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 κ.

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 κ. 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. 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 main results with empirical implications for the persistence and variation of algorithmic bias. The analysis hinges on the firm's optimization problem, where the profit-maximizing value function $V(b)$ is strictly concave. This ensures a unique optimum for the 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 ($b^* > 0$).*

Proof. The proof in Appendix A.1 shows that at zero bias, the marginal value of increasing bias is strictly positive ($dV/db|_{b=0} > 0$). Because the function is increasing at $b = 0$, the optimum cannot be at zero. Furthermore, since the value function $V(b)$ is strictly concave (see Lemma 3), there exists a unique optimal bias b^* which must therefore be strictly positive. □

This proposition implies that firms may rationally choose to employ biased algorithms even when protected groups have identical average productivity and debiasing is technologically costless.

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, κ.*

Proof. See Appendix A.2. The proof uses the Implicit Function Theorem on the first-order condition to show that $\partial b^*/\partial \kappa > 0$. □

This result generates clear, testable predictions about how algorithmic bias should vary with the technological environment.

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, 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.

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

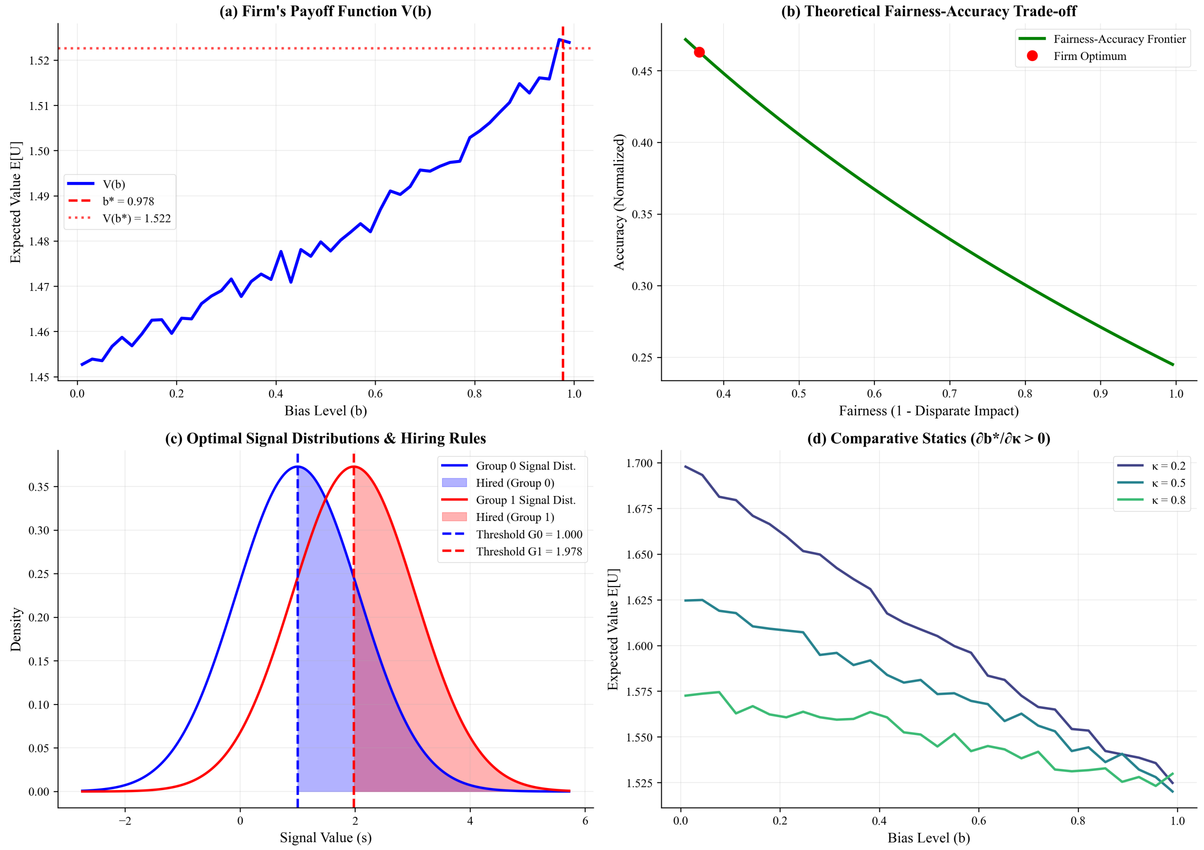

Figure 1: **Model mechanics and results.** Panel (a): The firm's simulated value function $V(b)$ is maximized at a strictly positive bias $b^* = 0.978$. Panel (b): The respective fairness-accuracy frontier, with the firm's privately optimal choice marked. Panel (c): The optimal signal distributions for 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^*.

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. *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:*

$$\frac{\partial E[\theta|s,g,b]}{\partial b} = -\frac{\sigma_\theta^2 \mathbf{1}_{g=1}}{\sigma_s^2(b)} + \kappa \frac{E[\theta|s,g,b] - \mu}{\sigma_s^2(b)} \tag{8}$$

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

Proof. These follow from applying the quotient rule to (6) 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.1 Proof of Proposition 1: 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.

$$\text{Distortion} = \pi_1 \, \mathbb{E}_{s_1} \left[-\frac{\sigma_\theta^2}{\sigma_s^2(0)} \mathbf{1}_{s_1 \geq \mu} \right] = -\frac{\pi_1 \sigma_\theta^2}{2\sigma_s^2(0)} \tag{10}$$

$$\text{Precision} = \sum_g \pi_g \mathbb{E}_{s_g} \left[\kappa \frac{E[\theta | s_g, g, 0] - \mu}{\sigma_s^2(0)} \mathbf{1}_{s_g \geq \mu} \right] = \frac{\kappa \sigma_\theta^2}{\sigma_s^2(0) \, \sigma_s(0) \, \sqrt{2\pi}} \tag{11}$$

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

$$\kappa \geq \pi_1 \sigma_s(0) \sqrt{\tfrac{\pi}{2}} \;\equiv\; \kappa_{\min} \tag{12}$$

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.2 Proof of Proposition 2: Comparative static ($\frac{\partial b^*}{\partial \kappa} > 0$)

The optimal bias b^* satisfies the first-order condition $\partial V(b^*)/\partial b = 0$. To analyze how b^* changes with κ, we use the Implicit Function Theorem. The conditions for the theorem are met: $V(b)$ is twice continuously differentiable in b and κ because it is constructed from integrals of Normal PDFs, which are smooth functions. Furthermore, the denominator in the resulting expression, $\partial^2 V/\partial b^2$, is strictly negative from Lemma 3, ensuring it is non-zero at the optimum. By the Implicit Function Theorem:

$$\frac{\partial b^*}{\partial \kappa} = -\frac{\partial^2 V/\partial b \partial \kappa}{\partial^2 V/\partial b^2}\bigg|_{b=b^*} \tag{13}$$

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:

$$\frac{\partial b^*}{\partial \kappa} = -\frac{(+)}{(-)} > 0 \tag{14}$$

References

Arrow, K. J. (1973). The theory of discrimination. In Ashenfelter, O. and Rees, A., editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

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

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

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.

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

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

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware learning through regularization. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 643–650.

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

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

Wick, M., Panda, S., and Tristan, J.-B. (2019). Unlocking fairness: a trade-off revisited. *arXiv preprint arXiv:1906.06653*.