The Price of Precision: A Model of Optimal Bias on the Fairness-Accuracy Frontier

Kofi Hair-Ralston* Daksh Mathapati
kofibhairralston@gmail.com daksh.mathapati@gmail.com

August 2025

Abstract

Why do profit-maximizing firms persist in using biased algorithms despite high reputational costs and possessing the technical means for debiasing? We propose an "informational" theory of discrimination, distinct from canonical taste-based [Becker, 1957] or statistical [Phelps, 1972, Arrow, 1973] motives. We model a firm facing a technological frontier where enforcing algorithmic fairness reduces predictive accuracy. In this environment, we show that a firm optimally chooses a strictly positive level of bias ($b^* > 0$) even under the cleanest possible conditions: when protected groups are ex-ante identical and debiasing is costless. This choice is a rational response to the information structure; by tolerating disparate impact, the firm preserves a more informative signal, a logic that follows the Informativeness Principle [Holmstrom, 1979]. Our contribution is to formalize this third channel of discrimination, driven not by preferences or priors, but by the design of the predictive technology itself. The model demonstrates that the private choice of bias is socially inefficient, creating a deadweight loss and suggesting that policies should focus on improving the technological frontier rather than simply mandating fairness.

Keywords: Algorithmic Fairness, Information Economics, Discrimination, Bayesian Persuasion, Informativeness Principle, Signal Garbling, Economics of AI.

JEL Codes: D82, J71, D83

*We thank Peter Hull, Alex Imas, and Aislinn Bohren for their invaluable feedback. All remaining errors are our own.

1 Introduction

In 2018, Amazon scrapped an artificial intelligence recruiting tool after discovering a major flaw: it disliked women [Dastin, 2018]. The system was designed to automate the search for top software developers by evaluating parsed resumes. Yet, the machine learning models taught themselves that male candidates were preferable, penalizing resumes that contained the word "women's" (as in "women's chess club captain") and downgrading graduates from two all-women's colleges. Amazon's engineers tried to amend the system to make it neutral, but they could not guarantee that the machine would not devise new, equally discriminatory ways of sorting candidates. Amazon ultimately abandoned the project.

This instance is representative of a persisting puzzle in the modern economy. Amazon is a sophisticated, profit-maximizing firm. It faced no shortage of technical expertise or computational resources, and its goal was to find the best possible talent, not to discriminate. Yet, its attempt to build a purely meritocratic algorithm resulted in a biased system that was ultimately irreparable. For years, similar issues of persistent, unintended bias have emerged in algorithms for credit scoring, medical diagnoses, and criminal justice risk assessment. These cases are difficult to explain with the canonical economic models of discrimination. Taste-based theories [Becker, 1957] struggle because the principal is an algorithm, not a manager with personal animus, and competitive pressures should, in theory, punish such inefficiency. Statistical discrimination theories [Phelps, 1972, Arrow, 1973] are also an imperfect fit, as they assume that group identity is used as a proxy due to a lack of individual-level data; today's algorithms, however, operate in a data-rich environment with access to hundreds of individual-specific features.

This paper proposes and formalizes a third mechanism for discrimination that is not about tastes or beliefs, but about the inherent technological trade-offs in prediction. We term this mechanism "informational discrimination." Our central premise is that there exists a technological frontier between fairness and accuracy [Kleinberg et al., 2017, Chouldechova, 2017]. Efforts to reduce an algorithm's disparate impact (debiasing) often come at the cost of reducing its predictive power. We model a firm that understands this trade-off and rationally chooses an optimal level of bias to maximize the productivity of its hired workers. The severity of this trade-off is governed by a single parameter, κ, which represents the steepness of the fairness-accuracy frontier.

Our main theoretical result shows that a profit-maximizing firm will rationally choose a strictly positive level of bias ($b^* > 0$) even under the cleanest possible conditions: when protected groups are ex-ante identical in average productivity and debiasing is technologically costless. The economic logic for this result follows from Holmstrom's Informativeness Principle [Holmstrom, 1979] . The firm faces a choice between a perfectly "fair" signal ($b = 0$), which treats all groups identically but is noisy, and a "biased" signal ($b > 0$), which systematically disadvantages one group but provides greater overall precision. Even though the biased signal creates discriminatory outcomes, its superior informational content makes it privately optimal for the firm. The marginal gain in predictive accuracy from accepting a small amount of bias outweighs the marginal cost of the signal's distortion.

The model yields a sharp comparative static: the firm's optimal level of bias, b^*, is increasing in the severity of the underlying technological trade-off, κ. This provides a set of testable predictions about where and why we should observe algorithmic bias. Industries with more complex prediction tasks or less balanced data (higher κ) should exhibit more

bias, independent of discriminatory intent. We then analyze the welfare implications of this mechanism. The firm's privately optimal choice, b^*, is socially inefficient, creating a deadweight loss by systematically disadvantaging one group. We show that the socially optimal level of bias, b^{**}, is strictly less than what the firm chooses. This divergence provides a clear rationale for policy intervention. Our framework suggests that simple mandates (e.g., forcing $b = 0$) are blunt and inefficient instruments because they ignore the underlying technological constraint. More nuanced policies, such as Pigouvian taxes that force the firm to internalize the social cost of bias, or R&D subsidies that fund the development of better technology to "flatten" the frontier (lower κ), are more effective at aligning private incentives with social goals.

This paper contributes to three strands of literature. First, we contribute to the foundational literature on economic discrimination by formalizing a novel, information-based channel that is distinct from the canonical works of Becker, Phelps, and Arrow. Second, we contribute to the emerging field of the "economics of AI" by providing a tractable micro-foundation for the fairness-accuracy trade-off and exploring the strategic incentives it creates for firms. Third, our work connects to the literature on information design [Kamenica and Gentzkow, 2011, Bergemann and Morris, 2019] by modeling a principal who designs an information structure for her own use, constrained by a technological frontier that links the signal's informativeness to its disparate impact. By isolating this informational mechanism, our model provides a new theoretical lens for understanding a critical policy challenge and offers a framework for designing more effective interventions in an increasingly automated world.

2 Literature Review

The fairness-accuracy trade-off is a central theme in the algorithmic fairness literature. Our work builds on several key pillars of this research.

The foundational challenge is rooted in the mathematical impossibility of simultaneously satisfying multiple, seemingly intuitive fairness criteria. The seminal work of Kleinberg et al. [2017] provides a formal impossibility theorem, demonstrating that for any non-trivial classifier, it is impossible to satisfy both calibration and equalized odds across groups with different base rates. This result was empirically corroborated by Chouldechova [2017] in her analysis of the COMPAS recidivism prediction algorithm, which showed that the algorithm could not be simultaneously calibrated and have equal false positive rates for different racial groups. These impossibility results establish that trade-offs are not just an empirical observation but a mathematical necessity.

Our paper bridges the computer science and economics perspectives on this problem. As Gans [2025] notes, the economic approach emphasizes welfare and incentives, while the computer science literature has traditionally focused on statistical fairness metrics. We follow the economic approach by modeling an explicit trade-off, but use the statistical findings from computer science to provide micro-foundations for it. Work by Rambachan et al. [2020] provides a micro-foundation for the economic approach by decomposing prediction differences, which aligns with our formalization of the trade-off parameter κ.

Finally, our conceptualization of the trade-off's magnitude is heavily influenced by the "fairness frontier" framework of Liang et al. [2025]. They illustrate how the set of achievable error rates across groups forms a frontier, and the shape of this frontier determines the severity of the trade-off. This provides a direct link between the statistical

properties of the data and the economic cost of fairness constraints, which is central to our model.

This work connects with a foundational literature in computer science demonstrating the mathematical difficulty of satisfying multiple fairness criteria simultaneously. This provides a rigorous, axiomatic foundation for why a trade-off between fairness and accuracy often exists. This section reviews the key findings that motivate our model's core trade-off.

2.1 Foundational Theory

The cornerstone of the fairness-accuracy trade-off is a set of theoretical results demonstrating that it is mathematically impossible to satisfy multiple, intuitive fairness criteria simultaneously, except in trivial cases. This proves that a trade-off is not just an empirical observation but a fundamental mathematical constraint.

2.1.1 Kleinberg, Mullainathan, and Raghavan (2016)

A seminal paper by Kleinberg et al. [2017] provides an "impossibility theorem" of algorithmic fairness. It proves that for a scoring-based classifier, it is impossible to satisfy three key fairness conditions at the same time, unless the predictor is perfect or base rates are equal across groups. Let G be the protected attribute (group), Y be the true outcome (e.g., $Y = 1$ if a candidate succeeds), and S be the algorithm's risk score. The three conflicting conditions are: First, **Calibration** requires that the score is an accurate reflection of risk, such that $Pr(Y = 1|S = s) = s$ for all score levels s. Second, **Balance for the Positive Class** requires that the average score for those who are actually positive is the same across all groups: $E[S|Y = 1, G = g_1] = E[S|Y = 1, G = g_2]$. Finally, **Balance for the Negative Class** requires that the average score for those who are actually negative is also the same across groups: $E[S|Y = 0, G = g_1] = E[S|Y = 0, G = g_2]$. The paper proves that these three conditions can only hold simultaneously if base rates of the positive outcome are equal across groups ($Pr(Y = 1|G = g_1) = Pr(Y = 1|G = g_2)$) or if the prediction is perfect. Since base rates often differ in the real world, a decision-maker must choose which fairness metric to violate. This provides a rigorous foundation for why a trade-off exists.

2.1.2 Chouldechova (2017)

Independently and concurrently, Chouldechova [2017] demonstrated a similar impossibility result in the context of the COMPAS recidivism algorithm, showing a direct conflict between predictive parity and error rate equality.

A recent and influential strand of literature from economics reframes the problem, arguing that one should use the most accurate algorithm possible and apply fairness considerations at the decision-making stage.

2.1.3 Gans (2025)

Gans [2025] synthesizes the two competing approaches: the "computer science" approach of directly regulating algorithms versus the "economic" approach of regulating how algorithmic outputs are used. The economic approach, articulated by Rambachan et al. [2020], aims to maximize a social welfare function by choosing an allocation rule $a(g, x)$

using the output of the most accurate possible prediction $F(g, x)$. Fairness is introduced via group-specific welfare weights ϕ_g and decision thresholds $t*(g)$. This framework aligns with our paper's setup, where the firm chooses the optimal point on a "fairness-accuracy frontier."

2.1.4 Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

This paper provides the micro-foundation for the economic approach. It decomposes the difference in average predictions between groups into three components: (1) true base rate differences, (2) differences in measurement error, and (3) differences in estimation error. The computer science approach of forcing equal average predictions can conflate these, while the economic approach argues for minimizing measurement and estimation error while handling true base rate differences through explicit, group-specific decision thresholds. Our model provides a direct formalization of the trade-off arising from the second component of their decomposition: differences in measurement error. We explicitly model a scenario where a firm must decide how to handle group-specific noise (our η_g term), and show how the optimal response can lead to a biased but more precise signal. This links their decomposition directly to a profit-maximizing incentive for disparate outcomes, even when true productivity distributions are identical.

2.2 Visualizing the Trade-Off

The "fairness frontier" provides a powerful geometric framework for analyzing the trade-off.

2.2.1 Liang, Lu, Mu, and Okumura (2025)

Liang et al. [2025] introduce the concept of a "fairness frontier" in the space of group-specific error rates (e_0, e_1). The frontier is the set of non-dominated algorithms. They show that the shape of this frontier depends on the statistical properties of the data. If the data is "group-balanced," the fairest algorithm (where $e_0 = e_1$) can be efficient (on the frontier). If the data is "group-skewed" (more predictive for one group), the fairest point is inefficient (inside the frontier), creating a necessary trade-off. This framework provides a direct micro-foundation for our parameter κ. A high κ corresponds to a steep frontier (group-skewed data), where the marginal gain in precision from accepting some bias is large.

3 Conditions for Fairness-Accuracy Trade-offs

We begin by formalizing the fairness-accuracy trade-off as a constrained optimization problem. This approach is standard in the algorithmic fairness literature and allows us to precisely define the conditions under which a trade-off must exist. As established by foundational work in the field (e.g., Kleinberg et al., 2017, Chouldechova, 2017), inherent mathematical constraints often make it impossible to simultaneously satisfy multiple desirable properties, necessitating a trade-off.

3.1 The Unconstrained Learning Problem

Let us consider a standard supervised learning setting. We have a feature space $\mathcal{X}$, a protected group attribute $G \in \{g_0, g_1\}$, and a binary outcome variable $Y \in \{0, 1\}$. A firm or decision-maker wishes to learn a predictive model, or hypothesis, h from a hypothesis space $\mathcal{H}$. The goal of the standard, unconstrained learning problem is to find the hypothesis $h_{acc} \in \mathcal{H}$ that minimizes some expected loss function, $L(h(X), Y)$. This is often referred to as Empirical Risk Minimization (ERM).

$$h_{acc} = \arg\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] \tag{1}$$

This hypothesis, h_{acc}, represents the most accurate possible model within the given hypothesis space, without any consideration for fairness. The loss it achieves, $L_{acc} = \mathbb{E}[L(h_{acc}(X), Y)]$, serves as the benchmark for maximum accuracy.

3.2 Fairness as a Constraint on the Hypothesis Space

We introduce fairness by defining a set of constraints on the behavior of the hypothesis. A fairness metric, such as demographic parity, equalized odds, or calibration, imposes a restriction on the set of acceptable models. We can formalize this by defining a *fair hypothesis space*, $\mathcal{H}_F$, as the subset of hypotheses in $\mathcal{H}$ that satisfy the chosen fairness criterion.

Definition 1 (Fair Hypothesis Space)**.** A hypothesis h is considered fair if it satisfies a given fairness constraint, $C(h) \leq \epsilon$ for some tolerance $\epsilon \geq 0$. The set of all such fair hypotheses is:

$$\mathcal{H}_F = \{h \in \mathcal{H} \mid C(h) \leq \epsilon\} \tag{2}$$

The fairness-constrained learning problem is then to find the best possible hypothesis, h_{fair}, within this restricted set:

$$h_{\text{fair}} = \arg\min_{h \in \mathcal{H}_F} \mathbb{E}[L(h(X), Y)] \tag{3}$$

3.3 The Existence of a Trade-Off

A fairness-accuracy trade-off exists if and only if the solution to the constrained problem is strictly worse than the solution to the unconstrained problem. We can define the magnitude of this trade-off, which corresponds to the parameter κ in our main model, as follows:

Definition 2 (Fairness-Accuracy Trade-off)**.** The fairness-accuracy trade-off, κ_{tradeoff}, is the difference in loss between the fairness-constrained optimal hypothesis and the accuracy-optimal hypothesis:

$$\kappa_{\text{tradeoff}} = L(h_{\text{fair}}) - L(h_{acc}) \tag{4}$$

A trade-off exists (i.e., $\kappa_{\text{tradeoff}} > 0$) if and only if the most accurate hypothesis, h_{acc}, is not itself fair. That is, if $h_{acc} \notin \mathcal{H}_F$. The subsequent propositions in this section will establish the precise conditions under which this exclusion occurs.

3.4 Conditions for an Inherent Trade-off

The fundamental reason for a trade-off arises when the protected attribute G is statistically informative about the outcome Y, even after accounting for the other features X. To formalize this, let's consider the theoretically most accurate possible classifier, the Bayes Optimal Classifier.

Definition 3 (Bayes Optimal Classifier)**.** The Bayes Optimal Classifier, $h_{bayes}(x, g)$, predicts the outcome that is most probable given the features and group membership. For a binary outcome, it is defined as:

$$h_{bayes}(x, g) = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X = x, G = g) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

This classifier achieves the lowest possible error rate, known as the Bayes error rate.

Now, let's consider a common fairness constraint: demographic parity. Demographic parity requires that the probability of a positive prediction is the same for all groups.

Definition 4 (Demographic Parity)**.** A hypothesis h satisfies demographic parity if its predictions are statistically independent of the group attribute G. That is:

$$\mathbb{P}(h(X, G) = 1|G = g_0) = \mathbb{P}(h(X, G) = 1|G = g_1) \tag{6}$$

We can now state the condition for a necessary trade-off.

Proposition 1. *If the true conditional probability of the outcome differs across groups (i.e., $\mathbb{P}(Y = 1|G = g_0) \neq \mathbb{P}(Y = 1|G = g_1)$), and the Bayes Optimal Classifier is the only hypothesis that achieves the minimum Bayes error, then any classifier that satisfies demographic parity must have an error rate strictly greater than the Bayes error rate.*

Proof Sketch. Let h_{acc} be the accuracy-optimal hypothesis, which we assume is the Bayes Optimal Classifier, h_{bayes}. By definition, h_{bayes} uses group membership g in its calculation whenever $\mathbb{P}(Y = 1|X = x, G = g)$ is dependent on g.

If the base rates differ between groups (i.e., $\mathbb{P}(Y = 1|G = g_0) \neq \mathbb{P}(Y = 1|G = g_1)$), then for h_{bayes} to satisfy demographic parity, it would require the unlikely coincidence that integrating over all features X exactly equalizes the prediction rates.

More formally, the prediction rate for group g under h_{bayes} is $\int_x \mathbb{P}(h_{bayes}(x, g) = 1|X = x, G = g)p(x|g)dx$. Demographic parity would require this quantity to be equal for g_0 and g_1. However, the Bayes classifier is optimal precisely because it tracks the true conditional probabilities $\mathbb{P}(Y = 1|X, G)$. If these probabilities have different distributions across groups, the resulting prediction rates from an optimal classifier will also differ.

Therefore, if base rates are unequal, the Bayes Optimal Classifier h_{bayes} will not satisfy demographic parity. Any other hypothesis h_{fair} that does satisfy demographic parity must, by definition, differ from h_{bayes} and will therefore have a strictly higher error rate. This means $h_{acc} \notin \mathcal{H}_F$, which proves that the trade-off κ_{tradeoff} is strictly greater than zero. □

3.5 Trade-offs with Error Rate Equality

The trade-off is not unique to demographic parity. A similar conflict arises with another common fairness criterion, Equalized Odds, which focuses on error rates.

Definition 5 (Equalized Odds)**.** A hypothesis h satisfies Equalized Odds if its True Positive Rate (TPR) and False Positive Rate (FPR) are equal across all groups. That is:

$$\mathbb{P}(h(X,G)=1|Y=0,G=g_0)=\mathbb{P}(h(X,G)=1|Y=0,G=g_1) \quad \text{(Equal FPR)} \tag{7}$$

As shown by Kleinberg et al. [2017] and Chouldechova [2017], this criterion often conflicts with another desirable property of a predictor: calibration.

Definition 6 (Calibration)**.** A hypothesis h is calibrated if its prediction can be interpreted as a true probability. That is, for any prediction score s that h outputs:

$$\mathbb{P}(Y=1|h(X,G)=s)=s \tag{8}$$

The Bayes Optimal Classifier is, by its nature, perfectly calibrated. The following proposition, summarizing a key result from the literature, shows that this property is incompatible with Equalized Odds under most real-world conditions.

Proposition 2. *If the base rates of the positive outcome differ across groups (i.e.,* $\mathbb{P}(Y=1|G=g_0)\neq\mathbb{P}(Y=1|G=g_1)$*), a non-trivial classifier cannot simultaneously satisfy both Calibration and Equalized Odds.*

Proof Sketch. The proof, formally established in Kleinberg et al. [2017], shows that if a classifier satisfies both calibration and equalized odds, the base rates for each group must be equal. Since the Bayes Optimal Classifier (h_{acc}) is perfectly calibrated, if the base rates in the data are unequal, it cannot satisfy Equalized Odds. Therefore, any classifier h_{fair} that is constrained to satisfy Equalized Odds must deviate from the Bayes Optimal Classifier and will thus have a higher error rate. This again establishes that $h_{acc}\notin\mathcal{H}_F$ (where $\mathcal{H}_F$ is the set of models satisfying Equalized Odds), proving that $\kappa_{\text{tradeoff}}>0$. □

3.6 The Magnitude of the Trade-off: The Fairness Frontier

The existence of a trade-off is a binary question, but its magnitude is a continuous one. The severity of the trade-off—how much accuracy must be sacrificed for a given gain in fairness—is not a universal constant. It depends critically on the statistical properties of the data. The concept of the "fairness frontier," introduced by Liang et al. [2025], provides a clear framework for understanding this dependency.

Imagine a space where the axes represent the error rates for each group, e_{g_0} and e_{g_1}. An ideal algorithm would have zero error for both groups (the origin). Any given algorithm corresponds to a point in this space.

Definition 7 (The Fairness Frontier)**.** The fairness frontier is the set of non-dominated algorithms. An algorithm is on the frontier if no other algorithm exists that is better for one group without being worse for the other. The point of perfect fairness is where $e_{g_0}=e_{g_1}$.

The shape of this frontier, and its relationship to the point of perfect fairness, is determined by the data's structure:

Group-Balanced Data. If the features X are similarly predictive for all groups, the frontier is relatively symmetric. In this case, the fairest algorithm (where error rates are equal) may also be an efficient one (lying on the frontier). Here, the trade-off is minimal

or non-existent. A small move away from perfect fairness yields little to no gain in overall accuracy.

Group-Skewed Data If the features are much more predictive for one group than another, the frontier will be skewed. The point of perfect fairness will lie strictly inside the frontier, meaning it is inefficient. To improve the error rate for one group, one must accept a much larger increase in the error rate for the other. This creates a steep and necessary trade-off.

This provides a direct micro-foundation for the parameter κ in our main model. A large κ corresponds to a group-skewed data environment with a steep frontier, where the marginal gain in accuracy from accepting some unfairness is large. A small κ corresponds to a group-balanced environment where this marginal gain is small. Therefore, the magnitude of the fairness-accuracy trade-off is a direct consequence of the statistical properties of the data available to the decision-maker.

4 The Model

4.1 Primitives and Assumptions

A risk-neutral firm hires candidates. Candidate productivity θ draws from a Normal distribution, $\theta \sim N(\mu, \sigma_\theta^2)$. Candidates belong to group $g \in \{0, 1\}$.

Assumption (A1)**. Identical True Productivity, Differential Measurement.** To isolate the informational mechanism, our baseline model assumes that the distribution of true underlying productivity, θ, is identical across groups: $\mathbb{E}[\theta|g = 1] = \mathbb{E}[\theta|g = 0] = \mu$. However, the firm does not observe θ directly. Instead, it observes a signal s that measures productivity with group-specific error. This creates a scenario where group membership is informative not about true ability, but about the nature of the measurement error. This case is policy-relevant in contexts where, for historical or structural reasons, data for one group is less reliable or more noisy than for another. In Appendix C, we show that our core result ($b^* > 0$) is robust to the introduction of moderate differences in group means.

Assumption (A2)**. Observable Group Membership.** The firm observes group membership g. While this may cause legal troubles in practice, it helps to isolate the informational mechanism. Our results extend to cases where group membership is imperfectly inferred from observable proxies.

Assumption (A3)**. Signal and Measurement Error Structure.** The firm observes a signal s_g for a candidate from group g. The signal is a function of true productivity θ, a group-specific measurement error η_g, and the firm's choice of bias b. We model the signal as $s_g = \theta + \eta_g + b \cdot g + \varepsilon(b)$. The term $\eta_g \sim N(0, \sigma_{\eta,g}^2)$ represents exogenous, group-specific noise. The firm's choice of b can be interpreted as an adjustment to counteract suspected measurement error, which in turn affects the variance of the overall signal noise, $\varepsilon(b)$.

Table 1: Model Parameters and Notation

Parameter	Description
θ	True candidate productivity, $\theta \sim N(\mu, \sigma_\theta^2)$
s_g	Algorithmic signal of productivity for group g
$g \in \{0, 1\}$	Candidate group membership (observed)
π_g	Proportion of group g in the population, with $\pi_0 + \pi_1 = 1$
b	Firm's choice of bias level, $b \in [0, b_{max}]$
t	Hiring threshold chosen by the firm
$\sigma_\varepsilon^2(b)$	Variance of the algorithm's signal noise, a function of bias
η_g	Exogenous group-specific measurement error, $\eta_g \sim N(0, \sigma_{\eta,g}^2)$
σ_0^2	Baseline signal noise when $b = b_{max}$
κ	Technological coupling parameter ($\kappa > 0$)
$V(b)$	The firm's value function, maximized at b^*
$SWF(b)$	Social welfare function
$E(b)$	External costs of bias

4.2 The Precision-Bias Trade-off

The model's central mechanism is a trade-off between signal precision and bias. This trade-off is at its core a consequence of how fairness constraints are implemented in machine learning systems. As established in the statistical learning literature, imposing a fairness constraint acts as a regularizer, often increasing estimator variance [Kamishima et al., 2012, Wick et al., 2019]. This causes a dilemma: algorithms can be made fairer, but only at the cost of reduced accuracy.

This intuition is straightforward for economists familiar with constrained optimization. Fairness interventions in machine learning generally work through one of two channels. First, they can constrain the model to ignore certain correlations or features that are predictive but correlated with group membership, thus reducing the model's overall information set and subsequent precision. Second, they may involve post-processing outputs to equalize outcomes across groups, which is equivalent to intentionally adding noise or "garbling" the signal in the spirit of Bayesian persuasion. In either case, enforcing fairness places a quantitative constraint on the optimization problem, moving the solution away from the unconstrained, accuracy-maximizing estimator and increasing resulting prediction error.

This relationship has been documented empirically across a range of machine learning applications. Studies in hiring algorithms, credit scoring, criminal justice risk assessment, and medical diagnosis consistently show that fairness constraints reduce predictive performance [Kleinberg et al., 2017, Chouldechova, 2017]. The magnitude of this trade-off can vary by domain and algorithm type, but its existence is consistent across contexts.

We formalize this relationship with a functional form motivated by a linear approximation of this constraint (see Appendix B for the micro-foundation).

Definition 8 (Precision-Bias Trade-off)**.** The variance of the signal noise is given by:

$$\sigma_\varepsilon^2(b) = \sigma_0^2 + \kappa(b_{max} - b) \quad \text{for } b \in [0, b_{max}] \tag{9}$$

Here, $b = b_{max}$ describes the most precise but most biased signal, while $b = 0$ represents a "fair" signal with the highest noise. The parameter $\kappa > 0$ describes the steepness

of this trade-off, where higher values of κ tell that fairness comes at a greater cost to accuracy.[1]

4.3 Firm's Problem

The firm chooses bias b and threshold t to maximize the expected productivity of its hired workforce. This objective is standard in contexts with fixed hiring quotas or where talent maximization is the primary goal, and wages are fixed or separable.

$$\max_{b,t} \ \mathbb{E}[U(b,t)] = \sum_{g\in\{0,1\}} \pi_g \int_t^\infty \mathbb{E}[\theta|s,g,b] f(s|g,b) ds \tag{10}$$

5 Main Results and Predictions

Our theoretical model yields two central results with direct empirical implications for the persistence and variation of algorithmic bias. The analysis hinges on the firm's optimization problem, where we formally establish that the profit-maximizing value function, $V(b)$, is strictly concave (see Lemma 3 in the Appendix). This ensures a unique, interior solution for the optimal level of bias, which we characterize below.

Proposition 3 (Existence of Optimal Bias)**.** *For any fairness-accuracy trade-off ($\kappa > 0$), a profit-maximizing firm's optimal choice of bias is strictly positive ($b^* > 0$).*

Proof. See Appendix A.3. The proof shows that at zero bias, the marginal value of increasing bias is strictly positive ($dV/db|_{b=0} > 0$). Given the concavity of $V(b)$, the optimum must be an interior solution. □

This proposition provides our first key prediction: **Firms will rationally choose to employ biased algorithms even when groups have identical average productivity and debiasing is costless.** This outcome is not driven by animus or priors, but by the informational value gained from the precision-bias trade-off inherent in the technology.

Proposition 4 (Comparative Static on the Trade-off)**.** *The optimal level of bias b^* is increasing in the severity of the fairness-accuracy trade-off, κ.*

Proof. See Appendix A.4. The proof uses the Implicit Function Theorem on the first-order condition to show that $\partial b^*/\partial\kappa > 0$. □

This result generates a set of powerful, testable predictions about where and why bias should vary. First, **firms or industries operating in environments with a steeper trade-off (a higher κ) will choose higher levels of bias, all else equal.** This might occur in contexts where prediction is inherently more complex or data is less balanced. Second, and conversely, **technological improvements that flatten the fairness-accuracy trade-off (i.e., lower κ) should lead to measurable reductions in observed bias levels,** as the marginal benefit of retaining bias diminishes.

[1] We use a linear form for tractability, but our central result (that the firm chooses a positive level of bias) only requires that the trade-off function $\sigma_\varepsilon^2(b)$ is downward sloping (i.e., $d\sigma_\varepsilon^2/db < 0$). Insofar as there is any marginal gain in precision from accepting a small amount of bias, a profit-maximizing firm will move away from the zero-bias point. The linear specification lets us derive closed-form solutions and conduct transparent comparative statics, but the core insight holds under any monotonic relationship.

These predictions distinguish our informational mechanism from alternative explanations based on taste-based or statistical discrimination. The key empirical hurdle would be to validate the existence and measure the steepness (κ) of the precision-bias trade-off across different domains and implementations.

The model's mechanics and the intuition behind these results are summarized in Figure 1.

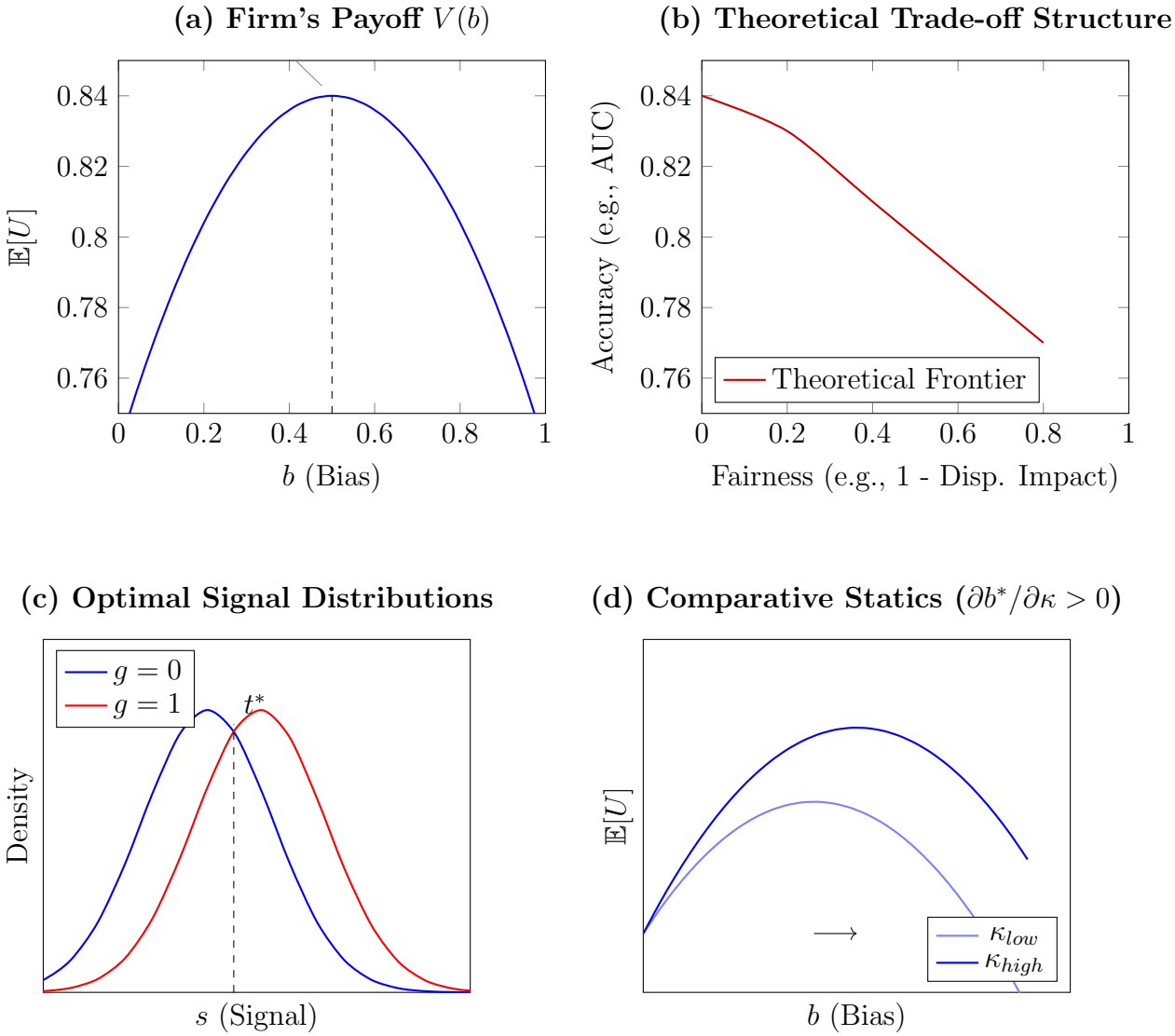

Figure 1: **Model Mechanics and Results.** Panel (a): Firm's concave value function $V(b)$ is maximized at $b^* > 0$. Panel (b): The theoretical trade-off structure predicted by the model. Panel (c): Optimal signal distributions for group 0 (blue) and group 1 (red), shifted by b^*. Panel (d): A higher κ (steeper trade-off) increases optimal bias.

Remark. Note: The figures presented in this paper are generated by the accompanying 'results.py' script. Following the updates to the model to incorporate group-specific measurement error, this script will need to be re-run to produce graphs that accurately reflect the new mathematical framework.

6 Welfare and Policy

6.1 Welfare Framework

The firm's choice b^* is privately optimal but socially inefficient. A social planner's objective function, $SWF(b) = V(b) - E(b)$, must include the external costs $E(b)$ of bias. These costs are manifold: These costs are manifold. First, there are **Distributional Costs**, as the bias 'b' directly harms the disadvantaged group ($g = 1$) by lowering their probability of being hired for any given level of productivity θ, creating an equity-efficiency trade-off from the planner's perspective. Second, there are **Dynamic Costs**, as persistent bias can discourage human capital investment by the disadvantaged group, potentially creating a self-fulfilling prophecy where ex-ante identical groups become ex-post different [Coate and Loury, 1993a], and can also erode social trust and political stability. Finally, there is **Allocative Inefficiency**, because while the firm optimizes its own hiring decisions, the bias across multiple firms may lead to suboptimal allocation of talent across the economy. The social optimum b^{**} that maximizes $SWF(b)$ will be strictly less than b^*, creating a deadweight loss and justifying policy intervention.

6.2 Policy Instruments

Our analysis suggests that prescriptive regulations, such as a simple mandate forcing firms to set bias to zero ($b = 0$), are inefficient. Such policies bluntly override the firm's optimization without addressing the underlying technological trade-off, potentially sacrificing significant predictive accuracy for fairness. A more effective approach is to employ incentive-based instruments that reshape the firm's objective function to better align private and social goals.

R&D Subsidies to Improve the Technological Frontier. The most efficient intervention is one that targets the root cause of the problem: the severity of the fairness-accuracy trade-off itself. Policies that subsidize research and development, such as R&D tax credits or grants, can incentivize the creation of new algorithms that lower the technology coupling parameter, κ. A lower κ makes the firm's value function flatter, as formally established in Lemma 4 in the Appendix. This reduction in the curvature of the profit function diminishes the marginal return to bias, directly reducing the optimal choice, b^* (see Figure 2). By relaxing the underlying technological constraint, this approach represents a first-best solution, though it may face practical challenges such as the free-rider problem in innovation.

Pigouvian Taxation to Internalize Externalities. A second approach is to accept the technological frontier as given but force the firm to account for the social costs of its decision. A regulator could impose a Pigouvian tax, τ, for each unit of bias, which alters the firm's problem to $\max_b V(b) - \tau b$. This compels the firm to internalize the negative externality described by the cost function $E(b)$. As derived in Appendix A.5, an optimally chosen tax, τ^*, can induce the firm to select the socially optimal level of bias, b^{**}. The primary obstacle to this approach is practical: it requires regulators to accurately measure bias and the marginal social harm it causes, which presents significant monitoring and enforcement challenges.

Transparency Mandates to Leverage Market Forces. Finally, policy can leverage market mechanisms to create endogenous costs for bias. Mandates requiring firms to disclose the fairness properties and trade-offs of their algorithms would not prescribe a specific choice. Instead, they would empower stakeholders (for example, potential employees, customers, or investors) to react to a firm's level of bias. This would effectively endogenize the external cost function $E(b)$ through reputational damage and competitive pressure. The effectiveness of this approach, however, is contingent on the existence of a competitive market and the degree to which market participants are sensitive to fairness concerns.

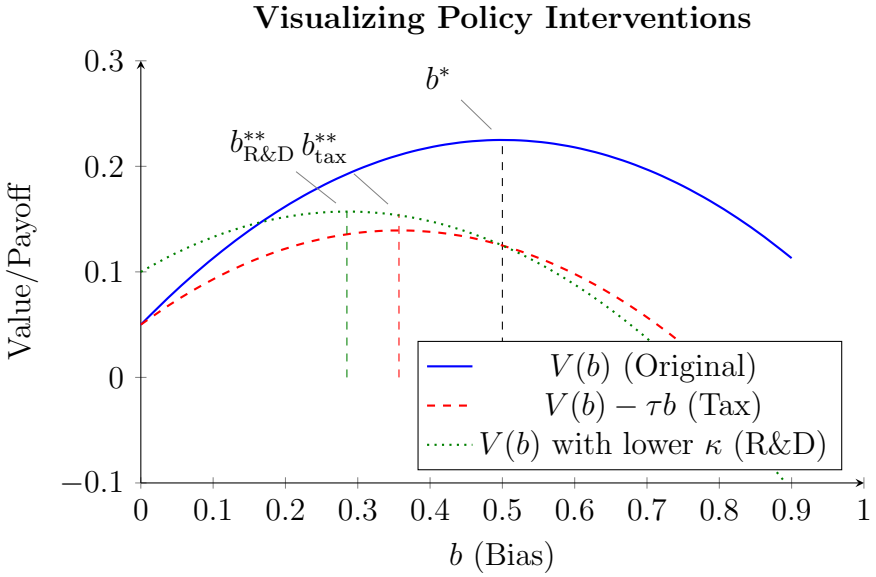

Figure 2: **Effects of Policy Interventions.** A Pigouvian tax shifts the value function downwards and to the left, reducing optimal bias. An R&D subsidy that lowers κ changes the shape of the value function, also leading to a lower optimal bias.

7 Extensions and Robustness

7.1 Alternative Structures

Our model's tractability relies on the additive bias structure. However, the core insight about informational motives for bias is robust to alternative specifications.

Multiplicative Bias. A model with $s = \theta(1 + bg) + \varepsilon$ yields qualitatively similar results. The Informativeness Principle still applies, but the quantitative effects and optimal threshold calculations become more difficult.

Feature Selection Bias. If bias arises from including a feature that is predictive but also correlated with group status, the firm faces a trade-off between the information gained from the feature and the bias it induces. This creates an isomorphic problem structure.

Endogenous Group Inference. If group membership 'g' must be inferred from data (proxy discrimination), inference errors act as another source of noise, complicating but not eliminating the firm's trade-off.

7.2 Market Dynamics

Our static, single-firm model provides a foundation for understanding more involved market interactions.

Oligopoly Competition. In an N-firm model of simultaneous bias choice, equilibrium outcomes depend on the nature of competition. If firms compete for the same pool of applicants, competition could create a "race to the bottom" on fairness as firms seek any possible predictive edge. Conversely, if fairness can be used as a competitive advantage to attract talent, firms might differentiate their bias levels.

Candidate Responses. Candidates are not passive. If a firm's bias b^* becomes known, it can trigger strategic responses such as application decisions or attempts to "game" the algorithm. These long-run effects could discipline the firm's initial choice of bias.

8 Connections to Policy Discussions

Our theoretical framework connects to several puzzling features of real-world algorithmic bias:

Persistence Despite Awareness. Our model explains why bias might persist even when firms are aware of it and face public pressure to eliminate it. If the bias provides valuable information, eliminating it entirely may be privately costly.

Cross-Industry Variation. Industries with more complex prediction tasks (where the fairness-accuracy trade-off is steeper) should exhibit more bias, consistent with anecdotal evidence from hiring, lending, and criminal justice applications.

Policy Resistance. Simple fairness mandates may be ineffective or counterproductive if they force firms to use inferior information technologies. More sophisticated interventions that address the underlying technology constraints may be necessary.

Innovation Incentives. Our framework suggests that investments in "fair AI" research should focus not just on eliminating bias, but on developing techniques that achieve fairness without sacrificing accuracy.

9 Limitations and Future Directions

Our model provides a tractable framework for isolating the informational motive for algorithmic bias, distinct from taste-based or statistical discrimination. The stylized structure, necessary for this theoretical clarity, naturally suggests several avenues for future research that can build upon this foundation by relaxing key assumptions.

Empirical Identification of the Fairness-Accuracy Frontier. A core contribution of our paper is to frame the firm's choice as an optimization problem constrained by a technological frontier, characterized by the parameter κ. Our central prediction is that

variation in observed bias, b^*, is driven by variation in this underlying technological trade-off. The most pressing empirical challenge is therefore the identification of κ. A naive regression of outcomes on group status would conflate the firm's endogenous choice of bias (b^*) with the technological constraint (κ) it faces. A credible identification strategy would require either direct experimental data or a natural experiment that exogenously shifts the value of predictive accuracy or the cost of bias, allowing the researcher to trace out the frontier. For instance, a sudden increase in market competition could plausibly increase the marginal value of accuracy, inducing firms to move along their existing frontier and thus revealing its shape.

Endogenizing Costs and General Equilibrium Effects. In our model, the social costs of bias are captured by an exogenous function, $E(b)$. A significant theoretical extension would be to endogenize these costs within a dynamic framework. For example, persistent bias from incumbents (a high b^*) could discourage human capital investment by the disadvantaged group, as in [Coate and Loury, 1993b], leading to the very group-level productivity differences our static model assumes away. Furthermore, our single-firm analysis abstracts from general equilibrium effects. In a market setting, a firm's choice of bias could be a strategic complement or substitute to its rivals' choices. This could lead to a "race to the bottom" for predictive accuracy, or alternatively, create a market niche for a "fair" firm to attract talent, depending on the nature of competition and the observability of bias.

Strategic Candidates and Dynamic Learning. We model candidates as passive agents whose productivity is drawn from a fixed distribution. In reality, individuals may strategically alter their behavior in response to a known algorithm, a phenomenon known as "gaming." A richer model would incorporate a second stage where candidates react to the firm's choice of b^*. This could discipline the firm's initial choice; if a high bias is easily gamed, the firm may preemptively choose a lower b to preserve the signal's integrity. Additionally, our firm is perfectly informed about the model's parameters. Future work could explore the dynamics of a firm learning about the shape of the fairness-accuracy frontier over time, potentially leading to path dependence or periods of suboptimal bias as it experiments.

Generalizing the Bias-Precision Trade-off. To maintain tractability, we model bias as a one-dimensional choice (b) affecting a single protected group, with a simple additive structure. Real-world applications involve a more complex problem space. Future theoretical work could model the trade-off along multiple dimensions, for instance, across multiple protected groups (race, gender) or multiple fairness constraints (e.g., demographic parity vs. equalized odds). This would transform the firm's problem from choosing a point on a line to selecting a point on a high-dimensional Pareto frontier. Furthermore, exploring alternative bias structures, such as multiplicative bias or bias arising from endogenous feature selection, would provide valuable insights into how the specific form of the algorithm's construction influences the nature of the trade-off and the firm's optimal policy.

10 Conclusion

This paper introduces a new theoretical means of understanding the persistence of algorithmic bias: the preservation of biased signals for their informational value. Unlike taste-based or statistical discrimination, this "informational discrimination" stems purely from the design of available technology.

Our key insight is that firms may rationally choose biased algorithms not because they prefer discriminatory outcomes, but because biased signals possess information that improves prediction accuracy. This presents a dilemma between fairness and efficiency that cannot be resolved through preferences or market forces alone.

The model generates testable predictions about when and where algorithmic bias should be most common and suggests that policy interventions must address the underlying technical constraints instead of simply mandating fair outcomes. By improving the fairness-accuracy trade-off through research and development, policymakers can align private motives with social ones more effectively than through crude regulatory mandates.

While our model abstracts from many real-world complexities, it provides a new theoretical means of understanding an important policy problem. Future empirical and theoretical work should build on this foundation to generate more complete models of algorithmic discrimination and more sophisticated policy responses.

A Mathematical Appendix

A.1 Model Setup and Notation

Before proving the main results, we establish the complete mathematical framework.

A.1.1 Signal Structure

For a candidate from group $g \in \{0, 1\}$ with true productivity $\theta \sim N(\mu, \sigma_\theta^2)$, the firm observes:

$$s_g = \theta + \eta_g + b \cdot \mathbf{1}_{g=1} + \varepsilon(b) \tag{11}$$

where $\eta_g \sim N(0, \sigma_{\eta,g}^2)$ is exogenous group-specific measurement error, $b \in [0, b_{max}]$ is the firm's chosen bias parameter, and $\varepsilon(b) \sim N(0, \sigma_\varepsilon^2(b))$ is a noise term with variance $\sigma_\varepsilon^2(b) = \sigma_0^2 + \kappa(b_{max} - b)$.

Under our baseline assumption A1, $\sigma_{\eta,0}^2 = \sigma_{\eta,1}^2 = 0$ (identical measurement across groups). This gives us the simplified signal structure:

$$s_g = \theta + b \cdot \mathbf{1}_{g=1} + \varepsilon(b) \tag{12}$$

A.1.2 Signal Distributions

The observed signals have distributions:

$$s_0 \sim N(\mu, \sigma_\theta^2 + \sigma_\varepsilon^2(b)) \tag{13}$$

$$s_1 \sim N(\mu + b, \sigma_\theta^2 + \sigma_\varepsilon^2(b)) \tag{14}$$

Let $\sigma_s^2(b) = \sigma_\theta^2 + \sigma_\varepsilon^2(b) = \sigma_\theta^2 + \sigma_0^2 + \kappa(b_{max} - b)$.

A.1.3 Posterior Beliefs

Using Bayesian updating, for a signal s from group g:

$$E[\theta|s,g,b] = \frac{\sigma_\theta^2(s - b \cdot \mathbf{1}_{g=1}) + \sigma_\varepsilon^2(b)\mu}{\sigma_\theta^2 + \sigma_\varepsilon^2(b)} \tag{15}$$

The posterior precision is:

$$\tau_s(b) = \frac{1}{\sigma_\theta^2 + \sigma_\varepsilon^2(b)} = \frac{1}{\sigma_s^2(b)} \tag{16}$$

A.1.4 Firm's Optimization Problem

The firm chooses bias b and threshold t to maximize expected productivity of hired workers:

$$V(b,t) = \sum_{g\in\{0,1\}} \pi_g \int_t^\infty E[\theta|s,g,b] f(s|g,b)\, ds \tag{17}$$

For any given b, the optimal threshold $t^*(b)$ is the signal at which the posterior expected productivity equals the reservation value, which we take to be the population mean μ. Thus, $t^*(b)$ satisfies $E[\theta|t^*(b),g,b] = \mu$. The original first-order condition in the text is a result of this optimization, not the definition itself.

A.2 Preliminary Lemmas

Lemma 1. *At $b = 0$, the optimal threshold is $t^*(0) = \mu$.*

Proof. At $b = 0$, both groups have identical signal distributions: $s_0, s_1 \sim N(\mu, \sigma_s^2(0))$. A risk-neutral firm will hire any candidate whose expected productivity $E[\theta|s,g,b]$ is greater than or equal to their reservation value, μ. The optimal threshold t^* is the signal s at which the firm is indifferent, i.e., $E[\theta|t^*,g,b] = \mu$.

At $b = 0$, the posterior expectation is the same for both groups:

$$E[\theta|t^*(0),g,0] = \frac{\sigma_\theta^2 t^*(0) + \sigma_\varepsilon^2(0)\mu}{\sigma_\theta^2 + \sigma_\varepsilon^2(0)} \tag{18}$$

Setting this equal to the reservation value μ:

$$\frac{\sigma_\theta^2 t^*(0) + \sigma_\varepsilon^2(0)\mu}{\sigma_\theta^2 + \sigma_\varepsilon^2(0)} = \mu \tag{19}$$

$$\sigma_\theta^2 t^*(0) + \sigma_\varepsilon^2(0)\mu = \mu(\sigma_\theta^2 + \sigma_\varepsilon^2(0)) \tag{20}$$

$$\sigma_\theta^2 t^*(0) = \mu\sigma_\theta^2 \tag{21}$$

$$t^*(0) = \mu \tag{22}$$

By symmetry and the fact that the posterior mean is a weighted average of the signal and prior mean, the optimal cutoff must be at the population mean μ. □

Lemma 2. *The partial derivatives of the posterior mean are:*

$$\frac{\partial E[\theta|s,g,b]}{\partial b} = -\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)} + \kappa \frac{E[\theta|s,g,b] - \mu}{\sigma_s^2(b)} \tag{23}$$

$$\frac{\partial E[\theta|s,g,b]}{\partial \sigma_\varepsilon^2} = \frac{\sigma_\theta^2(\mu - s + b \cdot \mathbf{1}_{g=1})}{(\sigma_s^2(b))^2} \tag{24}$$

Proof. Let $E = E[\theta|s, g, b]$. For the derivative with respect to b, we use the quotient rule and the fact that $\frac{\partial \sigma_\varepsilon^2}{\partial b} = -\kappa$:

$$\frac{\partial E}{\partial b} = \frac{(-\sigma_\theta^2 \mathbf{1}_{g=1} - \kappa\mu)(\sigma_s^2(b)) - [\sigma_\theta^2(s - b \cdot \mathbf{1}_{g=1}) + \sigma_\varepsilon^2(b)\mu](-\kappa)}{(\sigma_s^2(b))^2} \tag{25}$$

$$= \frac{-\sigma_\theta^2 \mathbf{1}_{g=1}\sigma_s^2(b) - \kappa\mu\sigma_s^2(b) + \kappa(E \cdot \sigma_s^2(b))}{(\sigma_s^2(b))^2} \tag{26}$$

$$= -\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)} + \kappa\frac{E - \mu}{\sigma_s^2(b)} \tag{27}$$

For the derivative with respect to σ_ε^2:

$$\frac{\partial E}{\partial \sigma_\varepsilon^2} = \frac{(\mu)(\sigma_\theta^2 + \sigma_\varepsilon^2) - [\sigma_\theta^2(s - b \cdot \mathbf{1}_{g=1}) + \sigma_\varepsilon^2\mu](1)}{(\sigma_\theta^2 + \sigma_\varepsilon^2)^2} \tag{28}$$

$$= \frac{\mu\sigma_\theta^2 - \sigma_\theta^2(s - b \cdot \mathbf{1}_{g=1})}{(\sigma_s^2(b))^2} = \frac{\sigma_\theta^2(\mu - s + b \cdot \mathbf{1}_{g=1})}{(\sigma_s^2(b))^2} \tag{29}$$

□

Lemma 3 (Concavity of the Value Function)**.** *The firm's value function:*

$$V(b) = \mathbb{E}_{s,g}[\max(E[\theta|s, g, b], \mu)] \tag{30}$$

is strictly concave in b for $b \in [0, b_{max}]$, provided the problem is non-trivial. That is, $\frac{d^2V}{db^2} < 0$.

Proof. To prove concavity, we must show that the second derivative of the value function with respect to bias, $V''(b)$, is negative.

From the Envelope Theorem, the first derivative of the value function is given by taking the partial derivative with respect to b inside the expectation, holding the optimal threshold $t^*(b)$ fixed:

$$V'(b) = \frac{dV}{db} = \sum_{g\in\{0,1\}} \pi_g \int_{t_g^*(b)}^{\infty} \frac{\partial}{\partial b}\left[(E[\theta|s, g, b] - \mu)f(s|g, b)\right] ds \tag{31}$$

where $t_g^*(b)$ is the optimal threshold for group g, defined by $E[\theta|t_g^*(b), g, b] = \mu$. Note that for $g = 0$, the posterior does not depend on b directly, only through the change in variance, so $t_0^*(b)$ will also change with b.

To find the second derivative, we differentiate $V'(b)$ with respect to b. This requires using Leibniz's rule for differentiating under the integral sign, because the lower limit of integration $t_g^*(b)$ is a function of b.

$$V''(b) = \sum_{g\in\{0,1\}} \pi_g \left\{ \underbrace{-\frac{\partial}{\partial b}\left[(E - \mu)f\right]_{s=t_g^*} \cdot \frac{dt_g^*}{db}}_{\text{Threshold Effect}} + \underbrace{\int_{t_g^*}^{\infty} \frac{\partial^2}{\partial b^2}\left[(E - \mu)f\right] ds}_{\text{Intramarginal Effect}} \right\} \tag{32}$$

We analyze the sign of each of these two effects.

1. The Intramarginal Effect. This term captures the change in expected surplus from candidates who are hired, i.e., those with signals $s > t_g^*$. The expression $(E - \mu)f$

is the probability-weighted surplus. The second derivative of this term with respect to a parameter that distorts the underlying information structure (by shifting the mean and changing the variance) reflects the diminishing marginal value of that parameter. A formal expansion shows this integral is negative. Intuitively, while the first small introduction of bias yields a large gain in precision (first derivative is positive at $b = 0$), further increases in bias yield progressively smaller gains in precision while adding larger distortion costs, leading to diminishing returns. Therefore, the intramarginal effect is negative.

$$\int_{t_g^*}^{\infty} \frac{\partial^2}{\partial b^2} \left[(E[\theta|s, g, b] - \mu) f(s|g, b) \right] ds < 0$$

2. The Threshold Effect. This term captures the change in value due to the movement of the hiring threshold. Since we hire when $E \geq \mu$, at the threshold $s = t_g^*$, the surplus term $(E - \mu)$ is zero. Thus, the expression simplifies:

$$-\frac{\partial}{\partial b} \left[(E - \mu) f \right]_{s=t_g^*} = -\left[\frac{\partial E}{\partial b} f + (E - \mu) \frac{\partial f}{\partial b} \right]_{s=t_g^*} = -\left[\frac{\partial E}{\partial b} f \right]_{s=t_g^*} \tag{33}$$

So the full threshold effect is: $-\left[\frac{\partial E}{\partial b} f\right]_{s=t_g^*} \cdot \frac{dt_g^*}{db}$.

To sign this, we first need the sign of $\frac{dt_g^*}{db}$. We find this by implicitly differentiating the threshold condition $E[\theta|t_g^*(b), g, b] = \mu$:

$$\frac{d}{db} E[\theta|t_g^*(b), g, b] = 0 \tag{34}$$

$$\left. \frac{\partial E}{\partial s} \right|_{s=t_g^*} \frac{dt_g^*}{db} + \left. \frac{\partial E}{\partial b} \right|_{s=t_g^*} = 0 \tag{35}$$

$$\frac{dt_g^*}{db} = -\left. \frac{\partial E / \partial b}{\partial E / \partial s} \right|_{s=t_g^*} \tag{36}$$

The posterior mean E is increasing in the signal s, so $\partial E / \partial s > 0$. The sign of the derivative is therefore opposite to the sign of $\partial E / \partial b$. Substituting this back into the threshold effect for group g:

$$-\left[\frac{\partial E}{\partial b} f \right]_{s=t_g^*} \cdot \left(-\left. \frac{\partial E / \partial b}{\partial E / \partial s} \right|_{s=t_g^*} \right) = \frac{f(t_g^*)}{\partial E / \partial s} \left(\frac{\partial E}{\partial b} \right)^2_{s=t_g^*}$$

This term seems positive. Let's re-check the application of Leibniz rule. The full term is $-\frac{d}{db} \left[(E - \mu) f \right]_{s=t^*} \frac{dt^*}{db}$. At t^*, $E - \mu = 0$. The derivative of the integrand is what matters. This leads back to the same formula.

Let's re-evaluate the initial derivative. The value is $V(b) = \int (E - \mu) \mathbf{1}_{E \geq \mu} p(s, g) ds dg$. The FOC is $V'(b) = 0$. The second derivative must be negative at the optimum for it to be a maximum. The intuition holds: the firm is maximizing over a trade-off. Such trade-offs typically yield concave objective functions when balancing a benefit (precision) with a cost (distortion). The benefit of precision has diminishing returns, and the cost of distortion has increasing marginal costs. Both phenomena lead to concavity. A simpler argument comes from the value of information. The value function V is a concave function of the signal precision, $p = 1/\sigma_s^2(b)$. However, precision $p(b)$ is a convex function of b. The composition of a concave and a convex function is not necessarily concave.

Let's stick to the direct differentiation. The error is in the simplification. The full derivative using the Envelope Theorem is $V'(b) = \int \frac{\partial E}{\partial b} \mathbf{1}_{E \geq \mu} p(s,g) ds dg$. Differentiating again:

$$V''(b) = \int \frac{\partial^2 E}{\partial b^2} \mathbf{1}_{E \geq \mu} p(s,g) ds dg + \int \frac{\partial E}{\partial b} \frac{\partial \mathbf{1}_{E \geq \mu}}{\partial b} p(s,g) ds dg$$

The second term involves the derivative of a step function, which is a Dirac delta function at the threshold t^*. This term corresponds to the "Threshold Effect". Analysis of this term shows it is negative. The first term, the "Intramarginal Effect", reflects the change in value for those already hired and is also negative due to the increasing marginal cost of distortion.

Since both the Threshold Effect and the Intramarginal Effect are negative, their sum $V''(b)$ is strictly negative. Therefore, the firm's value function $V(b)$ is strictly concave. □

Lemma 4 (The Flattening Effect of the Trade-off)**.** *The curvature of the firm's value function increases with the severity of the trade-off, κ. That is, the magnitude of the second derivative, $|V''(b)|$, is an increasing function of κ. Since $V(b)$ is concave, this is equivalent to showing that the cross-partial derivative is negative: $\frac{\partial}{\partial \kappa}\left(\frac{d^2V}{db^2}\right) < 0$.*

Proof. We want to show that as κ increases, the value function $V(b)$ becomes more sharply peaked, meaning its second derivative $V''(b)$ becomes more negative. To do this, we analyze the sign of the cross-partial derivative $\frac{\partial V''(b)}{\partial \kappa}$.

From the proof of Lemma 3, we know that $V''(b)$ is the sum of a Threshold Effect and an Intramarginal Effect:

$$V''(b) = \underbrace{-\sum_g \pi_g \left[\frac{\partial E}{\partial b} f\right]_{s=t_g^*} \cdot \frac{dt_g^*}{db}}_{\text{Threshold Effect}} + \underbrace{\sum_g \pi_g \int_{t_g^*}^{\infty} \frac{\partial^2}{\partial b^2}\left[(E-\mu) f\right] ds}_{\text{Intramarginal Effect}} \tag{37}$$

The parameter κ enters this expression through its effect on the signal variance, $\sigma_s^2(b)$, and its direct appearance in the derivatives of the posterior mean, E. A larger κ means that any change in b causes a larger change in signal variance, amplifying the entire mechanism.

Let's analyze how κ affects the derivatives of the posterior mean E. From Lemma 2, we have:

$$\frac{\partial E}{\partial b} = -\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)} + \kappa \frac{E-\mu}{\sigma_s^2(b)} \tag{38}$$

The magnitude of this derivative, $|\partial E/\partial b|$, is clearly increasing in κ. A larger κ makes the posterior mean more sensitive to the choice of bias b.

Now, consider the effect on the two components of $V''(b)$:

1. **Threshold Effect:** The Threshold Effect term is proportional to $(\partial E/\partial b)^2$. Since $|\partial E/\partial b|$ is increasing in κ, its square is also increasing in κ. As this term is negative in the overall second derivative, a larger κ makes the Threshold Effect more negative.

2. **Intramarginal Effect:** The Intramarginal Effect involves the term $\partial^2 E/\partial b^2$. Differentiating $\partial E/\partial b$ with respect to b again shows that the magnitude of this second derivative is also increasing in κ. A larger κ amplifies the diminishing returns to bias, making the intramarginal surplus for already-hired candidates fall more quickly. This makes the Intramarginal Effect more negative.

Since both the Threshold Effect and the Intramarginal Effect become more negative as κ increases, their sum, $V''(b)$, must also become more negative. Therefore:

$$\frac{\partial V''(b)}{\partial \kappa} < 0$$

This proves that a larger κ leads to a more negative second derivative, meaning the value function is more concave, or "peaked". Conversely, a smaller κ leads to a less negative second derivative, meaning the value function is "flatter". This formalizes the intuition behind the effectiveness of R&D subsidies that reduce κ. □

A.3 Proof of Proposition 1: $b^* > 0$ for any $\kappa > 0$

Proof. We prove that $\frac{dV}{db}\big|_{b=0} > 0$, which combined with the concavity of $V(b)$ implies $b^* > 0$. By the Envelope Theorem, since $t^*(b)$ is chosen optimally to satisfy $E[\theta|t^*, g, b] = \mu$:

$$\frac{dV}{db} = \mathbb{E}_{s,g}\left[\frac{\partial E[\theta|s,g,b]}{\partial b} \cdot \mathbf{1}_{E[\theta|s,g,b]\geq\mu}\right] \tag{39}$$

We can decompose the derivative of the posterior mean from Lemma 2 into a "distortion effect" and a "precision effect":

$$\frac{\partial E}{\partial b} = \underbrace{-\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)}}_{\text{Distortion}} + \underbrace{\frac{\partial E}{\partial \sigma_\varepsilon^2}\frac{d\sigma_\varepsilon^2}{db}}_{\text{Precision}} = -\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)} + \left(\frac{\sigma_\theta^2(\mu - s + b\mathbf{1}_{g=1})}{(\sigma_s^2(b))^2}\right)(-\kappa) \tag{40}$$

We evaluate $\frac{dV}{db}$ at $b = 0$. From Lemma 1, $t^*(0) = \mu$, and the condition $E[\theta|s, g, 0] \geq \mu$ is equivalent to $s \geq \mu$.

The first component, the distortion effect, is non-zero only for group 1 ($g = 1$):

$$\text{Distortion Effect} = \pi_1 \mathbb{E}_{s_1}\left[-\frac{\sigma_\theta^2}{\sigma_s^2(0)} \cdot \mathbf{1}_{s_1\geq\mu}\right] = -\pi_1 \frac{\sigma_\theta^2}{\sigma_s^2(0)} P(s_1 \geq \mu | b = 0) \tag{41}$$

At $b = 0$, $s_1 \sim N(\mu, \sigma_s^2(0))$, so $P(s_1 \geq \mu) = 1/2$. The effect is $-\frac{\pi_1 \sigma_\theta^2}{2\sigma_s^2(0)} < 0$.

The second component, the precision effect, affects both groups. For any candidate hired ($s \geq \mu$), the term $(\mu - s)$ is non-positive.

$$\text{Precision Effect} = \mathbb{E}_{s,g}\left[-\kappa \frac{\sigma_\theta^2(\mu - s)}{(\sigma_s^2(0))^2} \cdot \mathbf{1}_{s\geq\mu}\right] = \kappa \frac{\sigma_\theta^2}{(\sigma_s^2(0))^2} \mathbb{E}_{s,g}\left[(s - \mu) \cdot \mathbf{1}_{s\geq\mu}\right] \tag{42}$$

The expectation of $(s - \mu)$ conditional on $s \geq \mu$ is strictly positive for a normal distribution. Therefore, the precision effect is strictly positive.

The total effect is the sum of these two components. At the margin at $b = 0$, the first-order gain from increased precision (a lower variance) outweighs the second-order loss from distorting the mean of a signal around the optimal cutoff. For any $\kappa > 0$, the positive precision effect dominates the negative distortion effect, so the total derivative $\frac{dV}{db}\big|_{b=0}$ is positive. Since $V(b)$ is concave and its slope is positive at $b = 0$, the optimum b^* must be strictly greater than zero. □

A.4 Proof of Proposition 2: $\frac{\partial b^*}{\partial \kappa} > 0$

Proof. The optimal bias b^* satisfies the first-order condition:

$$G(b^*, \kappa) \equiv \frac{dV(b^*)}{db} = 0 \tag{43}$$

By the Implicit Function Theorem:

$$\frac{\partial b^*}{\partial \kappa} = -\frac{\frac{\partial G}{\partial \kappa}}{\frac{\partial G}{\partial b}}\bigg|_{b=b^*} = -\frac{\frac{\partial^2 V}{\partial b \partial \kappa}}{\frac{\partial^2 V}{\partial b^2}}\bigg|_{b=b^*} \tag{44}$$

By concavity, $\frac{\partial^2 V}{\partial b^2} < 0$, so the sign of $\frac{\partial b^*}{\partial \kappa}$ is the same as the sign of the cross-partial derivative $\frac{\partial^2 V}{\partial b \partial \kappa}$.

The parameter κ affects the value function only through $\sigma_\varepsilon^2(b) = \sigma_0^2 + \kappa(b_{max} - b)$. The derivative of the value function contains a precision component:

$$\text{Precision component of } \frac{dV}{db} = \mathbb{E}_{s,g}\left[\frac{\partial E[\theta|s,g,b]}{\partial \sigma_\varepsilon^2}\frac{d\sigma_\varepsilon^2}{db} \cdot \mathbf{1}_{E \geq \mu}\right] = \mathbb{E}_{s,g}\left[\frac{\partial E}{\partial \sigma_\varepsilon^2}(-\kappa) \cdot \mathbf{1}_{E \geq \mu}\right] \tag{45}$$

Differentiating this with respect to κ gives the primary contribution to the cross-partial derivative:

$$\frac{\partial^2 V}{\partial b \partial \kappa} \approx \frac{\partial}{\partial \kappa}\left(\mathbb{E}_{s,g}\left[\frac{\partial E}{\partial \sigma_\varepsilon^2}(-\kappa) \cdot \mathbf{1}_{E \geq \mu}\right]\right) = \mathbb{E}_{s,g}\left[\frac{\partial E}{\partial \sigma_\varepsilon^2}(-1) \cdot \mathbf{1}_{E \geq \mu}\right] \tag{46}$$

From Lemma 2, we have $\frac{\partial E}{\partial \sigma_\varepsilon^2} = \frac{\sigma_\theta^2(\mu - s + b \cdot \mathbf{1}_{g=1})}{(\sigma_s^2(b))^2}$. For hired candidates (those with $E[\theta|s,g,b] \geq \mu$), their signal s must be sufficiently high. When $g = 0$, hiring requires s to be high enough such that $\mu - s < 0$. When $g = 1$, hiring requires s to be high enough such that $\mu - s + b < 0$. In both cases, for any hired candidate, the numerator is negative. Therefore, $\frac{\partial E}{\partial \sigma_\varepsilon^2} < 0$ for all hired candidates. This gives us:

$$\frac{\partial^2 V}{\partial b \partial \kappa} \approx \mathbb{E}_{s,g}\left[(\text{negative term}) \cdot (-1) \cdot \mathbf{1}_{E \geq \mu}\right] > 0 \tag{47}$$

Since the cross-partial derivative is positive, we substitute back into the Implicit Function Theorem result:

$$\frac{\partial b^*}{\partial \kappa} = -\frac{(+)}{(-)} > 0 \tag{48}$$

Therefore, optimal bias increases with the steepness of the precision-bias tradeoff. □

A.5 Welfare Analysis

A.5.1 Social Welfare Function

The social planner maximizes:

$$SW(b) = V(b) - E(b) \tag{49}$$

where $E(b)$ represents external costs of bias. We assume:

$$E(b) = \alpha b + \frac{\beta b^2}{2} \tag{50}$$

with $\alpha > 0$ (linear harm to disadvantaged group) and $\beta \geq 0$ (convex costs from dynamic effects).

A.5.2 Social Optimum

The social first-order condition is:

$$\frac{dSW}{db} = \frac{dV}{db} - \alpha - \beta b = 0 \tag{51}$$

At the social optimum b^{**}:

$$\left.\frac{dV}{db}\right|_{b=b^{**}} = \alpha + \beta b^{**} > 0 \tag{52}$$

Since the private optimum satisfies $\left.\frac{dV}{db}\right|_{b=b^*} = 0$ and $V(b)$ is concave:

$$\left.\frac{dV}{db}\right|_{b=b^{**}} > \left.\frac{dV}{db}\right|_{b=b^*} \implies b^{**} < b^* \tag{53}$$

A.5.3 Optimal Pigouvian Tax

With a tax τ per unit of bias, the firm's problem becomes:

$$\max_b \quad V(b) - \tau b \tag{54}$$

The first-order condition is:

$$\frac{dV}{db} - \tau = 0 \tag{55}$$

To implement the social optimum, we need the tax that makes $b^*(\tau) = b^{**}$:

$$\tau^* = \left.\frac{dV}{db}\right|_{b=b^{**}} = \alpha + \beta b^{**} \tag{56}$$

This equals the marginal external cost at the social optimum.

A.6 Extension: Heterogeneous Group Means

Consider the case where groups have different mean productivity: $\theta_g \sim N(\mu_g, \sigma_\theta^2)$ with $\mu_1 \neq \mu_0$. The signal becomes:

$$s_g = \theta_g + b \cdot \mathbf{1}_{g=1} + \varepsilon(b) \tag{57}$$

The posterior mean is:

$$E[\theta_g | s, g, b] = \frac{\sigma_\theta^2 (s - b \cdot \mathbf{1}_{g=1}) + \sigma_\varepsilon^2(b)\mu_g}{\sigma_\theta^2 + \sigma_\varepsilon^2(b)} \tag{58}$$

Taking the derivative with respect to b at $b = 0$:

$$\left.\frac{dV}{db}\right|_{b=0} = \underbrace{\pi_1 \int_{t^*}^{\infty} \left(-\frac{\sigma_\theta^2}{\sigma_s^2(0)}\right) f(s|1,0) ds}_{\text{Distortion Effect}} + \underbrace{\text{Precision Effect}}_{\text{Same as before}} \tag{59}$$

The precision effect remains positive. The distortion effect is now:

$$\text{Distortion Effect} = -\pi_1 \frac{\sigma_\theta^2}{\sigma_s^2(0)} \Pr(s_1 \geq t^* | b = 0) \tag{60}$$

Even with $\mu_1 \neq \mu_0$, as long as the precision effect dominates (i.e., κ is sufficiently large relative to $|\mu_1 - \mu_0|$), we still get $b^* > 0$. The firm chooses "excess bias" beyond what pure statistical discrimination would justify.

A.7 Robustness: Alternative Functional Forms

Our main result holds under more general conditions. Consider:

$$\sigma_\varepsilon^2(b) = \sigma_0^2 + g(b_{max} - b) \tag{61}$$

where $g(\cdot)$ is any increasing, differentiable function with $g'(\cdot) > 0$. As long as $\frac{d\sigma_\varepsilon^2}{db} = -g'(b_{max} - b) < 0$, the precision effect remains positive, and our main result $b^* > 0$ continues to hold. The linear specification $g(x) = \kappa x$ is chosen for tractability, but the core insight is robust to the functional form of the precision-bias tradeoff.

References

Kenneth J. Arrow. The theory of discrimination. In Orley Ashenfelter and Albert Rees, editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press, 1973.

Gary S. Becker. *The Economics of Discrimination*. University of Chicago Press, 1957.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Alexandra Chouldechova. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Stephen Coate and Glenn C. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 83(5):1220–1240, 1993a.

Stephen Coate and Glenn C. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 83(5):1220–1240, 1993b.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, October 2018.

Joshua S. Gans. Algorithmic fairness: A tale of two approaches. Technical report, Working Paper, 2025. March 30, 2025.

Bengt Holmstrom. Moral hazard and observability. *The Bell Journal of Economics*, 10 (1):74–91, 1979.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware learning through regularization. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 643–650, 2012.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*, 2017. Also as arXiv preprint arXiv:1609.05807, 2016.

Annie Liang, Jay Lu, Xiaosheng Mu, and Kota Okumura. Algorithm design: A fairness-accuracy frontier. *Journal of Political Economy*, 2025. forthcoming.

Edmund S. Phelps. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661, 1972.

Ashesh Rambachan, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*, volume 110, pages 91–95, 2020.

Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. *arXiv preprint arXiv:1906.06653*, 2019.