Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

Online appendix to the paper

Algorithmic Design: A Fairness-Accuracy Frontier

Annie Liang Jay Lu Xiaosheng Mu Kyohei Okumura

December 16, 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 $e_g = \mathbb{E}_{D\sim a(X)}[\ell(D,Y) \mid G = g]$. 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

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

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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 $d(e) > d(e')$.*

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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.

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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 $a^k \in \varphi\left(e^k\right)$. Then by definition of φ, each $e\left(a^k\right) = e^k$, 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 $a^k \in \varphi\left(e^k\right)$ be the closest point in $\varphi\left(e^k\right)$ to a. Since $\varphi\left(e^k\right)$ is a linear subspace, a^k is unique and well-defined. We will show that $\left|a^k - a\right| \to 0$. 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 $a' \in \varphi\left(e^k\right)$. 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

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

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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

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

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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 $F_X \in \mathcal{P}(\mathcal{E})$.

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

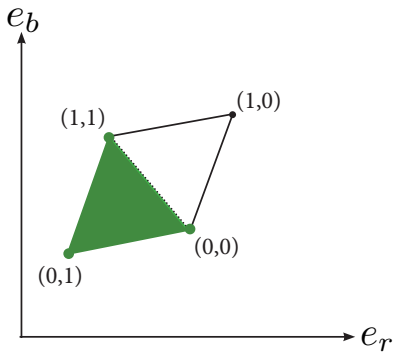

FIGURE 13. X fails generalized group balance. The fairness-accuracy set is the shaded green area. In this example, $r_X = b_X = (0,1)$ while F_X is the line from $(0,0)$ to $(1,1)$.

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.

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

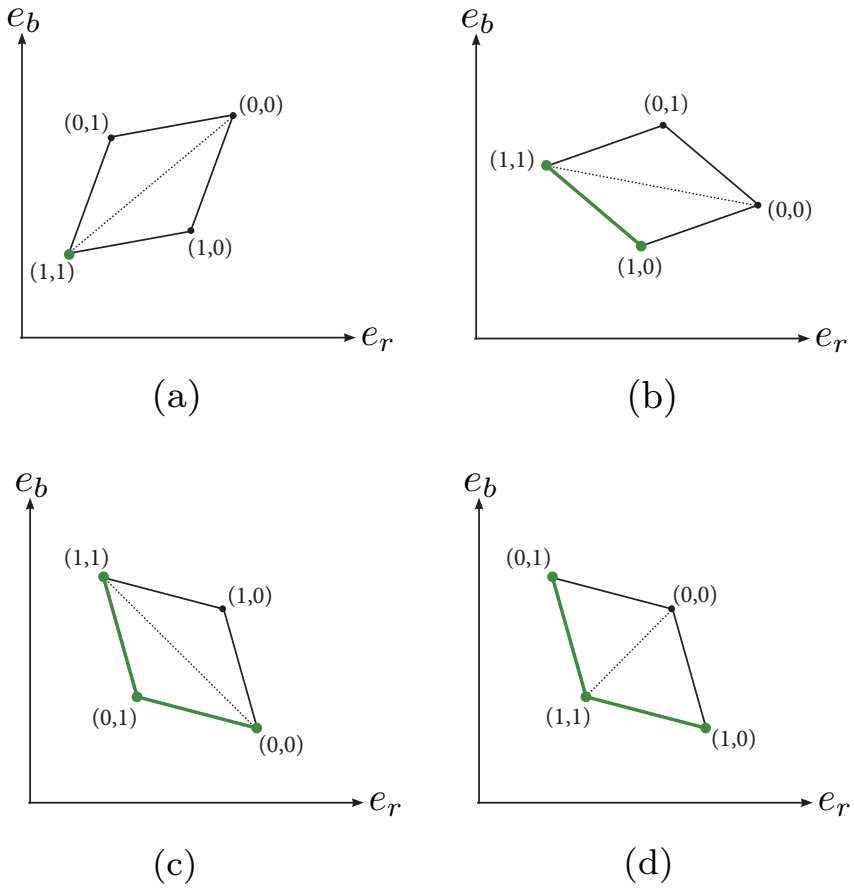

FIGURE 14. X satisfies generalized group balance. The fairness-accuracy set is the shaded green area. In all cases, $F_x = \{f_x\}$ is a singleton. In Panel (a), $r_X = b_X = f_X = (1,1)$. In Panels (b) and (c), $r_X = f_X = (1,1)$ while $b_X = (0,0)$. In Panel (d), $r_X = (0,1)$, $f_X = (1,1)$, and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

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

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

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

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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.

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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

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

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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

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

In this case, we obtain

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

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

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

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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

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

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.

Supplemental Material (not copyedited or formatted) for: Annie Liang, Jay Lu, Xiaosheng Mu, Kyohei Okumura<aff id="af3" />. 2026. "Algorithm Design: A Fairness-Accuracy Frontier."
Journal of Political Economy 134(5). DOI: https://doi.org/10.1086/739826.

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

Algorithms	$\hat{e}^b_b$	$\hat{e}^b_w$	$\hat{e}^w_b$	$\hat{e}^w_w$
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	$\hat{e}^w_w$	$\hat{e}^w_m$	$\hat{e}^m_w$	$\hat{e}^m_m$
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