that both the fully revealing garbling and the r-shaded garbling induce the agent to choose an algorithm that implements the designer's preferred point, and are thus both optimal.

Now consider the r-skewed case. By the same logic as for the group-balanced case, sending the agent the fully revealing signal leads the agent to choose $r_{X,G}$, which is optimal for the designer when $\gamma_f \leq \gamma_r$. This yields part (a) of the result. Part (b) of the result follows by noticing that the r-shaded garbling is precisely the garbling constructed in the proof of Lemma 1, which recommends an action to the agent and leads to the error pair $(\underline{e}_b, \min\{\underline{e}_b, \overline{e}_r\})$ assuming that the agent follows the recommendation. By Proposition 2 this error pair is the point $f_{X,G}$, which we assume belongs to H. So by the proof of Lemma 1, the agent optimally follows the recommendation and chooses $f_{X,G}$, which is optimal for designers with $\gamma_f \geq \gamma_r$. This completes the proof.

B.6. **Proof of Proposition 7.** Let $\underline{e}_g = \min\{e_g \mid e \in \mathcal{E}_{X,G}\}$ and $\overline{e}_g = \max\{e_g \mid e \in \mathcal{E}_{X,G}\}$ be the minimal and maximal feasible errors for group g given covariate vector (X, G), and define $\underline{e}_g^* = \min\{e_g \mid e \in \mathcal{E}_{X,G,X'}\}$ and $\overline{e}_g^* = \max\{e_g \mid e \in \mathcal{E}_{X,G,X'}\}$ to be the corresponding quantities given (X, G, X'). The following lemma says that additional access to X' reduces the minimal feasible error for group g relative to (X, G) if and only if X' is decision-relevant over X for group g.

Lemma B.2. *$\underline{e}_g^* < \underline{e}_g$ 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.,

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

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

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

By optimality of a_g^*, for all $x \in \mathcal{X}$ and $x' \in \mathcal{X}'$,

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

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 $(x,\tilde{x}')$. By taking the expectation of (B.2) conditional on $G=g$, we obtain

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

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

We now use Lemma B.1 and B.2 to prove Proposition 7. First suppose (X,G) is r-skewed, in which case $r_X = b_X$ lies strictly above the 45-degree line. By Proposition 2, the unconstrained fairness-accuracy frontier $\mathcal{F}_{X,G}$ is then the horizontal line segment from $r_{X,G} = b_{X,G}$ to $f_{X,G}$.

If X' is not decision-relevant over X for group b, then from Lemma B.2 we know that the minimal feasible error for group b is the same given (X,G,X') as given (X,G). By assumption that (X,G) is r-skewed, group b's minimal error given (X,G) exceeds group r's minimal error given (X,G). Since group b's minimal error is the same given (X,G) and (X,G,X'), while group r's minimal error is weakly smaller given (X,G,X') compared to (X,G), it must be that group b minimal error given (X,G,X') also exceeds the group r minimal error given (X,G,X'). In other words, $r_{X,G,X'} = b_{X,G,X'}$ also lies strictly above the 45-degree line, and the fairness-accuracy frontier $\mathcal{F}_{X,G,X'}$ is the horizontal line segment from $r_{X,G,X'} = b_{X,G,X'}$ to $f_{X,G,X'}$. Crucially, this line segment shares the same e_b as the line segment from $r_{X,G} = b_{X,G}$ to $f_{X,G}$. In addition, as $r_{X,G,X'}$ must have weakly smaller e_r than $r_{X,G}$, and $f_{X,G,X'}$ must be weakly closer to the 45-degree line than $f_{X,G}$, we deduce that the unconstrained fairness-accuracy frontier $\mathcal{F}_{X,G,X'}$ is a horizontal line segment that is a superset of the line segment $\mathcal{F}_{X,G}$. Thus, in particular, $r_{X,G} = b_{X,G}$ belongs to both of these frontiers. Lemma 1 thus imply that $r_{X,G} = b_{X,G}$ also belongs to the input-design fairness-accuracy frontiers $\mathcal{F}^*_{X,G}$ and $\mathcal{F}^*_{X,G,X'}$ ($r_{X,G} = b_{X,G}$ belongs to H because this point can be implemented by giving (X, G) to the agent, who will then minimize both groups' errors given this information). The result then follows from Lemma B.1.

If X' is decision-relevant over X for group b, then Lemma B.2 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, G, X') is r-skewed just as (X, G) is. Then the unconstrained fairness-accuracy frontier $\mathcal{F}_{X,G,X'}$ 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 $\mathcal{F}_{X,G}$. Thus $\mathcal{F}_{X,G}$ does not intersect $\mathcal{F}_{X,G,X'}$. As their subsets, the input-design fairness-accuracy frontiers $\mathcal{F}^*_{X,G}$ and $\mathcal{F}^*_{X,G,X'}$ also do not intersect. Thus the result follows from Lemma B.1. In the remaining case we have $\underline{e}_b^* \leq \underline{e}_r^*$, so that (X, G, X') is b-skewed. Then the unconstrained fairness-accuracy frontier $\mathcal{F}_{X,G,X'}$ is now a *vertical* line segment with $e_r = \underline{e}_r^*$. The points on this frontier have varying e_b, but any of the e_b does not exceed $\underline{e}_r^*$ because these points are below the 45-degree line. Because $\underline{e}_r^* \leq \underline{e}_r < \underline{e}_b$, we thus know that any point on the frontier $\mathcal{F}_{X,G,X'}$ has strictly smaller e_b compared to any point on $\mathcal{F}_{X,G}$. Once again these two unconstrained frontiers do not intersect, and nor do the input-design frontiers. This proves Proposition 7 when (X, G) is r-skewed.

A symmetric argument applies when (X, G) is b-skewed, so below we focus on the case where (X, G) is group-balanced. That is, $r_{X,G} = b_{X,G}$ lies on the 45-degree line. In this case the fairness-accuracy frontiers $\mathcal{F}_{X,G}$ and $\mathcal{F}^*_{X,G}$ are both this singleton point. If X' is not decision-relevant over X for group b, then Lemma B.2 tells us that $\underline{e}_b^* = \underline{e}_b = \underline{e}_r \geq \underline{e}_r^*$. When equality holds the fairness-accuracy frontiers $\mathcal{F}_{X,G,X'}$ and $\mathcal{F}^*_{X,G,X'}$ are also the singleton point $r_{X,G} = b_{X,G}$, and the result trivially holds. If we instead have strict inequality $\underline{e}_b^* = \underline{e}_b > \underline{e}_r^*$, then (X, G, X') is r-skewed and the unconstrained fairness-accuracy frontier $\mathcal{F}_{X,G,X'}$ is a horizontal line segment with one of the endpoints being $f_{X,G,X'} = r_{X,G} = b_{X,G}$. Thus $r_{X,G} = b_{X,G}$ belongs also to the input-design fairness-accuracy frontier $\mathcal{F}^*_{X,G,X'}$, showing that $\mathcal{F}^*_{X,G}$ and $\mathcal{F}^*_{X,G,X'}$ intersect. The result again follows from Lemma B.1.

Conversely, suppose X' is decision-relevant over X for both groups. Then by Proposition 2, the unconstrained frontier $\mathcal{F}_{X,X'}$ 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 $\mathcal{F}_X$ does not intersect with $\mathcal{F}_{X,X'}$. Hence $\mathcal{F}^*_X$ and $\mathcal{F}^*_{X,X'}$ also do not intersect, and Lemma B.1 concludes the proof. This completes the proof of Proposition 7.

B.7. **Proof of Corollary 2.** Omitted since it follows immediately from Proposition 6.

Appendix C. A Minimal Class Characterization of the FA Frontier

Here we show that the FA frontier can be fully recovered from the smaller class of parameterized utility functions that linearly trade off accuracy and fairness. Specifically, consider the class of functions

$$w(e) = -\gamma_r e_r - \gamma_b e_b - \gamma_f |e_r - e_b|$$

where $\gamma_r, \gamma_b > 0$ and $\gamma_f \geq 0$ are (respectively) the designer's weights on accuracy for each group and on fairness. A preference $\succeq$ is *simple* if it can be represented by a utility of this form, i.e. $e \succeq e'$ if and only if $w(e) \geq w(e')$. All simple preferences are FA preferences, but not all FA preferences are simple. For example, both Utilitarian and Rawlsian preferences are simple but Egalitarian preferences are not.[57] Given any preference $\succeq$ on error pairs, let

$$\mathcal{C}_X(\succeq) := \{e \in \mathcal{E}_X : e \succeq e' \text{ for all } e' \in \mathcal{E}_X\}$$

denote the set of $\succeq$-optimal points.

Proposition C.1. *The following statements are equivalent:*

(1) $e \in \mathcal{F}_X$.
(2) $e \in \mathcal{C}_X(\succeq)$ *for some FA preference* $\succeq$.
(3) $\{e\} = \mathcal{C}_X(\succeq)$ *for some FA preference* $\succeq$.
(4) $e \in \mathcal{C}_X(\succeq)$ *for some simple FA preference* $\succeq$.

The equivalence between (1) and (4) is in the spirit of Wald and Wolfowitz (1951)'s complete class theorem, and provides a class of utility functions which is sufficient for generating the FA frontier. Recalling that the usual Pareto frontier can be characterized as the set of optimal points for designers with utility functions of the form $\gamma_r e_r + \gamma_b e_b$, our characterization differs in additionally including the term $\gamma_f |e_r - e_b|$, which captures a concern for inequity across groups. As it turns out, the addition of this term is sufficient to characterize the entire FA frontier, so the frontier can be traced out simply by shifting the parameter weights γ_r, γ_b and γ_f.

[57]To see this for the Utilitarian designer, set $\gamma_r = p_r$, $\gamma_b = p_b$ and $\gamma_f = 0$. To see this for the Rawlsian designer, set $\gamma_r = \gamma_b = \gamma_f = 1$. Egalitarian preferences are not simple as they are not continuous.

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 $e \in \mathcal{C}_X(\succeq)$ for some FA preference $\succeq$ but $e \notin \mathcal{F}_X$. Thus, there exists some $e' >_{FA} e$ so $e' \succ e$ yielding a contradiction.

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

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

and let $\succeq$ be the corresponding preference. We will show that $\succeq$ is an FA preference. Suppose $e >_{FA} e'$ so $e >_{PD} e'$ which implies $\psi(e_r + e_b) > \psi(e'_r + e'_b)$. 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

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

Thus, $\succeq$ is an FA preference. Now, since $e^* \in \mathcal{F}_X$, 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 $\{e^*\} = \mathcal{C}_X(\succeq)$. 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 $e^* \in \mathcal{F}_X$, 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,[58] 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 (γ_r, γ_b) that is normal to this edge, such that e^* maximizes $\gamma_r e_r + \gamma_b e_b$ among all feasible points. Since this edge has negative slope, it is straightforward to see that $\gamma_r, \gamma_b < 0$. So e maximizes the simple utility $\gamma_r e_r + \gamma_b e_b$ 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

[58] Again we assume e_r to be on the x-axis in applying the definition of lower boundary.

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 γ_f to belong to the interval $(\beta_b, -\beta_r)$, which is in particular negative. Further define $\gamma_r = \beta_r + \gamma_f < 0$ and $\gamma_b = \beta_b - \gamma_f < 0$. Then $\beta_r e_r + \beta_b e_b$ can be rewritten as $\gamma_r e_r + \gamma_b e_b + \gamma_f(e_b - e_r)$. If we consider the simple utility $\gamma_r e_r + \gamma_b e_b + \gamma_f |e_b - e_r|$, then for any other feasible point e^{**} it holds that

$$\begin{aligned}
\gamma_r e_r^{**} + \gamma_b e_b^{**} + \gamma_f |e_b^{**} - e_r^{**}| &\leq \gamma_r e_r^{**} + \gamma_b e_b^{**} + \gamma_f (e_b^{**} - e_r^{**}) \\
&= \beta_r e_r^{**} + \beta_b e_b^{**} \\
&\leq \beta_r e_r^{*} + \beta_b e_b^{*} \\
&= \gamma_r e_r^{*} + \gamma_b e_b^{*} + \gamma_f (e_b^{*} - e_r^{*}) \\
&= \gamma_r e_r^{*} + \gamma_b e_b^{*} + \gamma_f |e_b^{*} - e_r^{*}|,
\end{aligned}$$

where the first inequality holds since $\gamma_f \leq 0$ and the last equality holds because $e^* \in \mathcal{F}_X$ 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. □

Appendix D. Supplementary Material to Section 5

D.1. **Bootstrap algorithm.** Here we describe the bootstrap algorithm used in Section 5. Let $(Y_i, X_i, G_i)_{i=1}^n$ be the test data and $\hat{P}$ be its empirical distribution. Fix a sample split and any algorithm obtained from the training data. To test the null hypothesis (5) against its alternative, we first compute the test statistic $\hat{T} \coloneqq \hat{e}_b^b - \hat{e}_w^b$, where $\hat{e}_g^b \coloneqq \frac{1}{n_g} \sum_{i:\, G_i = g} \ell(a(X_i), Y_i)$ is the sample estimate of e_g^b on the test data. Let $(Y_i^*, X_i^*, G_i^*)_{i=1}^n$ be a bootstrap sample drawn i.i.d. from the test data and denote the bootstrap analogue of $\hat{T}$ by $\hat{T}^*$. The critical value for the test is based on the quantile of the bootstrap distribution of the test statistics. Specifically, let Ψ be the (conditional) cumulative distribution function of $\hat{T}^* - \hat{T}$ given $\hat{P}$. In practice, these are not known exactly, but can be approximated via Monte-Carlo by taking Q independent bootstrap samples from $\hat{P}$ and computing the empirical distribution

$$\hat{\Psi}(x) \coloneqq \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{1}\{\hat{T}_q^* - \hat{T} \leq x\},$$

where $\hat{T}_q^*$ denotes the bootstrap test statistic computed on the qth sample. The p-value is given by $1 - \hat{\Psi}(\hat{T})$.

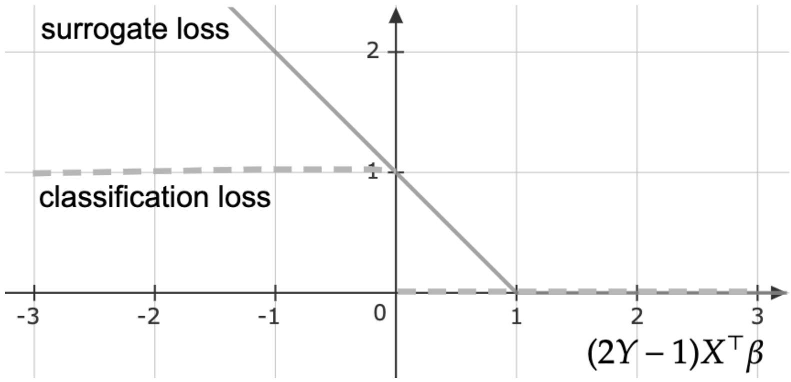


FIGURE 12. Classification and surrogate loss

To test the null hypothesis (6) against its alternative, let $\hat{T}^W$ and $\hat{T}^M$ be the empirical analogues of $|e_w^w - e_m^w| - \delta$ and $|e_w^m - e_m^w| - \delta$, respectively. The bootstrap analogues $\hat{T}^{W*}$ and $\hat{T}^{M*}$ and the cdfs $\hat{\Psi}^W$ and $\hat{\Psi}^M$ are defined in the same way as before. The p-value is then given by $\max\{\hat{\Psi}^W(\hat{T}^W), \hat{\Psi}^M(\hat{T}^M)\}$.

D.2. **Estimation of Feasible Sets via Linear Programming.** First, let us examine the case where α_r and α_b are nonnegative. When we compute the group errors, we substitute the classification loss $\mathbb{1}\left\{\mathbb{1}\{\beta^\top x \geq 0\} \neq y\right\} = 1 - \mathbb{1}\left\{(2y-1)\beta^\top x > 0\right\}$ (for $\beta^\top x \neq 0$) with the hinge surrogate loss $\max\left\{1 - (2y-1)\beta^\top x, 0\right\}$ (see also Figure 12.) This allows us to formulate the following alternative LP with linear constraints (without integer constraints):

$$
\text{(LP)} \quad \left[\begin{array}{lll}
\min\limits_{\beta, \hat{e}_r, \hat{e}_b, w} & \alpha_r \hat{e}_r + \alpha_b \hat{e}_b & \\
\text{s.t.} & \hat{e}_r = \dfrac{1}{n_r} \sum\limits_{i:\, G_i = r} w_i + \lambda \|\beta\|_1 & \\
 & \hat{e}_b = \dfrac{1}{n_b} \sum\limits_{i:\, G_i = b} w_i + \lambda \|\beta\|_1 & \\
 & w_i \geq 1 - (2Y_i - 1) X_i^\top \beta & \text{(for all } i) \\
 & w_i \geq 0 & \text{(for all } i) \\
 & \beta \in \mathbb{R}^{\dim \mathcal{X}}, \ \hat{e}_r \geq 0, \ \hat{e}_b \geq 0 &
\end{array}\right.
$$

Note that we also allow for the option of adding L_1 regularization into the definition of the sample error rates e_b and e_r in the definition of LP. Such regularization could be important to avoid over-fitting to the training data when X is high-dimensional. We set $\lambda := 0.001$ for the Obermeyer et al. (2019) dataset, and $\lambda := 0$ for the Strack et al. (2014) dataset.

When either α_r or α_b is strictly negative, we need a slight adjustment to the above LP. Suppose that $\alpha_r < 0$ and $\alpha_b \geq 0$ (other cases can be considered analogously.) The LP defined above becomes unbounded and has no optimal solution: for any $M \in \mathbb{R}$, there exists a feasible solution $(\beta, \hat{e}_r, \hat{e}_b, w)$ such that $\alpha_r \hat{e}_r + \alpha_b \hat{e}_b < M$. To see this is possible, observe that we can make the value of the objective function arbitrarily small by increasing w_i for some i with $G_i = r$, which increases $\hat{e}_r$ (note that w_i is not bounded from above).

To address the issue, we perform a change of variables before defining LP: let $\tilde{\alpha}_r \coloneqq -\alpha_r > 0$ and $\tilde{e}_r \coloneqq 1 - \hat{e}_r$. Then, we have $\alpha_r \hat{e}_r + \alpha_b \hat{e}_b = \tilde{\alpha}_r \tilde{e}_r + \alpha_b \hat{e}_b + \tilde{\alpha}_r$. Since $\tilde{\alpha}_r$ is a constant, the original optimization problem

$$\min \left\{ \alpha_r \hat{e}_r(a) + \alpha_b \hat{e}_b(a) \colon \forall g,\ \hat{e}_g(a) = \frac{1}{n_g} \sum_{i\colon G_i = g} \ell(a(X_i), Y_i),\ a \in \mathcal{A} \right\},$$

is equivalent to the following: [59]

$$\min \left\{ \tilde{\alpha}_r \tilde{e}_r(a) + \alpha_b \hat{e}_b(a) \colon \forall g,\ \hat{e}_g(a) = \frac{1}{n_g} \sum_{i\colon G_i = g} \ell(a(X_i), Y_i),\ \tilde{e}_r(a) = 1 - \hat{e}_r(a),\ a \in \mathcal{A} \right\},$$

As before, since we consider the class of linear algorithms, the latter optimization problem can be written as an MILP, which is the same as the previous MILP except for the constraint regarding $\hat{e}_r$: the constraint regarding $\hat{e}_r$ in the previous MILP is replaced by the following:

$$\tilde{e}_r = 1 - \hat{e}_r = \frac{1}{n_r} \sum_{i\colon G_i = r} \left(1 - \mathbb{1}\left\{\hat{Y}_i \neq Y_i\right\}\right) = \frac{1}{n_r} \sum_{i\colon G_i = r} \mathbb{1}\left\{\hat{Y}_i = Y_i\right\} = \frac{1}{n_r} \sum_{i\colon G_i = r} \mathbb{1}\left\{\hat{Y}_i \neq \tilde{Y}_i\right\},$$

where $\tilde{Y}_i \coloneqq \mathbb{1}\{Y_i = 0\}$ (the label is flipped.) This means that the new MILP has exactly the same form as the previous MILP if we flip the labels of group r. Then, the new MILP can be reduced to the LP by the same procedure as before.

In general, we can derive the LP by the following procedure:

(1) Formulate the LP as in the original case where $\alpha_b > 0$ and $\alpha_r > 0$;
(2) For group g with $\alpha_g < 0$, flip the labels of group-g observations, i.e., use $\tilde{Y}_i \coloneqq \mathbb{1}\{Y_i = 0\}$ for group g as if it were the true outcome variable;
(3) Replace α_g in the objective function with $|\alpha_g|$ for each g.

[59] The two problems are equivalent in the sense that the solutions of the two optimization problem have a clear one-to-one mapping: $1 - \hat{e}_r = \tilde{e}_r$.

Data Availability

The code that generates the tables and figures can be found in the Harvard Dataverse, https://doi.org/10.7910/DVN/4HURK4 (Liang et al., 2024).

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.

Agarwal, A., A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach (2018): "A Reductions Approach to Fair Classification," in *Proceedings of the 35th International Conference on Machine Learning: PMLR 80*, 60–69.

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

Andreoni, J. and J. Miller (2002): "Giving According to GARP," *Econometrica*, 70, 737–753.

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.

Auerbach, E., A. Liang, K. Okumura, and M. Tabord-Meehan (2024): "Testing Fairness-Accuracy Improvability," Working Paper.

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

Bertrand, M. and E. Kamenica (2023): "Coming Apart? Cultural Distances in the United States Over Time," *American Economic Journal: Applied Economics*, 15, 100–141.

Blattner, L., S. Nelson, and J. Spiess (2024): "Unpacking the Black Box: Regulating Algorithmic Decisions," Working Paper.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Caplin, A., D. Martin, and P. Marx (2024): "Modeling Machine Learning: A Cognitive Economic Approach," Working Paper.

Cerdeña, J. P., M. V. Plaisime, and J. Tsai (2020): "From Race-based to Race-conscious Medicine: How Anti-racist Uprisings Call Us to Act," *The Lancet*, 396, 1125–1128.

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

Charness, G. and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *The Quarterly Journal of Economics*, 117, 817–869.

Che, Y.-K., K. Kim, and W. Zhong (2019): "Statistical Discrimination in Ratings-Guided Markets," Working Paper.

Chohlas-Wood, A., M. Coots, H. Zhu, E. Brunskill, and S. Goel (2024): "Learning to be Fair: A Consequentialist Approach to Equitable Decision-Making," *Management Science*, forthcoming.

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

Corbett-Davis, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017): "Algorithmic Decision-making and the Cost of Fairness," in *KDD '17: Proceedings of the 23rd ACM Conference on Knowledge Discovery and Data Mining*, 797–806.

Cowgill, B. and M. T. Stevenson (2020): "Algorithmic Social Engineering," *AEA Papers and Proceedings*, 110, 96–100.

Cowgill, B. and C. E. Tucker (2020): "Algorithmic Fairness and Economics," Working Paper.

Curello, G. and L. Sinander (2024): "The Comparative Statics of Persuasion," Working Paper.

Delgado, C., M. Baweja, N. R. Burrows, D. C. Crews, N. D. Eneanya, C. A. Gadegbeku, L. A. Inker, M. L. Mendu, W. G. Miller, M. M. Moxey-Mims, G. V. Roberts, W. L. St. Peter, C. Warfield, and N. R. Powe (2021): "Reassessing the Inclusion of Race in Diagnosing Kidney Diseases: An Interim Report from the NKF-ASN Task Force," *Journal of the American Society of Nephrology (JASN)*, 32, 1305–1317.

Dessein, W., A. Frankel, and N. Kartik (2023): "Test-Optional Admissions," Working Paper.

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 *EC '21: Proceedings of the 22nd ACM Conference on Economics and Computation*, 371–389.

Doval, L. and A. Smolin (2024): "Persuasion and Welfare," *Journal of Political Economy*, forthcoming.

Dworczak, P., S. Kominers, and M. Akbarpour (2021): "Redistribution Through Markets," *Econometrica*, 89, 1665–1698.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012): "Fairness through Awareness," in *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Dwork, C. and A. Roth (2014): "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science*, 9, 211–407.

Ellison, G. and P. A. Pathak (2021): "The Efficiency of Race-Neutral Alternatives to Race-Based Affirmative Action: Evidence from Chicago's Exam Schools," *American Economic Review*, 111, 943–75.

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

Feigenberg, B. and C. Miller (2021): "Would Eliminating Racial Disparities in Motor Vehicle Searches have Efficiency Costs?" *The Quarterly Journal of Economics*, 137, 49–113.

Ferry, J., U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala (2023): "Improving Fairness Generalization Through a Sample-Robust Optimization Method," *Machine Learning*, 112, 2131–2192.

Fisman, R., S. Kariv, and D. Markovits (2007): "Individual Preferences for Giving," *American Economic Review*, 97, 1858–1876.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2021): "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *Journal of Finance*, 77, 5–47.

Garfinkel, S. L., J. M. Abowd, and S. Powazek (2018): "Issues Encountered Deploying Differential Privacy," in *WPES '18: Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–137.

Garg, N., H. Li, and F. Monachou (2021): "Standardized Tests and Affirmative Action: The Role of Bias and Variance," in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 261.

Hansen, V. P. B., A. T. Neerkaje, R. Sawhney, L. Flek, and A. Søgaard (2024): "The Impact of Differential Privacy on Group Disparity Mitigation," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 3952–3965.

Hardt, M., E. Price, and N. Srebro (2016): "Equality of Opportunity in Supervised Learning," in *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

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.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2, Springer.

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

——— (2023): "Information and Policing," Working Paper.

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

Kamenica, E. (2019): "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 11, 249–272.

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 *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576 – 586.

Klare, B. F., M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain (2012): "Face Recognition Performance: Role of Demographic Information," *IEEE Transactions on Information Forensics and Security*, 7, 1789–1801.

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 *ITCS '17: 8th Innovations in Theoretical Computer Science Conference*.

Liang, A., J. Lu, X. Mu, and K. Okumura (2024): "Replication Code for: 'Algorithm Design: A Fairness-Accuracy Frontier'," Harvard Dataverse, https://doi.org/10.7910/DVN/4HURK4.

Lipton, Z. C., A. Chouldechova, and J. McAuley (2018): "Does Mitigating ML's Impact Disparity Require Treatment Disparity?" in *NIPS '18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 8136–8146.

Little, C. O., M. Weylandt, and G. I. Allen (2022): "To the Fairness Frontier and Beyond: Identifying, Quantifying, and Optimizing the Fairness-Accuracy Pareto Frontier," Working Paper.

Liu, Y. and F. Molinari (2024): "Inference for an Algorithmic Fairness-Accuracy Frontier," Working Paper.

Loewenstein, G. F., L. Thompson, and M. H. Bazerman (1989): "Social Utility and Decision Making in Interpersonal Contexts," *Journal of Personality and Social Psychology*, 57, 426–441.

Ludwig, J. and S. Mullainathan (2024): "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, 139, 751–827.

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.

Manski, C. F. (2022): "Patient-Centered Appraisal of Race-free Clinical Risk Assessment," *Health Economics*, 31, 2109–2114.

Manski, C. F., J. Mullahy, and A. Venkataramani (2023): "Using Measures of Race to Make Clinical Predictions: Decision Making, Patient Health, and Fairness," *PNAS*, 120, e2303370120.

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.

Menon, A. K. and R. C. Williamson (2018): "The Cost of Fairness in Binary Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency:*

PMLR 81, 107–118.

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.

Persico, N. (2002): "Racial Profiling, Fairness, and Effectiveness of Policing," *American Economic Review*, 92, 1472–1479.

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.

Rose, E. K. (2021): "Who Gets a Second Chance? Effectiveness and Equity in Supervision of Criminal Offenders," *The Quarterly Journal of Economics*, 136, 1199–1253.

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

Rothstein, J. M. (2004): "College Performance Predictions and the SAT," *Journal of Econometrics*, 121, 297–317.

Saez, E. and S. Stantcheva (2016): "Generalized Social Marginal Welfare Weights for Optimal Tax Theory," *American Economic Review*, 106, 24–45.

Sagawa, S., P. W. Koh, T. B. Hashimoto, and P. Liang (2020): "Distributionally Robust Neural Networks," in *International Conference on Learning Representations*.

Strack, B., J. Deshazo, C. Gennings, J. L. O. Ortiz, S. Ventura, K. Cios, and J. Clore (2014): "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," *BioMed Research International*, 2014, 781670.

Sullivan, C. (2023): "Eliciting Preferences over Life and Death: Experimental Evidence from Organ Transplantation," Working Paper.

Systemwide Academic Senate (2020): *Report of the UC Academic Council Standardized Testing Task Force (STTF)*, University of California, Oakland, CA.

Viviano, D. and J. Bradic (2023): "Fair Policy Targeting," *Journal of the American Statistical Association*, 119, 730–743.

Vyas, D. A., L. G. Eisenstein, and D. S. Jones (2020): "Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms," *New England Journal of Medicine*, 383, 874–882.

Wald, A. and J. Wolfowitz (1951): "Characterization of the Minimal Complete Class of Decision Functions When the Number of Distributions and Decisions Is Finite," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, Berkeley, California: University of California Press, 149–157.

Wei, S. and M. Niethammer (2022): "The Fairness-Accuracy Pareto Front," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15, 287–302.

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: A Fairness-Accuracy Frontier

Annie Liang Jay Lu Xiaosheng Mu Kyohei Okumura

June 26, 2025

O.1. **Different Loss Functions.** In this section, we generalize Theorem 1 to cover cases where fairness and accuracy are evaluated using different loss functions.

Assume the set of covariate vectors $\mathcal{X}$ is finite, and let $a : \mathcal{X} \to \Delta(D)$ describe a generic algorithm and $\mathcal{A}$ denote the set of all algorithms. As in the main text, there is a loss function $\ell : \mathcal{Y} \times \mathcal{D} \to \mathbb{R}$ such that each group g's error rate under algorithm a is $e_g = \mathbb{E}_{D\sim a(X)}\left[\ell(D, Y) \mid G = g\right]$. Different from the main text, the *unfairness* of algorithm $a \in \mathcal{A}$ is measured by $|h(a)|$ where $h : \mathcal{A} \to \mathbb{R}_+$ is any linear function. This includes as a special case

$$h(a) = \mathbb{E}_{D\sim a(X)}\left[\tilde{\ell}(D, Y)\,|G = r\right] - \mathbb{E}_{D\sim a(X)}\left[\tilde{\ell}(D, Y)\,|G = b\right] \tag{O.1}$$

where $\widetilde{\ell} : \mathcal{Y} \times \mathcal{D} \to \mathbb{R}$ is a "fairness" loss function. Our previous approach is returned when h takes the formulation in (O.1) and $\widetilde{\ell}$ is identical to ℓ.

For each pair of error rates $e \in \mathcal{E}_X$, we define

$$d(e) := \min_{e(a)=e} |h(a)|$$

to be the minimal unfairness that can be achieved using an algorithm that yields error pair e. This is well defined since $|h(\cdot)|$ is continuous and the set of algorithms $\{a : e(a) = e\}$ is compact.

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

Definition O.1. Let $>_{FA}$ be the strict order on $\mathcal{E}_X$ where $e >_{FA} e'$ if $e < e'$ and $d(e) \leq d(e')$.

Definition O.2. The FA frontier is $\mathcal{F}_X := \{e \in \mathcal{E}_X : \text{no } e' \in \mathcal{E}_X \text{ such that } e' >_{FA} e\}$.

When $h(a)$ has the formulation (O.1) and the accuracy and fairness loss functions $\tilde{\ell} = \ell$ coincide, then $d(e) = |e_r - e_b|$, and so these definitions reduce to Definitions 6, 4 and 7.

Define

$$\underline{\delta} = \min_{e\in\mathcal{E}} d(e)$$

to be the minimal level of unfairness that is achievable by a feasible algorithm. For any $\delta \geq \underline{\delta}$,

$$\mathcal{E}_\delta := \{e \in \mathcal{E}_X : d(e) \leq \delta\}$$

is the set of errors achievable by an algorithm whose unfairness is weakly less than δ.

Definition O.3. For each $\delta \geq \underline{\delta}$, define $r_X(\delta) = \arg\min_{e \in \mathcal{E}_\delta} e_r$ and $b_X(\delta) = \arg\min_{e \in \mathcal{E}_\delta} e_b$ to be the group-optimal points within each set $\mathcal{E}_\delta$, where we break ties by choosing the point that minimizes the other group's error. Further define $R_X = (r_X(\delta))_{\delta \geq \underline{\delta}}$ and $B_X = (b_X(\delta))_{\delta \geq \underline{\delta}}$ to be the set of group-optimal points as we vary over the level of unfairness.

Definition O.4. For any convex set $E \subseteq \mathbb{R} \times \mathbb{R}$, let $\mathcal{P}(E)$ denote the usual Pareto frontier of E, i.e., all points $e \in E$ where no other $e' \in E$ is weakly smaller in each entry and strictly smaller in at least one.

Definition O.5. The fairness-optimal set is $F_X := \mathcal{P}(\mathcal{E}_{\underline{\delta}})$.

We now characterize the FA frontier for this general case.

Theorem O.1. *$\mathcal{F}_X$ is the closed set bounded by R_X, B_X, $\mathcal{P}_X$ and F_X.*

The property of group-balance generalizes as follows.

Definition O.6 (Generalized Group-Balance). Say that X is *generalized group-balanced* if $F_X \subseteq \mathcal{P}_X$.

That is, X is generalized group-balanced if the fairness-optimal set belongs to the usual Pareto frontier. This reduces to the condition in the main text when h takes the form given in (O.1) and $\widetilde{\ell} = \ell$. Several of our previous results extend under this generalization of group-balance. For example, group-balance again identifies when the fairness-accuracy frontier is equivalent to the usual Pareto frontier.

Proposition O.1. *If X is generalized group-balanced, then $\mathcal{F}_X = \mathcal{P}_X$.*

The related Corollary 1 also extends.

Corollary O.1. *X fails generalized group-balance if and only if there are points $e, e' \in \mathcal{F}_X$ such that $e < e'$ but $d(e) > d(e')$.*

In some cases, generalized group-balance reduces further. One such case is when $X \in \{0,1\}$ is binary, and h follows the formulation in (O.1) where the fairness loss function $\ell(d,y) = \mathbb{1}(d=0)$ is an indicator for whether the decision is equal to 0 (e.g., not getting hired); in this case, fairness is measured as the absolute difference in the conditional probability of being assigned $d=0$ given membership in either group. Let the accuracy loss function $\ell(d,y) = \mathbb{1}(d \neq y)$ be the standard misclassification loss. For each $g \in \{r,b\}$ define

$$a_g^x := \mathbb{1}\left[\mathbb{P}(Y=1 \mid X=x, G=g) \geq 1/2\right]$$

to be the optimal action for group g given signal realization x, breaking ties in favor of $d=1$. Then generalized group-balance reduces to the following easily checkable condition.

Claim 1. *X fails generalized group-balance if and only if $a_{r0} = a_{b0}$ and $a_{r1} = a_{b1}$ and these values are distinct—that is, the optimal action is the same for both groups given either covariate realization, and this common optimal action differs across covariate realizations.*

The proof of Claim 1 and all other results mentioned in this section are contained below.

O.1.1. *Proofs of Theorem O.1 and Proposition O.1.* To save on notation we suppress dependence on X in what follows, using $\mathcal{F}$ for the fairness-accuracy frontier, $\mathcal{E}$ for the feasible set and $\mathcal{A}$ for the set of algorithms. We first show that the fairness-accuracy frontier is the union of the Pareto frontiers of the unfairness sublevel sets.

Lemma O.1. *$d(\cdot)$ is continuous and convex.*

Proof. We first show convexity. Consider $e_1, e_2 \in \mathcal{E}_X$ and let a_i be the algorithm that minimizes unfairness among all algorithms yielding error pair e_i; that is, $d(e_i) = |h(a_i)|$ and $e(a_i) = e_i$ for $i \in \{1,2\}$. Since $e(\cdot)$ and $h(\cdot)$ are linear,

$$\begin{aligned} d(\lambda e_1 + (1-\lambda) e_2) &= d(\lambda e(a_1) + (1-\lambda) e(a_2)) = d(e(\lambda a_1 + (1-\lambda) a_2)) \\ &\leq |h(\lambda a_1 + (1-\lambda) a_2)| = |\lambda h(a_1) + (1-\lambda) h(a_2)| \\ &\leq \lambda |h(a_1)| + (1-\lambda) |h(a_2)| \\ &= \lambda d(e_1) + (1-\lambda) d(e_2) \end{aligned}$$

as desired.

We now show continuity. Consider the correspondence $\varphi : \mathcal{E} \rightrightarrows \mathcal{A}$ where

$$\varphi(e) := \{a \in \mathcal{A} : e(a) = e\}$$

Note this is compact-valued. We will show φ is continuous. To show upper hemicontinuity, consider sequences $e^k \to e$ and $a^k \to a$ where each $a^k \in \varphi(e^k)$. Then by definition of φ, each $e(a^k) = e^k$, which further implies $e(a) = e$ as $e(\cdot)$ is continuous. Thus, $a \in \varphi(e)$ proving upper hemicontinuity.

We now show lower hemicontinuity. Consider $e^k \to e$ and some $a \in \varphi(e)$. For each e^k, let $a^k \in \varphi(e^k)$ be the closest point in $\varphi(e^k)$ to a. Since $\varphi(e^k)$ is a linear subspace, a^k is unique and well-defined. We will show that $|a^k - a| \to 0$. Suppose otherwise, in which case we can find some n and $\varepsilon > 0$ such that for all $k > n$, $|a' - a| \geq \varepsilon$ for all $a' \in \varphi(e^k)$. But that means $\varphi(e)$ is also strictly separated from a yielding a contradiction. This proves φ is also lower hemicontinuous and thus continuous. Since $h(\cdot)$ is continuous, by the maximum theorem, $d(\cdot)$ is continuous. □

Lemma O.2. *$\mathcal{E}_\delta$ is closed and convex.*

Proof. Immediate from the fact that $d(\cdot)$ is convex and continuous (Lemma O.1). □

Lemma O.3. $\mathcal{F} = \bigcup_{\delta \geq 0} \mathcal{P}(\mathcal{E}_\delta)$

Proof. First, suppose $e \in \mathcal{P}(\mathcal{E}_\delta)$ for some $\delta \geq 0$. Suppose $e \notin \mathcal{F}$ so there exists some $e' \in \mathcal{E}$ that FA-dominates e. Thus, $d(e') \leq d(e)$ so $e' \in \mathcal{E}_\delta$. Note that if $e'_g < e_g$ for some group g, then this contradicts $e \in \mathcal{P}(\mathcal{E}_\delta)$. Thus, it must be that $e'_g = e_g$ for both groups g so $e' = e$ yielding a contradiction.

Now, let $e \in \mathcal{F}$ and consider $\delta = d(e)$. Clearly, $e \in \mathcal{E}_\delta$. Note that if $e \notin \mathcal{P}(\mathcal{E}_\delta)$, then there exists another $e' \in \mathcal{E}_\delta$ that Pareto dominates e. But since $d(e') = d(e)$, e' also FA-dominates e yielding a contradiction. □

Completion of the proof of Theorem O.1. We will first prove $r_X(\cdot)$ is continuous. First, let

$$\mathcal{A}_\delta := \{a \in \mathcal{A} : \ |h(a)| \leq \delta\}$$

and note that

$$\mathcal{A}_\delta = \{a \in \mathcal{A} : \ -\delta \leq h(a) \leq \delta\}$$

Since $h(\cdot)$ is linear, this is just a polytope in $A = [0,1]^{\mathcal{X}}$.

Fix some δ and define

$$e_r^* := \min_{a \in \mathcal{A}_\delta} e_r(a)$$

$$e_b^* := \min_{a' \in \arg\min_{a \in \mathcal{A}_\delta} e_r(a)} e_b(a')$$

We will show that $e^* = r_X(\delta)$. First, let a^* be the corresponding algorithm for e^* so $|h(a^*)| \leq \delta$. This implies that $d(e^*) \leq \delta$ so $e^* \in \mathcal{E}_\delta$. Thus, if we let $e = r_X(\delta)$, then $e_r \leq e_r^*$. Suppose the inequality is strict. That means we can find some algorithm a^{**} such that $e_r(a^{**}) < e_r(a^*)$ and $|h(a^{**})| \leq \delta$. But that implies $a^{**} \in \mathcal{A}_\delta$ contradicting the definition of e_r^* so it must be $e_r = e_r^*$. This implies that $e_b \leq e_b^*$. Suppose the inequaility is strict, so again we can find some algorithm a^{**} such that $e_r(a^{**}) = e_r(a^*)$, $e_b(a^{**}) < e_b(a^*)$ and $|h(a^{**})| \leq \delta$. This contradicts the definition of e_b^* so it must be that $e = e^*$. We can thus write

$$r_X(\delta) = \left(\min_{a \in \mathcal{A}_\delta} e_r(a), \min_{a' \in \arg\min_{a \in \mathcal{A}_\delta} e_r(a)} e_b(a') \right)$$

Continuity follows from the fact that $e_r(\cdot)$, $e_b(\cdot)$ and $h(\cdot)$ are all linear. That $b_X(\delta)$ is continuous follows symmetrically. Since $\mathcal{F} = \bigcup_{\delta \geq 0} \mathcal{P}(\mathcal{E}_\delta)$ and $\mathcal{P}(\mathcal{E}_\delta)$ is characterized by $r_X(\delta)$ and $b_X(\delta)$, the result follows.

Completion of the proof of Proposition O.1. Recall

$$\mathcal{A}_\delta = \{a \in A : \; -\delta \leq h(a) \leq \delta\}$$

Now, for $\lambda \in (0,1)$, define

$$e_\lambda := \lambda e_r + (1-\lambda) e_b$$

and

$$\mathcal{A}_\delta^*(\lambda) := \arg\min_{a \in \mathcal{A}_\delta} e_\lambda(a)$$

Define $\mathcal{A}_\delta^*(0)$ and $\mathcal{A}_\delta^*(1)$ similarily but with tie-breaking. For large enough δ where $\mathcal{A}_\delta = \mathcal{A}$, we can just let $\mathcal{A}^* = \mathcal{A}_\delta^*$. It is straightforward to show that

$$\mathcal{P}(\mathcal{E}_\delta) = \bigcup_{\lambda \in [0,1]} \{e(a) \; : \; a \in \mathcal{A}_\delta^*(\lambda)\}$$

We will now prove that if $\mathcal{P}(\mathcal{E}_{\delta_1}) \subset \mathcal{P}(\mathcal{E})$ and $\delta_1 \leq \delta_2$, then $\mathcal{P}(\mathcal{E}_{\delta_2}) \subset \mathcal{P}(\mathcal{E})$. Consider some $e \in \mathcal{P}(\mathcal{E}_{\delta_2})$ so we can find some $\lambda \in [0,1]$ such that $a_2 \in \mathcal{A}^*_{\delta_2}(\lambda)$ and $e = e(a_2)$. Let $a_1 \in \mathcal{A}^*_{\delta_1}(\lambda)$ and $\bar{a} \in \mathcal{A}^*(\lambda)$. Since $\mathcal{A}_\delta$ is increasing in δ, it must be that

$$e_\lambda(a_1) \leq e_\lambda(a_2) \leq e_\lambda(\bar{a})$$

Now, since $e(a_1) \in \mathcal{P}(\mathcal{E}_{\delta_1}) \subset \mathcal{P}(\mathcal{E})$, there must exist some $a_1' \in \mathcal{A}^*(\lambda_1)$ for some $\lambda_1 \in [0,1]$ such that $e(a_1') = e(a_1)$. That implies that

$$e_{\lambda_1}(a_1) = e_{\lambda_1}(a_1') \leq e_{\lambda_1}(a)$$

for all $a \in \mathcal{A}$ so $a_1 \in \mathcal{A}^*(\lambda_1)$. Note that $a_1 \in \mathcal{A}^*(\lambda_1)$ and $\bar{a} \in \mathcal{A}^*(\lambda)$ are on the boundary of $\mathcal{A}$. Since a_2 is also on the boundary of $\mathcal{A}$, by continuity, we can find some λ_2 between λ_1 and λ such that $e(a_2) \in \mathcal{A}^*_{\delta_2}(\lambda_2)$. This implies $e(a_2) \in \mathcal{P}(\mathcal{E})$ as desired.

O.2. **Proof of Claim 1.** Since X is binary-valued, each algorithm can be identified with a pair (p_0, p_1) denoting the respective probabilities with which $X = 0$ and $X = 1$ are mapped into $d = 1$. From the proof of Lemma A.1, we know that the feasible set is a polygon whose vertices are the error rates derived from the deterministic algorithms $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. Moreover,

$$\begin{aligned}|\mathbb{E}(d=1 \mid G=r) - \mathbb{E}(d=1 \mid G=b)| &= |(\alpha_r p_0 + (1-\alpha_r)p_1) - (\alpha_b p_0 + (1-\alpha_b)p_1| \\ &= |(\alpha_r - \alpha_b)(p_0 - p_1))|\end{aligned}$$

where $\alpha_g := \mathbb{P}(X = 0 \mid G = g)$. So unfairness is minimized (and achieves the value zero) by setting $p_0 = p_1$. Thus F_X is the Pareto set of the line from the $(1,1)$ vertex to the $(0,0)$ vertex of the polygon.

Suppose $a_{r0} = a_{b0}$ and $a_{r1} = a_{b1}$ with distinct values. Then the deterministic algorithm $(p_0, p_1) = (a_{r0}, a_{r1}) \in \{(0,1), (1,0)\}$ maximizes accuracy for both groups. The corresponding vertex is simultaneously r_X and b_X, so it is also the Pareto set $\mathcal{P}(\mathcal{E})$. But this point does not intersect the line from $(0,0)$ to $(1,1)$, so F_X does not belong to $\mathcal{P}(\mathcal{E})$ (see Figure 13 for an example). We thus have the claim in one direction.

In the other direction, suppose first that $a_{r0} = a_{b0} = a_{r1} = a_{b0}$. In this case, $(p_0, p_1) = (a_{r0}, a_{r1}) \in \{(0,0), (1,1)\}$ is simultaneously r_X, b_X, and F_X, as depicted in Panel (a) of Figure 14. Clearly $F_X \in \mathcal{P}(\mathcal{E})$.

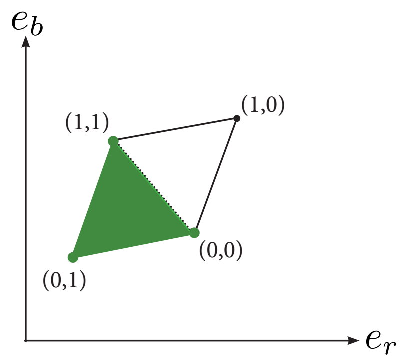


FIGURE 13. X fails generalized group balance. The fairness-accuracy set is the shaded green area. In this example, $r_X = b_X = (0,1)$ while F_X is the line from $(0,0)$ to $(1,1)$.

In all remaining cases, r_X is different from b_X, so the Pareto set $\mathcal{P}(\mathcal{E}_X)$ includes at least one non-degenerate line segment. This line segment must include at least one of the vertices $(0,0)$ and $(1,1)$. See Panels (b)-(d) of Figure 14 for the possible configurations. So the Pareto set intersects the line connecting $(1,1)$ and $(0,0)$, and F_X is precisely this point of intersection. Thus $F_X \in \mathcal{P}(\mathcal{E})$, completing the argument.

O.3. **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.[60]

Proposition O.2. *Let $\mathcal{F}'_X$ 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) $\mathcal{F}_X = \mathcal{F}'_X$ if X is group-balanced
(2) $\mathcal{F}_X \subseteq \mathcal{F}'_X$ if X is group-skewed and ϕ in concave
(3) $\mathcal{F}_X \supseteq \mathcal{F}'_X$ if X is group-skewed and ϕ in convex

[60]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.

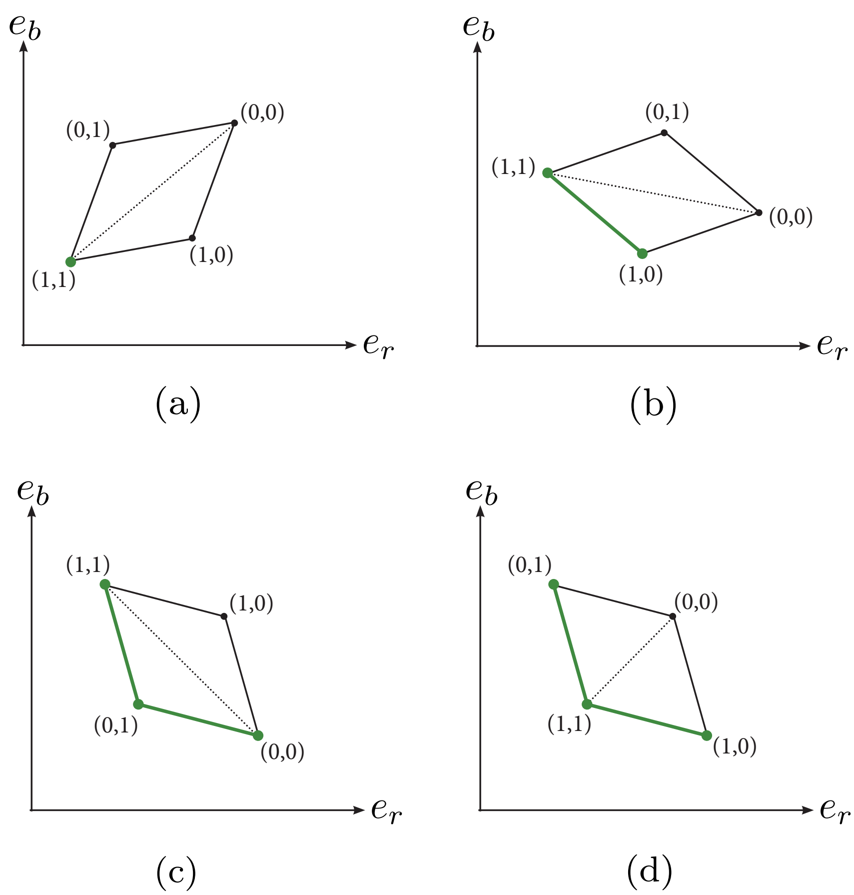


FIGURE 14. X satisfies generalized group balance. The fairness-accuracy set is the shaded green area. In all cases, $F_x = \{f_x\}$ is a singleton. In Panel (a), $r_X = b_X = f_X = (1,1)$. In Panels (b) and (c), $r_X = f_X = (1,1)$ while $b_X = (0,0)$. In Panel (d), $r_X = (0,1)$, $f_X = (1,1)$, and $b_X = (1,0)$.

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, $\mathcal{F}_X = \mathcal{F}'_X$ 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

$$e_b - e_r \leq e'_b - e'_r$$

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

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, $\mathcal{F}_X \subseteq \mathcal{F}_X'$. The case for when ϕ is convex is symmetric. □

O.4. **Adversarial Agents.** We now consider the problem outlined in Section 4, when one of the weights α_r, α_b is negative.[61] 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.

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

Lemma O.4. *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

$$g_X^* := \arg\min_{e \in \mathcal{E}_X^*} e_g$$

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

$$f_X^* := \arg\min_{e \in \mathcal{E}_X^*} |e_r - e_b|$$

to be the point that minimizes the absolute difference between group errors (breaking ties by minimizing either group's error).

Definition O.7. Covariate vector X is:

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

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

Theorem O.2. *The input-design fairness-accuracy (FA) frontier $\mathcal{F}_X^*$ 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 4.3.2.

Definition O.8. 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.3. *Suppose $\alpha_r > 0 > \alpha_b$ and X is strictly input-design group-balanced. Then excluding G uniformly worsens the frontier, in the sense that $\mathcal{F}_{X,G}^* >_{FA} \mathcal{F}_X^*$.*

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.5. **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.

O.5.1. *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

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

returns a relaxed version of this criterion, since

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

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 $d = 1$.

O.5.2. *False Positives.* Another common fairness criterion is equality of false positives across two groups (Angwin and Larson, 2016; 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

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

To see how we can accommodate this, consider the group-dependent loss function

$$\ell_g\left(d, y\right) = \frac{\mathbb{1}\left\{d = 1, y = 0\right\}}{\mathbb{P}\left(Y = 0 \mid G = g\right)},$$

In this case, we obtain

$$\begin{aligned} e_g\left(a\right) &= \frac{\mathbb{P}\left(a\left(X\right) = 1, Y = 0 \mid G = g\right)}{\mathbb{P}\left(Y = 0 \mid G = g\right)} \\ &= \frac{\mathbb{P}\left(a\left(X\right) = 1, Y = 0, G = g\right)}{\mathbb{P}\left(Y = 0, G = g\right)} \\ &= \mathbb{P}\left(a\left(X\right) = 1 \mid Y = 0, G = g\right) \end{aligned}$$

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.

An alternative (unconditional) false positive rate would be equalizing

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

This is achieved by setting the loss function

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

since in this case

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

O.5.3. *Equalized Odds.* Another popular fairness criterion asks for equalized odds (Hardt et al., 2016), which an algorithm a satisfies if

$$\mathbb{E}_Y[\mathbb{E}_X[a(X) \mid G = r, Y] - \mathbb{E}_X[a(X) \mid G = b, Y]] = 0 \tag{O.2}$$

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 group-dependent loss function

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

returns a relaxed version of this criterion, since

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

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 (O.2).

O.6. **Supplementary Material to Section 5.** As a robustness check, we compute point estimates of group-optimal points using three additional algorithms: a neural net with two hidden layers (NN-2), a neural net with three hidden layers (NN-3), and a support vector machine with a radial basis function kernel (SVM).[62] These three algorithms, like the random forest algorithm, are widely used to search within the class of all possible (non-linear) classifiers. Table 2 and Table 3 below summarize the point estimates generated by these algorithms. We find that the random forest algorithm performs similarly well to these algorithms, and all three yield the same result regarding group-balance/group-skew.

[62]For NN-2 and NN-3, each hidden layer has 32 nodes.

Table 2. Comparison of estimates for the Obermeyer et al. (2019) data

Algorithms	$\hat{e}_b^b$	$\hat{e}_w^b$	$\hat{e}_b^w$	$\hat{e}_w^w$
RF	0.041	0.018	0.044	0.017
NN-2	0.052	0.021	0.046	0.021
NN-3	0.054	0.021	0.048	0.021
SVM	0.057	0.022	0.062	0.024

Table 3. Comparison of estimates for the Strack et al. (2014) data

Algorithms	$\hat{e}_w^w$	$\hat{e}_m^w$	$\hat{e}_w^m$	$\hat{e}_m^m$
RF	0.393	0.394	0.393	0.393
NN-2	0.381	0.379	0.381	0.380
NN-3	0.384	0.384	0.382	0.381
SVM	0.380	0.377	0.381	0.377