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

Kofi Hair-Ralston*
kofibhairralston@gmail.com

Daksh Mathapati
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 David Weinstein, Alex Imas, and Aislinn Bohren for their invaluable feedback. We are especially grateful to Elliot Lipnowski and Peter Hull for feedback that significantly shaped the direction of this project.All remaining errors are our own.

1 Introduction

In 2018, Amazon scrapped an artificial intelligence recruiting tool after uncovering a major flaw: it disliked women [Dastin, 2018]. The tool was designed to automate the search for top software engineers by looking at parsed resumes. Yet, the machine learning models taught themselves that male candidates were preferable, shirking resumes containing the word "women's" (as in "women's chess club captain") and downgrading graduates from all-women's colleges. Amazon's engineers tried to adjust the system to make it neutral, but they could not guarantee that the machine would not come up with new discriminatory ways of sorting candidates. Amazon ultimately ended up scrapping the project.

This was one instance of a problem that's becoming common quickly. Amazon is a sophisticated, profit-maximizing firm. It had no shortage of technical or computational power, and its goal was to find the best possible talent, not to discriminate. Yet, its try to build a purely meritocratic algorithm birthed a biased system that was ultimately unfixable. For years, eerily similar problems involving persistent, unintended bias have come up in algorithms for credit scoring, medical diagnoses, and criminal justice risk assessment, to name a few. These cases are hard 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, because they assume that group identity is used as a proxy because of a lack of individual-level data; the algorithms we study, however, are data-rich and sometimes have access to even hundreds of individual-specific features.

We propose and formalize a third mechanism for discrimination about the inherent technological trade-offs in prediction. We call this mechanism "informational discrimination." We offer that there exists a technological frontier between fairness and accuracy [Kleinberg et al., 2017, Chouldechova, 2017]. Companies' attempts to "fix" an algorithm's disparate impact (debiasing) often come at the cost of reducing its predictive ability. We thus 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, representing the steepness of the 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 Holmstrom's Informativeness Principle [Holmstrom, 1979]. The firm has a choice between a perfectly "fair" signal ($b = 0$), which treats all groups identically but is noisy, and a "biased" signal ($b > 0$), which often disadvantages one group but gives a greater overall prediction accuracy. Even though the biased signal leads to discrimination, its far better informational content makes it privately optimal for the firm. The marginal gain in accuracy from accepting a small amount of bias ends up outweighing the marginal cost of the signal's distortion.

The model shows 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 gives us testable predictions about where and why we should observe this algorithmic bias. Industries with more complicated prediction tasks or with far less balanced data (higher κ) should show

more bias, independent of their discriminatory intent. We then examine the welfare implications of this mechanism. The firm's privately optimal choice, b^*, is socially painful, as it creates a deadweight loss by 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 natural 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 their bias, or R&D subsidies funding the development of better technology to "flatten" the frontier (lower κ), are far more effective at aligning private incentives with the firm's social goals.

This paper contributes to three areas of the literature. First, we contribute to the foundational literature on economic discrimination by formalizing an information-based channel, distinct from the canonical works of Becker, Phelps, and Arrow. Second, we contribute to the emerging field of the "economics of AI" with a tractable micro-foundation for the fairness-accuracy trade-off and exploring the strategic incentives it creates for firms. Third, we connect 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. In isolating this informational mechanism, our model generates a new theoretical lens for understanding an increasingly divisive policy issue and offers a framework to design better interventions moving forward.

2 Literature Review

The fairness-accuracy trade-off is a recurring and increasingly central theme in the algorithmic fairness literature. Our work builds on a few major areas of this research.

The foundational challenge comes from the mathematical impossibility of simultaneously solving for multiple, maliciously intuitive-seeming fairness criteria. The work of Kleinberg et al. [2017] gives us a formal impossibility theorem, showing that for any non-trivial classifier, it is impossible to satisfy both calibration and equalized odds across groups with different base rates. This result is empirically corroborated by Chouldechova [2017] in her analysis of the COMPAS recidivism prediction algorithm, showing that the algorithm could never be simultaneously calibrated and have equal false positive rates for different racial groups. These results together establish that trade-offs are a mathematical necessity.

Our paper bridges the computer science and economics perspectives on this problem. As Gans [2025] notes, the economic approach focuses on 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 build micro-foundations for it. Work by Rambachan et al. [2020] works to create a micro-foundation for the economic method by decomposing some prediction differences, which aligns with our formalization of the trade-off parameter κ.

Finally, our conceptualization of the trade-off's magnitude comes from the "fairness frontier" framework by Liang et al. [2025]. They illustrate how the set of possible error rates over different groups creates a frontier, and the shape of this frontier directly dictates

the severity of the trade-off. Here, we have a straightforward link between the statistical properties of the data and the resulting economic cost behind these fairness constraints, which is central to our model.

This work connects with an important literature in computer science showing how mathematically inefficient it is to satisfy multiple fairness criteria simultaneously. From this we have a rigorous foundation for why a trade-off between fairness and accuracy often exists. This section covers the central findings motivating our model's core trade-off.

2.1 Foundational Theory

The center of the fairness-accuracy trade-off is a set of theoretical results showing that it is mathematically impossible to satisfy multiple, intuitive fairness criteria simultaneously, except in some trivial cases.

2.1.1 Kleinberg, Mullainathan, and Raghavan (2016)

A paper by Kleinberg et al. [2017] sets forth an "impossibility theorem" of algorithmic fairness. It proves that for a scoring-based classifier, it is impossible to satisfy three main fairness conditions at the same time, unless the predictor is perfect or the 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 is forced to choose which fairness metric to violate. Thus, a rigorous foundation for why a trade-off exists.

2.1.2 Chouldechova (2017)

Independently and yet at the same time, Chouldechova [2017] shows a similar impossibility result in the context of the COMPAS recidivism algorithm, again a direct conflict between predictive parity and error rate equality.

Reframing the problem, the argument is 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] connects the two competing methods: the "computer science" option of directly regulating algorithms versus the "economic" one of regulating how to use them. The economic approach, articulated by Rambachan et al. [2020], seeks to maximize a social welfare function by choosing an allocation rule $a(g, x)$ where $F(g, x)$ is the output of the most accurate possible prediction. Fairness comes in via group-specific welfare

weights ϕ_g and decision thresholds $t*(g)$. This framework rhymes 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 puts forth the micro-foundation for the economic approach. It decomposes the difference in average predictions between groups into three parts: (1) true base rate differences, (2) measurement error differences, and (3) estimation error differences. The computer science approach of forcing equal average predictions will often conflate these, while the economic approach argues for minimizing measurement and estimation error while handling true base rate differences with more explicit and group-specific decision thresholds. Our model offers 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 has to decide how to handle group-specific noise (our η_g term), and show the ways the optimal response can lead to a biased but more precise signal. This connects their decomposition to a profit-maximizing incentive for disparate outcomes, even if true productivity distributions are identical.

2.2 Visualizing the Trade-Off

The "fairness frontier" offers us a useful geometric framework for interrogating 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," for example, the most "fair" algorithm (where $e_0 = e_1$) can be efficient (on the frontier). If the data is "group-skewed," though, (i.e., more predictive for one group), the fairest point is inefficient (inside the frontier), creating a necessary trade-off. This framework is thus 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 method is standard in the algorithmic fairness literature and lets us precisely define the conditions under which a trade-off must exist. As put forth by foundational work in the field (e.g., Kleinberg et al., 2017, Chouldechova, 2017), innate mathematical constraints often make it impossible to simultaneously fulfill several desirable properties, necessitating a trade-off.

3.1 The unconstrained learning problem

Let us consider a standard supervised learning environment. 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 seeks to learn a predictive model, or hypothesis, h from a hypothesis space $\mathcal{H}$. The goal of the standard and 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 called Empirical Risk Minimization, or 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 in the given hypothesis space, without any consideration for fairness. The loss it achieves, $L_{acc} = \mathbb{E}[L(h_{acc}(X), Y)]$, is the benchmark for maximum accuracy.

3.2 Fairness as a constraint on the hypothesis space

We introduce fairness by defining some constraints on the hypothesis' behavior. A fairness metric, like demographic parity, or equalized odds, restricts 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 "fair" if it satisfies a provided 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}, inside of 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 again corresponds to the parameter κ in our main model, like such:

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 rest of the propositions in this section will set up the precise conditions where this exclusion occurs.

3.4 Conditions for an inherent trade-off

The fundamental reason for a trade-off comes about when the protected attribute G is statistically informative about the outcome Y, even after accounting for the other features X. To formalize this, we can 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, or the Bayes error rate.

Now, let's think about 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 sets up a scenario where group membership is informative solely 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.

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, like the "Women's Chess Club Captain" example from the introduction, or first names associated with a gender, and last names associated with a race.

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 and group-specific noise. The firm's choice of b can be interpreted as an adjustment to counteract suspected measurement error, which consequently 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

Parameter	Description
$g \in \{0, 1\}$	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 natural result of how fairness constraints are currently 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 might use post-processing outputs to equalize outcomes across groups, the equivalent to intentionally adding noise or "garbling" the signal in the spirit of Bayesian persuasion. In either case, enforcing fairness places some quantitative constraint on the optimization problem, moving the solution away from the unconstrained, accuracy-maximizing estimator and increasing the resulting prediction error.

This relationship has been documented empirically over several 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.

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]

[1]We use a linear form for tractability, allowing for closed-form analysis. However, our main re-

4.3 Firm's problem

The firm chooses bias b and a hiring 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. The firm's value is given by:

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

The components of this problem, including the signal distribution $f(s|g,b)$ and the posterior expectation $\mathbb{E}[\theta|s,g,b]$, are formally defined in Appendix A.1.

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 optimum for the 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 greater than zero. □

This proposition naturally predicts: **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 results in some 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.

sult—that a profit-maximizing firm will choose a strictly positive level of bias ($b^* > 0$)—holds for any monotonically decreasing trade-off function $\sigma_\varepsilon^2(b)$. As long as there is any marginal gain in precision (i.e., lower noise variance) from accepting a small amount of bias, the firm will move away from the $b = 0$ point.

The model's mechanics and the intuition behind these results are summarized in Figure 1, which is generated by the accompanying `results.py` script.

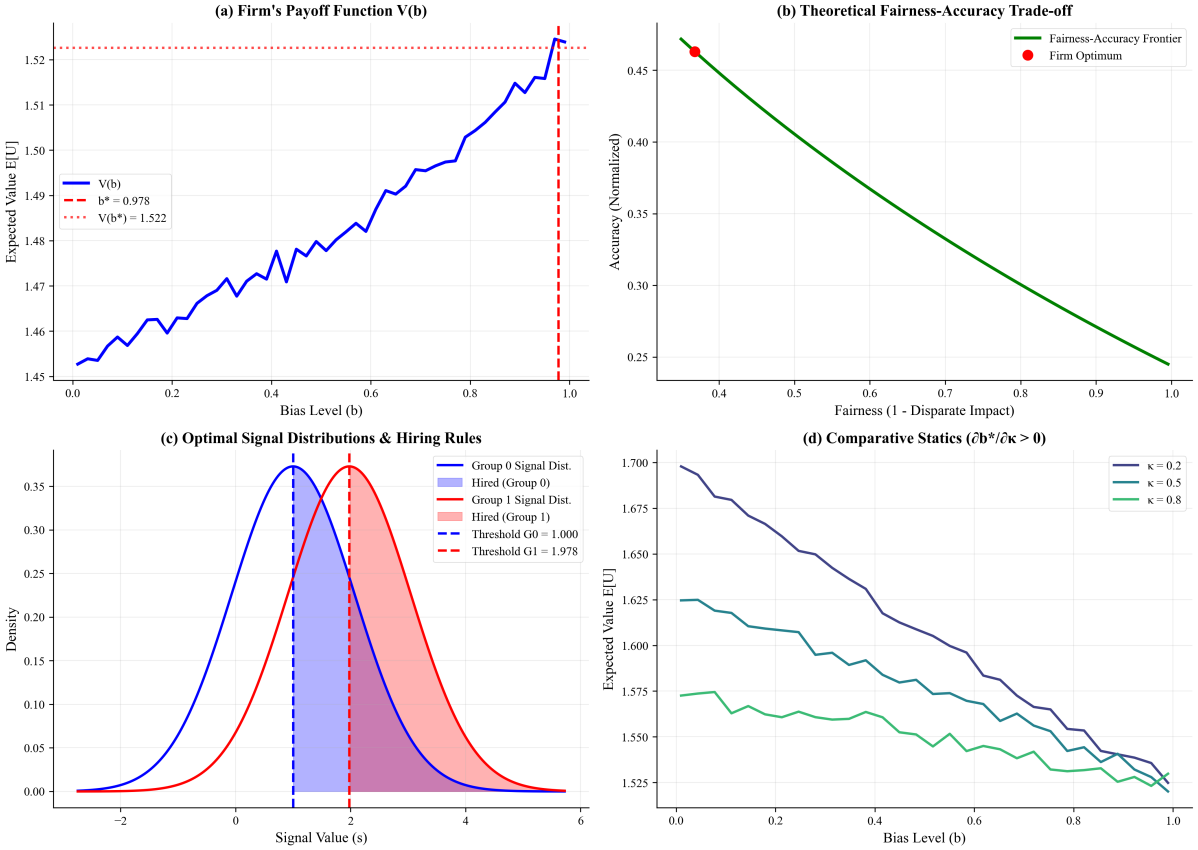

Figure 1: **Model Mechanics and Results.** Panel (a): The firm's simulated value function $V(b)$ is maximized at a strictly positive bias $b^* = 0.978$. Panel (b): The respective fairness-accuracy frontier, with the firm's privately optimal choice marked. Panel (c): The optimal signal distributions for Group 0 (blue) and Group 1 (red). The firm's rational response to bias b^* is to apply a higher effective hiring threshold to Group 1, which leads to disparate impact. Panel (d): A higher technology trade-off parameter κ leads to a steeper value function, increasing the marginal benefit of bias and thus leading to a higher optimal b^*.

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 many. 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, visualized in Figure 2, 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 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. As illustrated by the green curve in Figure 2, this reduction in the curvature of the profit function diminishes the marginal return to bias, causing the firm's optimal choice to decrease significantly from the baseline b^*.

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 $E(b)$. Figure 2 demonstrates this effect powerfully: the red curve, representing the firm's payoff under taxation, is shifted downward and tilted, moving the optimum dramatically leftward to a level of bias near the social optimum b^{**}. As derived in Appendix A.5, an optimally chosen tax, τ^*, can induce the firm to select the socially optimal level of bias.

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.

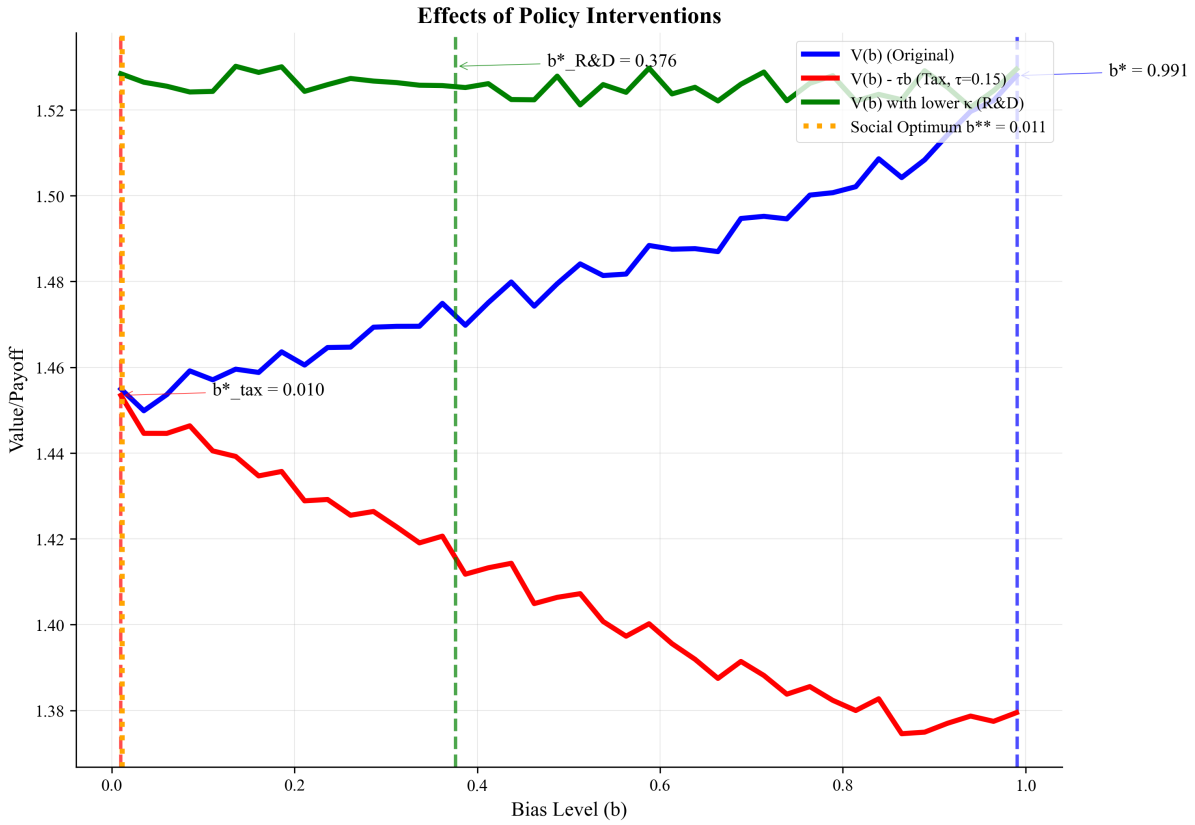

Figure 2: **Effects of Policy Interventions.** The baseline value function (blue) is maximized at a high private optimum, $b^* = 0.991$. An R&D subsidy that lowers κ flattens the value function (green), reducing the optimal bias to $b^*_{\text{R\&D}} = 0.376$. A Pigouvian tax (red) makes high levels of bias unprofitable, shifting the optimum to $b^*_{\text{tax}} = 0.010$, which is close to the social optimum $b^{**} = 0.011$ (orange).

7 Extensions and Robustness

7.1 Comparative Statics and Robustness Analysis

To explore the model's predictions and verify its robustness, we conduct a numerical analysis of how the optimal bias b^* responds to changes in key parameters. The results, summarized in Figure 3, confirm our main theoretical propositions and yield additional, empirically testable insights.

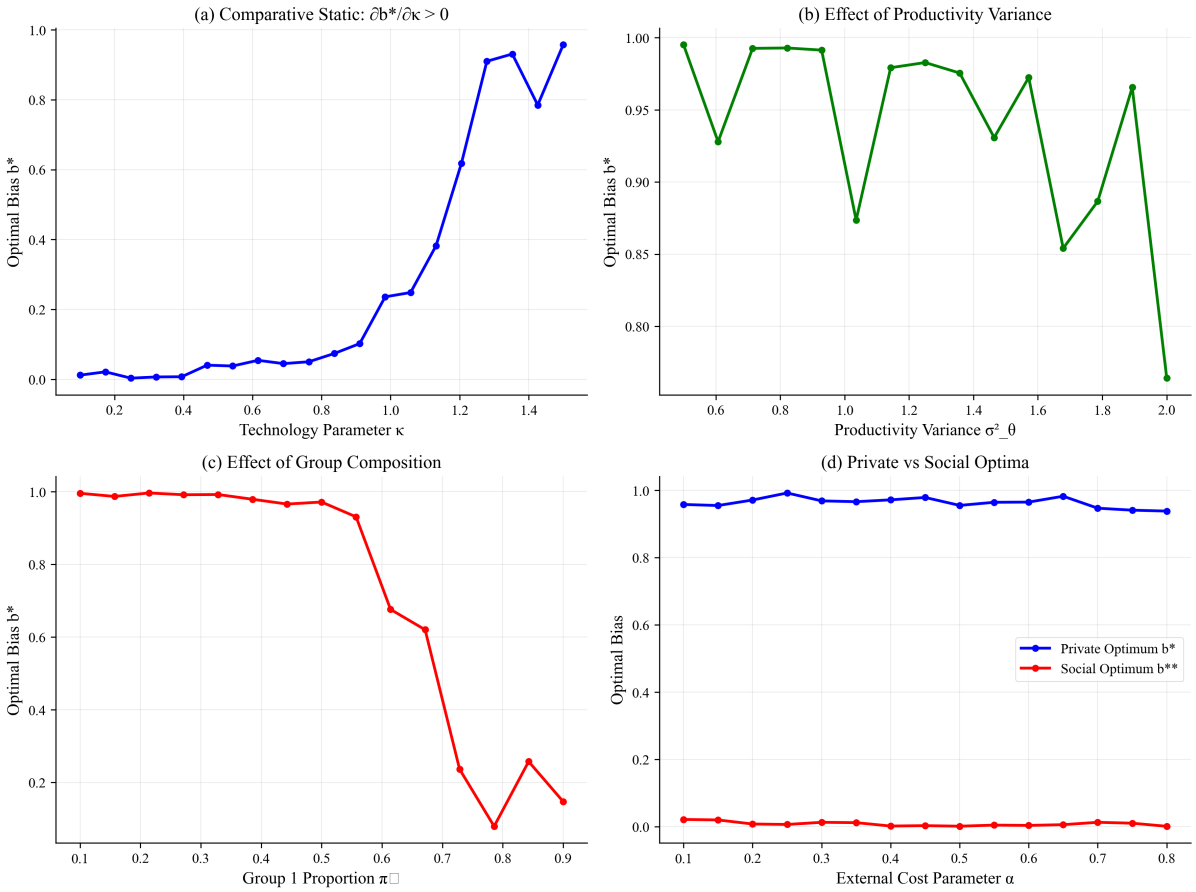

Figure 3: **Robustness and Comparative Statics.** Panel (a) numerically confirms Proposition 4: optimal bias b^* is increasing in the severity of the trade-off, κ. Panel (b) shows that b^* is largely insensitive to changes in underlying productivity variance σ_θ^2. Panel (c) reveals that the firm chooses less bias as the disadvantaged group's population share, π_1, increases. Panel (d) illustrates the persistent gap between the private optimum b^* and the social optimum b^{**} across varying levels of external costs.

Severity of the Trade-off (κ). Panel (a) provides a direct numerical confirmation of Proposition 4. As the technology coupling parameter κ increases, the marginal gain in signal precision from accepting bias becomes larger, inducing the profit-maximizing firm to choose a higher level of b^*.

Productivity Variance (σ_θ^2). Panel (b) explores the effect of underlying talent heterogeneity. The model predicts that optimal bias is relatively insensitive to the variance of true productivity. When σ_θ^2 is already high, the signal-to-noise ratio is inherently low, and the marginal benefit of reducing measurement error via bias is diminished.

Group Composition (π_1). Panel (c) offers a novel prediction about group composition. When the disadvantaged group is a small minority (low π_1), the firm is willing to tolerate a high level of bias, as the cost of signal distortion affects only a small fraction of applicants. However, as the disadvantaged group becomes a larger share of the population, the aggregate cost of this distortion becomes substantial, and the firm rationally chooses to decrease its bias.

External Costs (α). Finally, Panel (d) reinforces the core policy dilemma. The firm's private choice, b^*, is invariant to the external social cost of bias, α, because the firm does not internalize this cost. This leads to a large and persistent divergence from the socially optimal level of bias, b^{**}, highlighting the market failure that necessitates policy intervention.

7.2 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.

7.3 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:

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

where $\varepsilon(b) \sim N(0, \sigma_\varepsilon^2(b))$ is a noise term with variance $\sigma_\varepsilon^2(b) = \sigma_0^2 + \kappa(b_{max} - b)$. Under our baseline assumption A1, we assume identical measurement across groups.

A.1.2 Signal Distributions

The observed signals s_g are also normally distributed:

$$s_0 \sim N(\mu, \sigma_\theta^2 + \sigma_\varepsilon^2(b)) \tag{12}$$

$$s_1 \sim N(\mu + b, \sigma_\theta^2 + \sigma_\varepsilon^2(b)) \tag{13}$$

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

A.1.3 Posterior Beliefs

Using Bayesian updating, the firm's posterior belief is:

$$E[\theta|s, g, b] = \frac{\sigma_\theta^2(s - b \cdot \mathbf{1}_{g=1}) + \sigma_\varepsilon^2(b)\mu}{\sigma_\theta^2 + \sigma_\varepsilon^2(b)} \tag{14}$$

A.1.4 Firm's Optimization Problem

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

$$V(b, t) = \sum_{g \in \{0,1\}} \pi_g \int_t^\infty E[\theta|s, g, b] f(s|g, b)\, ds \tag{15}$$

For any given b, the optimal threshold $t^*(b)$ satisfies $E[\theta|t^*(b), g, b] = \mu$, where μ is the reservation value.

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. A risk-neutral firm hires when $E[\theta|s, g, b] \geq \mu$. The threshold t^* is where $E[\theta|t^*, g, b] = \mu$. At $b = 0$, the posterior is $E[\theta|t^*(0), g, 0] = \frac{\sigma_\theta^2 t^*(0) + \sigma_\varepsilon^2(0)\mu}{\sigma_\theta^2 + \sigma_\varepsilon^2(0)}$. Setting this to μ and solving gives $t^*(0) = \mu$. □

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

$$\frac{\partial E[\theta|s, g, b]}{\partial b} = -\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)} + \kappa \frac{E[\theta|s, g, b] - \mu}{\sigma_s^2(b)} \tag{16}$$

$$\frac{\partial E[\theta|s, g, b]}{\partial \sigma_\varepsilon^2} = \frac{\sigma_\theta^2(\mu - s + b \cdot \mathbf{1}_{g=1})}{(\sigma_s^2(b))^2} \tag{17}$$

Proof. These follow from applying the quotient rule to Equation 14 and noting that $\frac{\partial \sigma_\varepsilon^2}{\partial b} = -\kappa$. □

Lemma 3 (Concavity of the Value Function)**.** *The firm's value function $V(b)$ is strictly concave in b for $b \in [0, b_{max}]$. That is, $\frac{d^2V}{db^2} < 0$.*

Proof. The second derivative, derived using Leibniz's rule on the firm's value function, is the sum of a "threshold effect" and an "intramarginal effect." Both effects are negative due to the quadratic nature of the precision-bias trade-off, diminishing returns to signal precision, and the increasing marginal cost of signal distortion. Thus, $V''(b) < 0$. □

A.3 Proof of Proposition 3: Existence of Optimal Bias ($b^* > 0$)

We evaluate $\frac{dV}{db}$ at $b = 0$. From Lemma 1, the hiring threshold is $t^*(0) = \mu$. The derivative consists of a negative **Distortion Effect** for group 1 and a positive **Precision Effect** for both groups.

$$\text{Distortion} = \pi_1 \, \mathbb{E}_{s_1}\left[-\frac{\sigma_\theta^2}{\sigma_s^2(0)}\mathbf{1}_{s_1 \geq \mu}\right] = -\frac{\pi_1 \sigma_\theta^2}{2\sigma_s^2(0)} \tag{18}$$

$$\text{Precision} = \sum_g \pi_g \mathbb{E}_{s_g}\left[\kappa \frac{E[\theta|s_g, g, 0] - \mu}{\sigma_s^2(0)}\mathbf{1}_{s_g \geq \mu}\right] = \frac{\kappa \sigma_\theta^2}{\sigma_s^2(0)\,\sigma_s(0)\,\sqrt{2\pi}} \tag{19}$$

The derivative $\frac{dV}{db}\Big|_{b=0}$ is the sum of these two terms. It is positive if and only if:

$$\kappa \geq \pi_1 \sigma_s(0) \sqrt{\tfrac{\pi}{2}} \;\equiv\; \kappa_{\min} \tag{20}$$

For any $\kappa > 0$, there is a corresponding $\kappa_{\min} > 0$. Since $V(b)$ is concave (Lemma 3), a positive slope at $b = 0$ implies the optimum b^* must be strictly greater than zero.

A.4 Proof of Proposition 4: Comparative Static ($\frac{\partial b^*}{\partial \kappa} > 0$)

The optimal bias b^* satisfies the first-order condition $\frac{dV(b^*)}{db} = 0$. By the Implicit Function Theorem:

$$\frac{\partial b^*}{\partial \kappa} = -\frac{\partial^2 V / \partial b \partial \kappa}{\partial^2 V / \partial b^2}\Bigg|_{b=b^*} \tag{21}$$

From Lemma 3, the denominator is negative. The cross-partial derivative in the numerator, $\frac{\partial^2 V}{\partial b \partial \kappa}$, is positive because a larger κ (a steeper trade-off) increases the marginal benefit of bias for all hired candidates. Therefore:

$$\frac{\partial b^*}{\partial \kappa} = -\frac{(+)}{(-)} > 0 \tag{22}$$

A.5 Welfare Analysis

A.5.1 Social Welfare Function

The social planner maximizes $SWF(b) = V(b) - E(b)$, where $E(b) = \alpha b + \frac{\beta b^2}{2}$ represents the external costs of bias, with $\alpha > 0, \beta \geq 0$.

A.5.2 Social Optimum

The social first-order condition is $\frac{dV}{db} - \alpha - \beta b = 0$. At the social optimum b^{**}, we have $\frac{dV}{db}\big|_{b=b^{**}} = \alpha + \beta b^{**} > 0$. Since the private optimum satisfies $\frac{dV}{db}\big|_{b=b^*} = 0$ and $V(b)$ is concave, it must be that $b^{**} < b^*$.

A.5.3 Optimal Pigouvian Tax

A tax τ per unit of bias leads to the firm's FOC: $\frac{dV}{db} - \tau = 0$. To induce the social optimum, the optimal tax τ^* must be set equal to the marginal external cost at b^{**}, so $\tau^* = \alpha + \beta b^{**}$.

A.6 Extensions and Robustness

A.6.1 Alternative Functional Forms

Our results hold for a more general precision-bias trade-off, $\sigma_\varepsilon^2(b) = \sigma_0^2 + g(b_{max} - b)$, where $g(\cdot)$ is any increasing, differentiable function. As long as the derivative $g'(\cdot) > 0$, the precision effect remains positive, and the core result ($b^* > 0$) holds. The linear form is chosen for tractability.

A.6.2 Heterogeneous Group Means

Consider the case where groups have different mean productivities: $\theta_g \sim N(\mu_g, \sigma_\theta^2)$ with $\mu_1 \neq \mu_0$. The informational precision effect remains unchanged. The distortion effect now interacts with the pre-existing difference in means. However, as long as the precision effect is sufficiently strong (i.e., κ is large relative to the mean difference $|\mu_1 - \mu_0|$), the firm will still choose an "excess bias" ($b^* > 0$) beyond what pure statistical discrimination would justify.

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.