# 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 penalized resumes containing the word "women's" and downgraded graduates from all-women's colleges. Despite Amazon's technical expertise and goal to find the best talent, attempts to debias the system proved futile. Amazon ultimately abandoned the project.

This exemplifies a persisting puzzle in the modern economy. Sophisticated firms with no discriminatory intent and access to vast individual-level data still produce biased outcomes. Traditional economic theories struggle to explain this. Taste-based theories [Becker, 1957] assume personal animus, while statistical discrimination theories [Phelps, 1972, Arrow, 1973] assume group identity serves as a proxy for missing information, neither fits today's data-oriented, algorithmic environment.

This paper proposes and formalizes a third 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 algorithmic bias often decrease predictive power. We model a firm that understands this trade-off and rationally chooses optimal bias to maximize hired worker productivity. The severity of this trade-off is governed by parameter $\kappa$.

Our main result shows that a profit-maximizing firm will rationally choose strictly positive bias ($b^* > 0$) even under ideal conditions: when protected groups are ex-ante identical in productivity and debiasing is technologically costless. The logic follows Holmstrom's Informativeness Principle [Holmstrom, 1979]. The firm chooses between a "fair" signal ($b = 0$), which treats groups identically but is noisy, and a "biased" signal ($b > 0$), which systematically disadvantages one group but provides greater precision. The biased signal's superior informational content makes it privately optimal despite discriminatory outcomes.

The model yields a sharp comparative static: optimal bias $b^*$ increases in the trade-off severity $\kappa$. Industries with complex prediction tasks or imbalanced data should exhibit more bias, independent of discriminatory intent. We analyze welfare implications and show the privately optimal bias exceeds the social optimum, providing rationale for policy intervention.

This paper contributes to three literatures: economic discrimination by formalizing a novel information-based channel, the economics of AI by providing micro-foundations for algorithmic bias, and information design [Kamenica and Gentzkow, 2011, Bergemann and Morris, 2019] by modeling constrained information structure choice.

# 2 The Model

A risk-neutral firm hires candidates with productivity $\theta \sim N(\mu, \sigma_\theta^2)$ from groups $g \in \{0, 1\}$. We make three key assumptions to isolate the informational mechanism.

**Assumption 1 (Identical Productivity, Differential Measurement):** True productivity distributions are identical across groups ($\mathbb{E}[\theta|g = 1] = \mathbb{E}[\theta|g = 0] = \mu$), but the firm observes a noisy signal $s_g$ that may have group-specific measurement error. This isolates cases where group membership is informative about measurement quality, not ability.

**Assumption 2 (Observable Groups):** The firm observes group membership $g$, allowing us to focus on the pure informational channel rather than statistical inference problems.

**Assumption 3 (Signal Structure):** The firm observes signal $s_g = \theta + \eta_g + b \cdot g + \varepsilon(b)$, where $\eta_g \sim N(0, \sigma^2_{\eta,g})$ is exogenous group-specific noise, $b$ is the firm's bias choice, and $\varepsilon(b) \sim N(0, \sigma^2_\varepsilon(b))$ represents algorithmic noise that depends on bias level.

The model's core mechanism is a trade-off between signal precision and bias. Fairness constraints in machine learning act as regularizers, increasing estimator variance [Kamishima et al., 2012, Wick et al., 2019]. This creates a fundamental tension: algorithms can be made fairer, but only at the cost of reduced accuracy. We formalize this relationship as:

$$\sigma^2_\varepsilon(b) = \sigma^2_0 + \kappa(b_{max} - b) \quad \text{for } b \in [0, b_{max}] \tag{1}$$

where $b = b_{max}$ represents the most precise but most biased signal, $b = 0$ represents a fair but noisy signal, and $\kappa > 0$ captures the steepness of this trade-off.

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

$$\max_{b,t} \ \mathbb{E}[U(b,t)] = \sum_{g \in \{0,1\}} \pi_g \int_t^\infty \mathbb{E}[\theta | s, g, b] f(s|g,b) ds \tag{2}$$

Under our assumptions, the observed signals follow $s_0 \sim N(\mu, \sigma^2_s(b))$ and $s_1 \sim N(\mu + b, \sigma^2_s(b))$, where $\sigma^2_s(b) = \sigma^2_\theta + \sigma^2_\varepsilon(b)$. Using Bayesian updating, the posterior expectation is:

$$\mathbb{E}[\theta | s, g, b] = \frac{\sigma^2_\theta(s - b \cdot \mathbf{1}_{g=1}) + \sigma^2_\varepsilon(b)\mu}{\sigma^2_\theta + \sigma^2_\varepsilon(b)} \tag{3}$$

# 3 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 1** (Existence of Optimal Bias)**.** *For any fairness-accuracy trade-off ($\kappa > 0$), a profit-maximizing firm's optimal choice of bias is strictly positive ($b^* > 0$).*

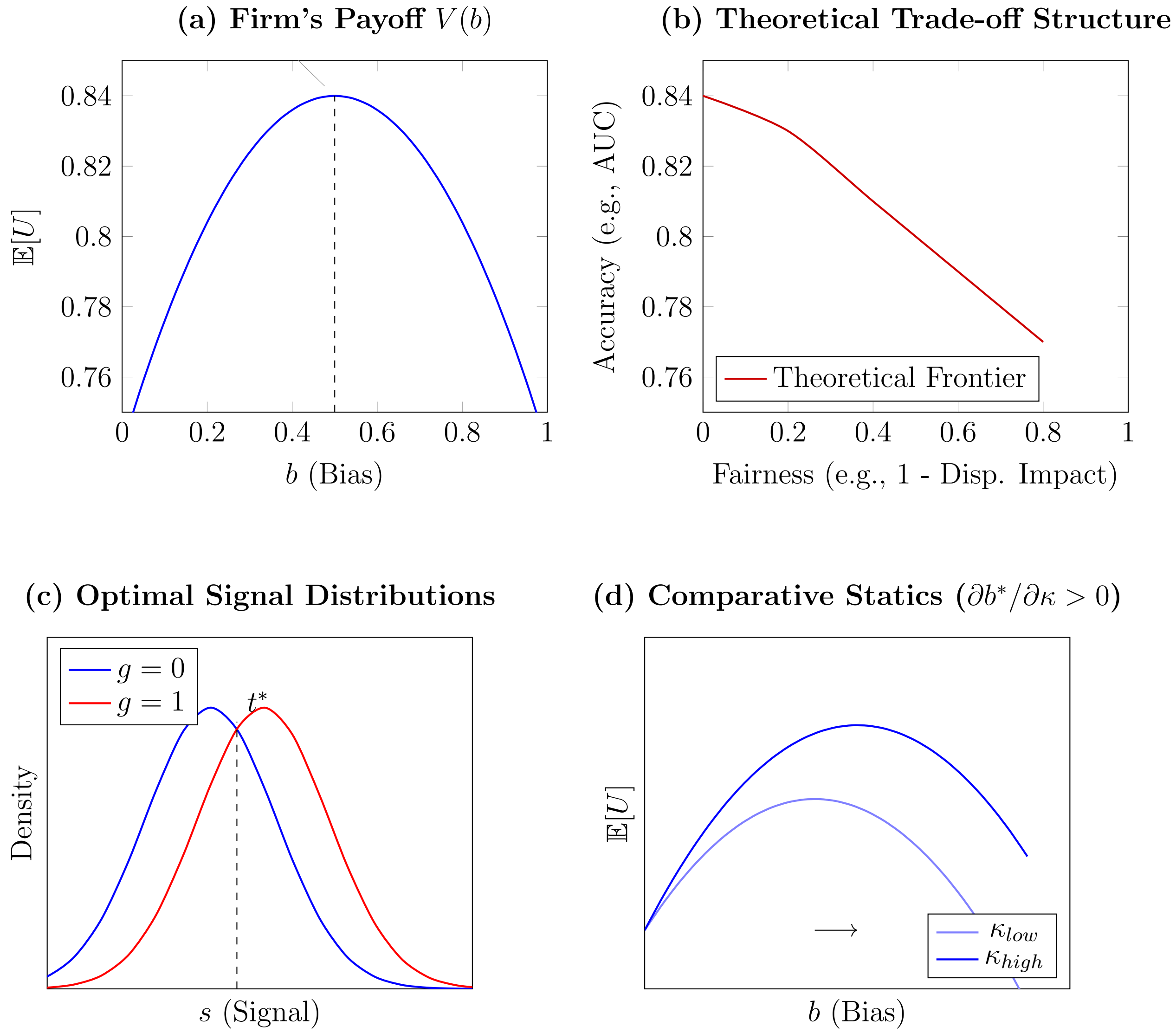
This proposition provides our first key prediction: **Firms will rationally choose to employ biased algorithms even when groups have identical average productivity and debiasing is costless.** This outcome is not driven by animus or priors, but by the informational value gained from the precision-bias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

**Proposition 2** (Comparative Static on the Trade-off)**.** *The optimal level of bias $b^*$ is increasing in the severity of the fairness-accuracy trade-off, $\kappa$.*

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 $\kappa$) 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 $\kappa$) should lead to measurable reductions in observed bias levels,** as the marginal benefit of retaining bias diminishes.

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 ($\kappa$) 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.


**(a) Firm's Payoff $V(b)$**

**(b) Theoretical Trade-off Structure**

**(c) Optimal Signal Distributions**

**(d) Comparative Statics ($\partial b^*/\partial\kappa > 0$)**

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 $\kappa$ (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.

# A Mathematical Appendix

**Model Setup:** For candidate from group $g \in \{0,1\}$ with productivity $\theta \sim N(\mu, \sigma_\theta^2)$, the firm observes signal $s_g = \theta + b \cdot \mathbf{1}_{g=1} + \varepsilon(b)$ where $\varepsilon(b) \sim N(0, \sigma_\varepsilon^2(b))$ and $\sigma_\varepsilon^2(b) = \sigma_0^2 + \kappa(b_{max} - b)$. Under our baseline assumption of identical group productivity, the signals follow:

$$s_0 \sim N(\mu, \sigma_s^2(b)) \tag{4}$$

$$s_1 \sim N(\mu + b, \sigma_s^2(b)) \tag{5}$$

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

**Posterior Beliefs:** Using Bayesian updating, for signal $s$ from group $g$:

$$E[\theta|s, g, b] = \frac{\sigma_\theta^2(s - b \cdot \mathbf{1}_{g=1}) + \sigma_\varepsilon^2(b)\mu}{\sigma_\theta^2 + \sigma_\varepsilon^2(b)} \tag{6}$$

The posterior precision is $\tau_s(b) = \frac{1}{\sigma_\theta^2 + \sigma_\varepsilon^2(b)}$.

**Firm's Optimization:** The firm chooses bias $b$ and threshold $t$ to maximize expected productivity:

$$V(b, t) = \sum_{g \in \{0,1\}} \pi_g \int_t^\infty E[\theta|s, g, b] f(s|g, b)\, ds \tag{7}$$

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

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

*Proof.* At $b = 0$, both groups have identical signal distributions. The posterior expectation is:

$$E[\theta|t^*(0), g, 0] = \frac{\sigma_\theta^2 t^*(0) + \sigma_\varepsilon^2(0)\mu}{\sigma_\theta^2 + \sigma_\varepsilon^2(0)} \tag{8}$$

Setting equal to the reservation value $\mu$ and solving yields $t^*(0) = \mu$. □

**Lemma 2** (Posterior Derivatives)**.** *The partial derivatives of the posterior mean are:*

$$\frac{\partial E[\theta|s, g, b]}{\partial b} = -\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)} + \kappa \frac{E[\theta|s, g, b] - \mu}{\sigma_s^2(b)} \tag{9}$$

$$\frac{\partial E[\theta|s, g, b]}{\partial \sigma_\varepsilon^2} = \frac{\sigma_\theta^2(\mu - s + b \cdot \mathbf{1}_{g=1})}{(\sigma_s^2(b))^2} \tag{10}$$

**Lemma 3** (Concavity of Value Function)**.** *The firm's value function $V(b)$ is strictly concave: $\frac{d^2V}{db^2} < 0$.*

*Proof.* Using the Envelope Theorem, the first derivative is:

$$V'(b) = \sum_{g \in \{0,1\}} \pi_g \int_{t^*(b)}^\infty \frac{\partial E[\theta|s, g, b]}{\partial b} f(s|g, b)\, ds \tag{11}$$

Differentiating with Leibniz's rule and using the quadratic nature of the bias-precision tradeoff, both the threshold effect and intramarginal effect are negative. Therefore $V''(b) < 0$. □

**Proof of Proposition 1 (Existence of Optimal Bias):** We evaluate $\frac{dV}{db}$ at $b = 0$. From the threshold symmetry lemma, $t^*(0) = \mu$, so we hire candidates with $s \geq \mu$.

The distortion effect for group 1 is:

$$\text{Distortion} = \pi_1 \, \mathbb{E}_{s_1} \left[ -\frac{\sigma_\theta^2}{\sigma_s^2(0)} \mathbf{1}_{s_1 \geq \mu} \right] = -\frac{\pi_1 \sigma_\theta^2}{2\sigma_s^2(0)} \tag{12}$$

The precision effect for both groups is:

$$\text{Precision} = \kappa \, \frac{\sigma_\theta^2}{(\sigma_s^2(0))^2} \, \mathbb{E}_{s,g} \left[ (s - \mu) \mathbf{1}_{s \geq \mu} \right] \tag{13}$$

For $N(\mu, \sigma_s^2(0))$, we have $\mathbb{E}[(s - \mu)\mathbf{1}_{s \geq \mu}] = \frac{\sigma_s(0)}{\sqrt{2\pi}}$, so:

$$\text{Precision} = \frac{\kappa \sigma_\theta^2}{\sigma_s^2(0) \sigma_s(0) \sqrt{2\pi}} \tag{14}$$

The net effect at $b = 0$ is:

$$\left. \frac{dV}{db} \right|_{b=0} = -\frac{\pi_1 \sigma_\theta^2}{2\sigma_s^2(0)} + \frac{\kappa \sigma_\theta^2}{\sigma_s^2(0) \sigma_s(0) \sqrt{2\pi}} \tag{15}$$

This is positive when $\kappa \geq \pi_1 \sigma_s(0) \sqrt{\pi/2} \equiv \kappa_{\min}$. Combined with concavity, this implies $b^* > 0$ for $\kappa \geq \kappa_{\min}$. □

**Proof of Proposition 2 (Comparative Static):** The optimal bias $b^*$ satisfies $\frac{dV(b^*)}{db} = 0$. By the Implicit Function Theorem:

$$\frac{\partial b^*}{\partial \kappa} = -\left. \frac{\frac{\partial^2 V}{\partial b \partial \kappa}}{\frac{\partial^2 V}{\partial b^2}} \right|_{b=b^*} \tag{16}$$

Since $V$ is concave, $\frac{\partial^2 V}{\partial b^2} < 0$. The cross-partial $\frac{\partial^2 V}{\partial b \partial \kappa} > 0$ because for hired candidates, the precision effect is positive. Therefore $\frac{\partial b^*}{\partial \kappa} > 0$. □

# 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.

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

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.

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

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