Queen's Economics Department Working Paper No. 1510

Can Unbiased Predictive AI Amplify Bias?

Tanvir Ahmed Khan

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

7-2023

Can Unbiased Predictive AI Amplify Bias?

Tanvir Ahmed Khan *
PhD Candidate, Queen's University

July 24, 2023

Preliminary Draft

Abstract

Predictive AI is increasingly used to guide decisions on agents. I show that even a bias-neutral predictive AI can potentially *amplify* exogenous (human) bias in settings where the predictive AI represents a cost-adjusted precision gain to unbiased predictions, and the final judgments are made by biased human evaluators. In the absence of perfect and instantaneous belief updating, *expected victims of bias* become less likely to be *saved by randomness* under more precise predictions. An increase in aggregate discrimination is possible if this effect dominates. Not accounting for this mechanism may result in AI being unduly blamed for *creating* bias.

1 Introduction

Predictive algorithms are increasingly used in important decisions with economic and policy consequences. For example, lending institutions use machine learning models to assess creditworthiness. Criminal justice systems use *predictive recidivism* algorithms to predict recidivism risks and guide bail decisions on defendants. Police departments use *predictive policing* algorithms to predict crimes and guide the deployment of law enforcement personnel into neighborhoods.[1]

How can predictive algorithms exacerbate discrimination? Existing literature recognizes two primary channels - biased model design or biased training data (see Barocas and Selbst, 2016; Cowgill et al., 2020). In this paper, I show the existence of a third channel. Predictive algorithms often represent an improvement over status-quo prediction technology in terms of cost-adjusted precision gain (noise-reduction). I show that a bias-neutral precision gain of the unbiased prediction technology can potentially amplify exogenous

*I thank Robert Clark, Nahim Zahur, and Shota Ichihashi for guidance and suggestions. I am also thankful for excellent feedback from James G. MacKinnon, Nicholas Brown, Luke Rawling, Chi Danh Dao, and others in the IO Working Group at Queen's University. Any limitations of the paper are solely my own.

[1] i.e., COMPAS is a predictive-recidivism AI used in the criminal justice system of several US states. GOTHAM is a predictive-policing software that has been used by the police departments of several US and European cities.

bias originating in human evaluators. The human-origin bias can be ascribed to either taste-based discrimination (Becker, 1957) or inaccurate statistical discrimination (Bohren et al., 2022).

This counter-intuitive result arises from human-machine interactions. In almost all settings where predictive algorithms are used, they function as 'decision-aids' or 'prediction machines' and not 'decision-makers', with ample room for human judgment (see Agarwal et al., 2018, Kleinberg et al., 2018).[2] However, when addressing the question of how predictive algorithms can exacerbate discrimination in the *realized outcome*, the literature has mostly ignored human-machine interactions and focused on *biased predictions* originating in the algorithm. In this paper, I build a simple model of this human-machine interaction - how machine predictions map to decision outcomes. I then show that an unbiased prediction technology *can* potentially amplify exogenous human-origin bias and exacerbate discrimination as it gets more precise.

Not accounting for this channel may result in predictive AI being unduly blamed for *'creating'* bias. Consider the following as a motivating example. In many settings, what often follows the adoption of predictive AI is an empirical assessment of its impact on discrimination. Discrimination can increase either due to *bias creation by predictive AI* or *amplification of exogenous human bias by predictive AI*. Existing literature does not disentangle the two. However, this distinction is important as it naturally leads to different conclusions regarding potential remedies. For example, if discrimination increases under predictive AI due to *bias creation*, potential solutions may include de-biasing the underlying model and the training data. If, on the other hand, discrimination increases due to *amplification of exogenous bias*, potential solutions may include reducing the scope of human discretion. In addition, this misattribution increases the risk of society preemptively moving away from a potentially beneficial technology.

Consider a simple model of algorithm-assisted decisions. An unbiased prediction technology produces signals of agents' qualifications (e.g. creditworthiness), and biased human evaluators make the final allocation decisions. I conceptualize the adoption of predictive AI as a bias-neutral precision gain for the prediction technology. Abstracting away from machine bias simplifies the demonstration of what happens to discrimination due to bias-neutral precision gain only, which is the focus of this paper. The earlier motivating example clarified that the relevant setting of this paper deals with short time horizons. Hence, I restrict human bias to be non-adaptive to the switch to predictive AI. This entails relaxing the assumption of perfect and instantaneous belief updating by human evaluators. Several features commonly found in relevant settings justify this modeling choice. For example, the 'black box', 'trade secret' and 'race blind' nature of predictive AI and

[2]This is perhaps partly explained by society's distrust of AI as the final decision-maker in many settings. Others argue that AI has a comparative advantage in predictions while humans have a comparative advantage in judgment (Korinek, 2023).

machine learning models, limited ex-ante information (regarding noisiness of predictions and extent of precision gain from predictive AI), and little or no feedback. Literature has found that belief updating may be slow, or fail altogether when these features are present (see, Bohren et al., 2022).

In the model, there are agents (i.e. loan applicants) with two observable group identities (i.e. race). Qualification to receive a penalty (adversarial treatment, i.e. being denied a loan) is based on an unobserved qualifying variable (i.e. default probability) that nature draws from the same distribution for both groups. An assigning authority (i.e. financial institution) deploys noisy prediction technologies to map the unobserved qualifications to predicted qualifications - which are sums of 'signals' (true qualification) and 'noise' (mean-zero normally distributed prediction errors). Human evaluators observe both the group identities and the *predicted* qualifications of agents and add their human judgment to make final allocation decisions. The representative human evaluator is biased against agents of one of the two group identities. Thus, bias originates in this model only through human evaluators and is exogenous to the prediction technology. Allocation decision is modeled as a cutoff rule on scores, which is a sum of human judgment (bias parameter) and model-predicted qualifications - comprising of signals and noise. Due to the bias in human judgment, agents from the discriminated group face a stricter cutoff on the realized predictions.

I show that, as the prediction technology becomes more precise (less noisy), *expected victims of bias* become (weakly) less likely to be *saved by randomness*. *Expected victims of bias* are agents from the discriminated group that are expected to be assigned to the penalty based on their true qualification draws, but would not be expected to be assigned to the penalty if they were from the non-discriminated group. *Expected victims of bias* get *saved by randomness* if they get negative noise draws large enough in magnitude to take their realized predictions below the cutoff.

As a result, discrimination (conditional on qualification) is non-decreasing for the *expected victims of bias* as the prediction technology becomes more precise. This result holds irrespective of the shape of the qualification distributions and parameter values. This is the main result of this paper. If the aforementioned effect dominates, it is possible for aggregate discrimination to increase under a more precise prediction technology.

As far as the net effect of increasing precision on aggregate discrimination is concerned, there are other effects at play. In general, discrimination may increase or decrease depending on parameter values for agents from the discriminated group other than the *expected victims of bias*. Thus, the net effect of precision gain on aggregate discrimination may get canceled out. I find that the net effect cancels out if the discrimination-precision sensitivity function [3] is symmetric. When this function is not symmetric, aggregate discrimination

[3] Marginal effect of precision gain on discrimination.

may increase or decrease, and this effect is amplified if the prediction technology before AI adoption was noisy, if the size of the precision gain is large, and if the intensity of human bias is large. The symmetry of this function is sensitive to parameter values. In general, aggregate discrimination arising from exogenous human bias become more likely to increase with a bias-neutral precision gain if the relative mass of *expected victims of bias* is large and if the distribution of qualifications is centered near the cutoff. Whether this happens or not depends on parameter values and the shape of the qualification distribution.

In terms of relation to literature, there is a growing economics and computer science literature on algorithmic fairness as it has come to the forefront of policy debate. The prospect of increasing prediction accuracy explains the rapid adoption of prediction algorithms (see Agarwal et al., 2018, Brynjolfsson et al., 2018). Several empirical works claim to find evidence of bias being present in algorithm-assisted decisions in lending (Bartlett et al., 2022), criminal sentencing (Arnold et al., 2018), health (Obermeyer et al., 2019), and hiring (Datta et al., 2015). In contrast, other empirical and theoretical works have shown algorithms do not necessarily lead to more discrimination and can be welfare-improving and bias-reducing (see Kleinberg et al.,2018; Rambachan et al. 2020, Avery et al.2023). Overall, the question of whether predictive AI increases or decreases discrimination has been an empirical one.

The contribution of this paper is in showing the existence of a channel through which even unbiased predictive AI can exacerbate discrimination. My analysis also sheds light on conditions when discrimination can increase or decrease in response to a bias-neutral precision gain. The rest of the paper is organized as follows: Section 2 presents a simple model of algorithm-assisted decision-making, taking into account human-machine interactions. Section 3 presents closed-form results. Section 4 concludes.

2 Model

In this section, I present a simple stylized model of the typical human-machine interactions through which predictions on agents get mapped to realized outcomes. This applies to settings where predictive algorithms are used as decision aids with human evaluators making the final call. To succinctly demonstrate how an unbiased predictive AI can amplify exogenous bias, I shut off both bias in the data and bias in the algorithm and allow bias to exist in human evaluators who are exogenous to the prediction model. The model is general and applies to a wide range of settings (i.e. lending, health, hiring, criminal justice, insurance). Without loss of generality and for expositional simplicity, let the setting be in lending.

Loan applicants. Consider loan applicants (agents) indexed by i with observable group identities (i.e. race) $R_i \in \{\mathrm{W}, \mathrm{M}\}$ denoting whites and minorities. Loan denial $T_i \in \{0, 1\}$

(treatment) is based on unobserved default risk Y_i (latent variable).[4] Let Y_i be absolutely continuous with respect to the Lebesgue measure. Without loss of generality, let nature draws Y_i from the same underlying distribution for both groups with support in the $[0, 1]$ interval. Restriction of Y_i to $[0, 1]$ interval is to give Y_i a probability interpretation. Also, note the simplifying assumption of Y_i draws coming from the same distribution for both groups shuts off 'accurate statistical discrimination'.

Financial Institutions. A financial institution (assigning authority) uses a two-step procedure for loan allocation. In step one, it uses a prediction technology that maps unobserved Y_i to predictions denoted by $\hat{Y}_i$. In step two, evaluators (humans) employed by the assigning authority observe the predictions $\hat{Y}_i$ and add human judgment to decide whether to assign T_i (loan denial) to applicants or not. [5]

Prediction Technology. Prediction technology $\hat{Y}$ maps unobserved Y_i of applicants to predicted $\hat{Y}_i$. Predictions $\hat{Y}_i$ can be additively separated into signals (true default risk) Y_i, and noises ϵ_i drawn from a normal distribution $N(0, \sigma)$. Thus, the prediction technology is unbiased as $\mathbb{E}[\hat{Y}_i] = Y_i$.[6] In other words, discrimination in this model does not originate from the prediction technologies.

Impact of Adoption of Predictive AI. Adoption of predictive AI makes the prediction technology more precise (less noisy). Consider a precision parameter $\eta = \frac{1}{\sigma}$. σ goes down and η as a result of the adoption of predictive AI. The predictive AI is bias-neutral. [7]

Evaluators. Evaluators observe both group identities R_i and predicted default risks $\hat{Y}_i$ of applicants. Representative evaluator applies subjective human judgment to the predictions $\hat{Y}_i$ to map them to scores S_i. Score S_i, therefore, is the sum of machine predictions and human judgment. Bias μ_i, arising from subjective human judgments, gets added in the mapping from predictions $\hat{Y}_i$ to scores S_i. Without loss of generality, I focus on discrimination against minority applicants. Let, $(\mu_i|M) = \mu$, $\mu > 0$ and $(\mu_i|W) = 0$.[8] Assignment to treatment is conceptualized as a cutoff rule on the scores S_i: assign to treatment if $S_i > \lambda \in (0, 1)$. Evaluators know that the prediction technology is unbiased and that the predictive AI is bias-neutral. However, the evaluator does not have perfect knowledge of

[4]To illustrate My model, I focus on penalties or 'adversarial treatments', which are unfavorable to the agents (i.e. loan denial, bail denial, hiring rejection). Y_i is a measure of 'disqualification' in this setup. This is without loss of generality, as any 'favorable treatment' can be reverse-coded.

[5]Note, there may be 'pure' fintech lenders where algorithms may make allocation decisions with no humans in the loop. This model does not apply to those settings.

[6]Legal requirements often explicitly prohibit prediction algorithms to condition on race or proxies of race that are uncorrelated with the underlying qualifications. For example, credit scoring models in the US, see the Equal Credit Opportunity Act.

[7]This assumption allows isolating the effect on discrimination attributable to precision gain only.

[8]Note, $(\mu_i|)$ is a slight abuse of notation with a more formal expression being $(\mu_i|R_i = M)$ Also, while the subjective human judgment can contribute more than just the bias parameter, for this stylized model any contribution other than the bias is normalized to zero.

what goes on as inputs into the prediction algorithms. [9] The human-origin bias μ can arise either from taste-based discrimination (Becker, 1957) or from inaccurate statistical discrimination (Bohren et al., 2022). The associated microfoundations and the microfoundations for the bias parameter μ and the cutoff λ are discussed later in this section.

Simplified model. The setup can be captured succinctly by the following equation:

$$\text{Score, } S_i = \overbrace{\mu \cdot \mathbb{I}(R_i = Minority)}^{\text{human judgment}} + \overbrace{\underbrace{Y_i + \epsilon_i}_{\hat{Y}_i};\ \ \epsilon_i \sim N(0, \sigma^2)}^{\text{machine prediction}},$$

$$\text{Cutoff Rule, } T_i = \mathbb{I}(S_i > \lambda),$$

$$\text{Adoption of AI: } \sigma \downarrow (\eta = \frac{1}{\sigma} \uparrow). \tag{1}$$

As can be seen from the model in equation (1), the scores are assumed to be additively separable into human judgment (first term) and machine predictions (second composite term). The machine predictions are further additively separable into signals (truths) and noises (prediction errors). Note, this is a static, short-run model, in the sense that μ does not adjust to the adoption of predictive AI.

Several features prevent μ from adjusting instantaneously to the adoption of predictive AI in the short run. First, due to the 'black box' nature (lack of interpretability) of AI and machine-learning algorithms, evaluators cannot foresee how the predictions $\hat{Y}_i$ would change under predictive AI. Second, designs of most algorithmic prediction models are guarded as trade secrets. Therefore, it is not an unrealistic assumption to think of human evaluators as lacking perfect information about what variables go into the construction of the predictive algorithms.[10] Third, perfect anticipation of how $\hat{Y}_i$ would change after the adoption of predictive AI requires knowing σ (which is inherently unobservable as Y_i are unobservables) and the extent of precision gain achieved by the predictive AI. Fourth, evaluators may not necessarily know or believe a priori that the predictive AI is more precise, as model improvements often take place at the back-end and such changes may not be communicated or understood well enough by the evaluators at the front-end.[11] Finally,

[9] Due to the black box nature of prediction algorithms and the trade secret nature of commercial prediction models, such information is generally undisclosed.

[10] For example, FICO scores are widely used by financial institutions in the United States to assess creditworthiness. The exact design and construction of the model are guarded as a trade secret. FICO only discloses five broad components that go into the construction of the model and their weights for the representative individual.

[11] Consider a prediction model that switched from using logistic regression to XGBoost, which is an advanced machine learning algorithm. For human evaluators, there is 'front-end equivalence' as they were getting predictions before and they would be getting predictions after. Due to the lack of interpretability of machine learning models, evaluators may not be able to understand the implications and may not necessarily believe a priori the new algorithm to be more precise. See, Monahan et al. (2020) for an example, which found that about 5% criminal judges in Virginia reported relying primarily on risk scores, compared to 38% reporting

in most institutional settings of relevance, human evaluators make fast decisions in high volumes with the possibility of exhaustion and little scope to receive feedback on their decisions (see Cowgill 2019). Literature has found that belief updating may be slow, or fail altogether in presence of these features (see, Bohren et al., 2022). Over long enough horizons, μ can still adjust to the increase in precision due to the adoption of predictive AI, but the presence of these features perhaps justifies considering μ to be static in the short run. Next, I highlight some trivial propositions that arise directly from the model.

Proposition 1. The bias parameter μ shifts the effective cutoff *on the realized predictions* $\hat{Y}_i$ to the left by μ for the discriminated group.

$$\begin{aligned}(T_i|W) &= \mathbb{I}(\hat{Y}_i > \lambda).\\ (T_i|M) &= \mathbb{I}(\hat{Y}_i > \lambda - \mu).\end{aligned} \tag{2}$$

From Proposition 1, while both groups face a cutoff of λ on the scores (S_i), for minorities, this translates to an 'effective cutoff' of $\lambda - \mu$ on the realized predictions ($\hat{Y}_i$).[12]

Proposition 2. The measure of minorities that is expected to be assigned to treatment (i.e. denied loan) is $1 - F(\lambda)$ and the associated measure for whites is $1 - F(\lambda - \mu)$. [13]

$$\begin{aligned}Pr(\mathbb{E}[S_i|W] > \lambda) = Pr(Y_i > \lambda) = 1 - F(\lambda).\\ Pr(\mathbb{E}[S_i|M] > (\lambda - \mu) = Pr(Y_i > \lambda - \mu) = 1 - F(\lambda - \mu).\end{aligned} \tag{3}$$

Thus, to fall in the group 'expected to be denied loan', whites face a cutoff (on Y_i), λ while minorities face, $\lambda - \mu$.

Proposition 3. The measure of *expected victims of bias* or *marginal minorities*, those expected to be denied loans based on their Y_i draws, but would not be so if they were white, is given by $F(\lambda) - F(\lambda - \mu)$.

Microfoundation for the cutoffs and the bias parameter. The bias parameter μ can arise either from taste-based discrimination (Becker, 1957) or from inaccurate statistical discrimination (Bohren et al., 2022). First, consider an expected profit-maximizing financial institution that wants to maximize a value function V. From an approved loan, it gets an expected payoff $\pi(Y_i)$. Financial institution's maximization problem:

relying on own judgment only, and 54% reporting relying equally on both.

[12] This follows directly from (1). Note, for whites, the same effective cutoff λ applies to both scores (S_i) and predictions ($\hat{Y}_i$).

[13] F denotes the CDF of the distribution of Y_i which is the same for both groups.

$$\max_{Y^*} V = E \int_0^{Y^*} \pi(Y_i) dY, \tag{4}$$

where, $\pi'(Y_i) < 0, \;\; \pi(0) > 0, \;\; \pi(1) < 0$. These assumptions ensure:

$$\begin{aligned} \exists : Y^* \in [0,1] \;\; s.t. \;\; & Y^* = \text{argmax} \;\; V, \\ & \pi(Y^*) = 0, \\ & \pi(Y_i) \geq 0 \;\; \forall \;\; Y_i \leq Y^*, \\ & \pi(Y_i) < 0 \;\; \forall \;\; Y_i > Y^*. \end{aligned} \tag{5}$$

The financial institution gets a positive payoff for approved loans to applicants with the lowest possible default risk ($Y_i = 0$), and a negative payoff for loans to applicants with the highest possible default risk ($Y_i = 1$). The payoff function is decreasing in default risk Y_i. Therefore, there exists an optimal cutoff default risk level below which the financial institution would want to approve all loans and above which it would want to approve none of the loans. The financial institution, however, cannot solve this problem as Y_i cannot be observed. It employs prediction technology to generate predictions $\hat{Y}_i$, and human evaluators (loan officers) to make the final allocation decision.

Now, consider an expected-utility-maximizing representative evaluator employed by the financial institution who cares to maximize the financial institution's value function. Thus, there is a mapping from the financial institution's expected profit function π to the representative evaluator's expected utility function u. The representative evaluator solves the following:

$$\max_{Y_W^*, Y_M^*} U = U_W + U_M = \int_0^{Y_W^*} u_w(E[\hat{Y}_i|W]) dY + \int_0^{Y_M^*} u_m(E[\hat{Y}_i|M])) dY. \tag{6}$$

U_W and U_M denote the aggregate utility functions from approving loans of white and minority applicants, respectively. Note, the representative evaluator's problem has observables as inputs; algorithm-predicted default risks $\hat{Y}_i$ and race R_i. For whites, the representative evaluator's incentives are perfectly aligned with those of the financial institution. I abstract away from any principal-agent problems for the non-discriminated group. For minorities, I restrict principal-agent problems to those originating only due to the bias of human evaluators (either arising from animus or incorrect beliefs). These restrictions allow me to illustrate the results in a tractable and succinct manner.

Let, the mappings from π to u be such that $u'(Y_i) < 0$, $u(0) > 0$, $u(1) < 0$. This ensures:

$$\exists : Y_W^*, Y_M^* \in [0,1]. \tag{7}$$

Under **(pure) taste-based discrimination**, (see Becker, 1957), the representative evaluator does not believe that minorities are more likely to default compared to whites when they have the same predictions $\hat{Y}_i$. However, the representative evaluator has animus towards minorities and receives additional disutility from approving their loan. Thus,

$$\begin{aligned} E[\hat{Y}_i|M] = E[\hat{Y}_i|W] = Y_i, \\ u_m(Y_i) < u_w(Y_i) \ \forall \ \hat{Y}_i. \end{aligned} \tag{8}$$

Under **(pure) inaccurate statistical discrimination**, (see Bohren et al., 2022), the representative evaluator does not have animus towards minorities but has an inaccurate belief that minorities are more likely to default, and this belief does not converge to the truth. Bohren et al., (2022) show how inaccurate subjective beliefs about the distributions of qualifications and signals by groups can persist even with learning and belief updating rules (both Bayesian and non-Bayesian). While learning can mitigate inaccurate beliefs in some settings, there is little or no feedback on the decisions being made in many other settings, leading to learning traps where inaccurate beliefs can persist (see Lepage, 2020; Bordalo et al., 2016). Bohren and Hauser (2021) showed inaccurate beliefs about the distributions of qualifications and signals may lead to 'incorrect learning' so that those beliefs may not converge to the true distribution. Inaccurate beliefs can arise due to heuristics (see Bohren and Hauser, 2021), or lack of information (i.e. failing to account for selection, see Hübert and Little, 2020). In this setup, this can translate to:

$$\begin{aligned} u_m(Y_i) = u_w(Y_i) \ \forall \ \hat{Y}_i, \\ E[\hat{Y}_i|M] > E[\hat{Y}_i|W] = Y_i. \end{aligned} \tag{9}$$

Recall in this regard the earlier assumptions that the representative evaluator knows that the prediction algorithms are blind to group identities (cannot condition on race and their proxies that are unrelated to creditworthiness due to legal restrictions). Furthermore, the evaluator does not know the exact set of variables that enter the prediction algorithms as inputs. [14] Hence, to account for the perceived differences in the distribution (mean) by group identities, the representative evaluator adds $E[\hat{Y}_i|M] - E[\hat{Y}_i|W]$ to the realized predictions of minorities.

Under either taste-based discrimination or inaccurate statistical discrimination, it fol-

[14] As a counterexample, if the representative evaluator knew that the predictive AI conditions on a previously unaccounted variable that is correlated with group identity (and creditworthiness), they may adjust the bias parameter accordingly.

lows that:

$$\begin{aligned} \exists : Y_W^*, Y_M^* \in [0,1] \;\; s.t. \;\; Y_W^* &= \text{argmax} \;\; U_W, \\ Y_M^* &= \text{argmax} \;\; U_M, \\ Y_M^* &< Y_W^*. \end{aligned} \tag{10}$$

Let, $Y_W^* = \lambda$ and $Y_W^* = \lambda - \mu$.

Thus, we have the cutoff λ and the bias parameter μ arising out of the representative evaluator's expected utility maximization problem.

Discussion on the Model. Existing literature recognizes two primary sources of biased allocations or amplification of bias under predictive algorithms - biased models or biased data. In this paper, I study whether a predictive AI that is unbiased by design and uses unbiased training data can still amplify exogenous human bias through noise reduction of the predictions. In constructing the model, I, therefore, shut off the possibility of bias arising from the predictions. Since I focus on showing the 'existence' of this counter-intuitive channel, the model is decidedly simple and abstracts from many complexities. The setup in the model is general and applies to a wide range of settings (i.e. financial institutions granting loans, criminal justice systems granting bail to defendants, law enforcement departments selecting neighborhoods to raid, similar settings in health, insurance, environmental regulations, food safety regulations etc.). The two-step allocation process with 'humans-in-the-loop' is almost ubiquitous in all settings where predictive AI is used. In the model, the bias parameter μ stays fixed in response to a switch to the less-noisy prediction technology (predictive AI). In this regard, My model is decidedly static and applies to short time horizons. Several features commonly observed in the relevant settings (highlighted earlier) motivate relaxing the assumption of accurate and instantaneous belief updating by human evaluators.

3 Results

In this section, I present closed-form results. But first, I define some key terms and provide graphical intuition for the main propositions of this paper. I define *discrimination* and *aggregate discrimination* as follows:

Definition 1 (Discrimination) Conditional on Y_i draws, the difference between whites and minorities in their loan approval (non-denial) probabilities.

$$\text{Discrimination} = [Pr(Y_i + \epsilon_i < \lambda) - Pr(Y_i + \epsilon_i < \lambda - \mu)|Y_i]. \tag{11}$$

Definition 2 (Aggregate Discrimination) Aggregate measure of discrimination obtained by integrating discrimination over the support of Y_i.

$$\text{Aggregate discrimination} = \int_{Y_i} (\text{discrimination}) \ dY. \tag{12}$$

The question I explore within the framework of this model is what happens to (1) *discrimination* (conditional on Y_i) and (2) *aggregate discrimination* as the prediction technology becomes more precise.

3.1 Response of discrimination to increasing signal precision

First, I address the question of what happens to discrimination as prediction technology becomes more precise. I illustrate the main ideas graphically to build intuition and follow it up with closed-form results afterward. Figure 1 shows how the two different cutoffs on realized predictions faced by whites and minorities create three partitions on the support of Y_i.

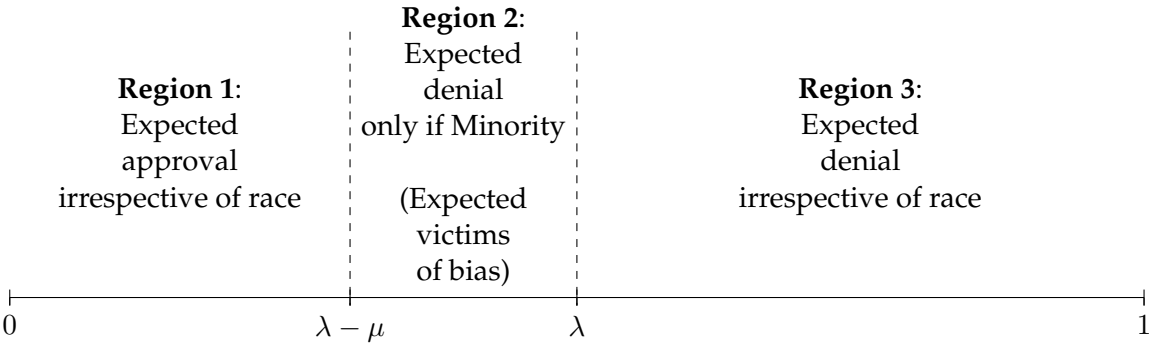

Figure 1: Three regions (partitions) on the support of Y_i

Recall, λ is the cutoff on predictions faced by whites, whereas $\lambda - \mu$ is the cutoff faced by minorities. Applicants are denied loans if their realized predictions $\hat{Y}_i$ fall above their effective cutoff. In other words, applicants are better off if their realized predictions fall below their group-specific cutoff. In the first region are applicants who are not expected to be denied loans irrespective of their race. Recall in this regard that $E[\hat{Y}_i] = Y_i$ as the prediction technology is unbiased. In the second region are *expected victims of bias* or *marginal minorities*. Note, in the absence of any prediction errors (perfect predictions), all of these agents would be denied loans if they are from a minority background but not if they are white. In the third region are applicants who fall above both cutoffs. They are expected to be denied loans irrespective of their group identities. Under perfectly precise predictions (zero prediction error) outcome of applicants will exactly match their expected outcome delineated in Figure 1. The presence of noise, however, may cause realized outcomes to differ from expected outcomes. In Figure 2, I plot 'true' Y_i draws (same as expected predictions) on the X-axis and realized predictions $\hat{Y}_i$, comprising of both signals (Y_i) and noises

(ϵ_i) on the Y-axis.

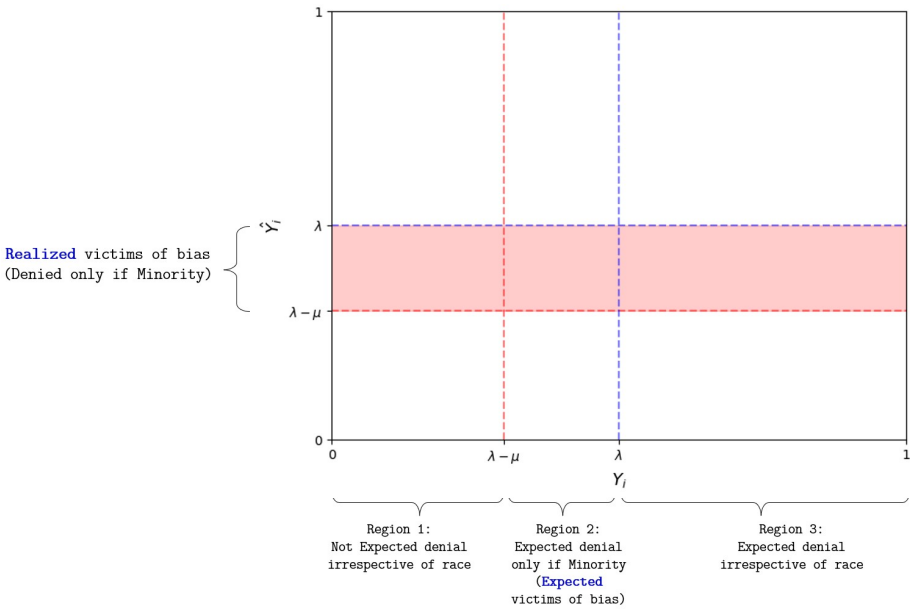

Figure 2: Y_i against $\hat{Y}_i$ and realized victims of bias

The vertical and horizontal lines at λ and $\lambda - \mu$ denote the two cutoffs on expected default risks $Y_i (= E(\hat{Y}_i))$ and realized predictions $\hat{Y}_i$ respectively. On the X-axis, the vertical lines also demarcate the three regions introduced earlier in Figure 1. The red-shaded area represents *realized discrimination* - applicants whose realized predictions are below the cutoff for minorities but above the cutoff for whites, meaning they would be denied loans conditional on their realized signals $\hat{Y}_i$ only if they are minorities but not if they are whites. Figure 3 builds upon Figure 2 and shows what happens to discrimination as σ approaches 0 in the limit with simulated data. Under perfect predictions (right panel), all applicants fall on the 45-degree line going through the origin.

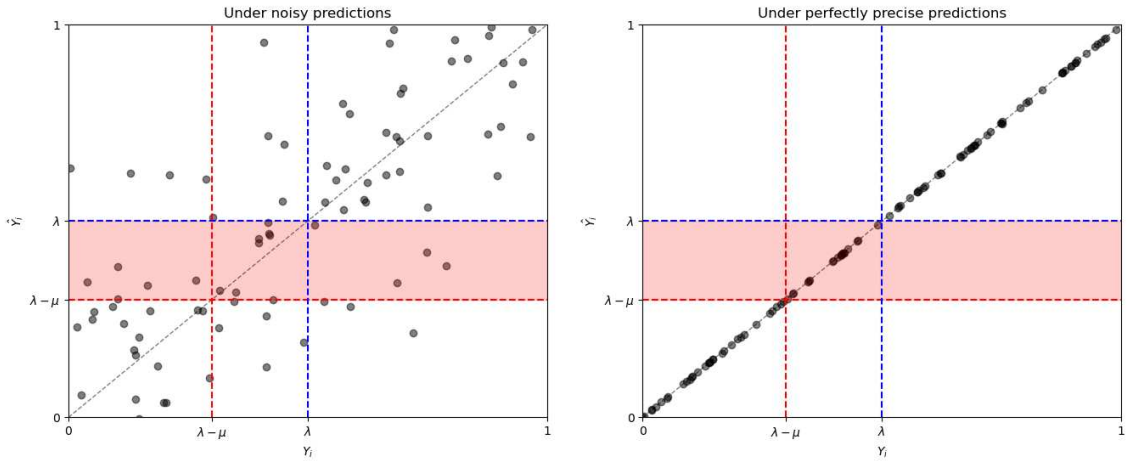

Figure 3: What happens to discrimination as $\sigma \to 0$

First, note that in the limiting case when $\sigma \to 0$, realized discrimination goes down in regions 1 and 3, but goes up in Region 2 (for marginal applicants). This shows that precision gain can have a disparate impact on discrimination, causing discrimination (conditional on Y_i draws) to go up for certain subgroups of minorities and go down for others. [15]. Second, note that under noisy predictions (left panel), some minority applicants from all three regions may become *realized* victims of bias, but under perfectly precise predictions (right panel), only marginal minority applicants who are *expected* victims of bias (Region 2) become realized victims of bias. This is because under noisy predictions (left panel) some minority applicants from Region 2 (marginal minorities) get lucky and get *saved by randomness*. This means they get a negative noise draw large enough in magnitude to take their realized predictions below the cutoff for minorities $\lambda - \mu$. This luck factor goes away as randomness decreases under precise predictions. Thus, it seems to indicate that there exists a subgroup of minorities - *marginal minorities* whose true Y_i draws are below the cutoff for whites but above the cutoff for minorities, for whom discrimination gets worse, at least weakly. Note, for the simulated case shown in Figure 3, the net effect on discrimination mostly cancels out, but this is not always the case. I shed light on when this does not cancel out so that an aggregate increase or decrease in discrimination becomes possible later in this section. These observations build the intuition for two main propositions of this paper. The two main propositions and their closed form proof follow after a brief analysis of the welfare effect on applicants by race as $\sigma \to 0$, which is shown in Figure 4.

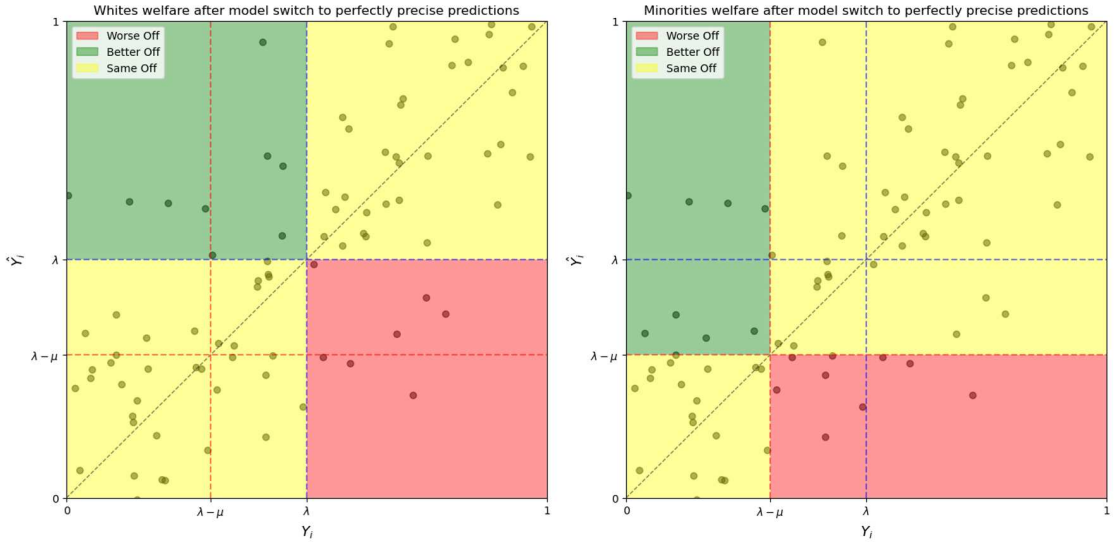

Figure 4: Welfare effect on applicants by race as $\sigma \to 0$

The left and right panels of Figure 4 shows the welfare effect on whites and minorities, respectively, in the limit as noise σ goes to zero. The points represent applicants under noisy predictions and use the same simulated data shown in Figure 3. Under noiseless

[15] Subgroups (regions) in this context are based on true Y_i draws

predictions, applicants find themselves on the 45-degree line going through the origin. The shaded boxes denote 'worse off', 'better off', and 'same off' areas. For example, in the left panel, white applicants who are in the top-left green box get denied loans under noisy predictions as their *realized* predictions are above the cutoff for whites, λ. However, if their *expected* predictions are realized, they would drop vertically along the Y-axis until they are on the 45-degree line. If the noise goes to zero, they would thus find themselves to be better off as they would be below the cutoff λ and no longer be denied loans. This deviation between realized and expected predictions under noisy predictions is due to randomness, and these applicants 'get unlucky' under noisy predictions. Overall, we see that in the limit as noise σ goes to zero, for this simulated data, the effects seem to almost cancel out on the extensive margin such that the aggregate effect is almost neutral. There are almost equal numbers of applicants who are better off and who are worse off from both whites and minorities. However, depending on parameter values, the effects may not always cancel out. The question of when it is more likely for aggregate effects to not cancel out is addressed later in this section. Now, based on the insights from Figure 3, I introduce the main propositions of this paper and some associated definitions.

Proposition 4. *Expected victims of bias* or *marginal minorities* become less likely to get *'saved by randomness'* as predictions become more precise.

Definition 3 (Expected Victims of Bias or Marginal Minorities) are minority applicants whose Y_i draws are below the cutoff on the prediction for whites (λ) but above that for minorities ($\lambda - \mu$).[16] They are 'marginal' minorities in the sense that they are expected to be denied loans but would not be so if they were white. This is expressed mathematically as follows:

$$Y_i \text{ s.t. } Y_i \in (\lambda - \mu, \lambda]; \;\; R_i = M. \tag{13}$$

Definition 4 (Saved by Randomness.) Expected victims of bias are *saved by randomness* if in the mapping from Y_i to $\hat{Y}_i$ they get negative noise draws (prediction errors) large enough in magnitude to take their realized predictions $\hat{Y}_i$ below their group-specific cutoff ($\lambda - \mu$).[17] This is expressed mathematically as follows:

$$Y_i \text{ s.t. } Y_i \in (\lambda - \mu, \lambda]; \;\; R_i = M; \;\; \hat{Y}_i = Y_i + \epsilon_i < \lambda - \mu. \tag{14}$$

Proof of Proposition 4.

$$\begin{aligned} &Pr(\text{expected victims of bias saved by randomness}) = \\ &Pr(\epsilon < \lambda - \mu - Y_i | Y_i \in (\lambda - \mu, \lambda]). \end{aligned} \tag{15}$$

[16] They correspond to Region 2 as shown in Figure 1.
[17] If this outcome materializes, they escape becoming *realized* victims of bias.

As $\lambda - \mu - Y_i < 0 \quad \forall \quad Y_i \in (\lambda - \mu, \lambda]$, this probability is always increasing in σ (decreasing in precision parameter η) for noise draws ϵ_i drawn from mean-zero distributions where extreme outcomes are less likely. i.e. For ϵ_i drawn from normal:

$$\frac{d\Phi\left(\frac{\lambda-\mu-Y_i}{\sigma}\right)}{d\sigma} = \phi\left(\frac{\lambda - \mu - Y_i}{\sigma}\right) \cdot \frac{-(\lambda - \mu - Y_i)}{\sigma^2} > 0. \tag{16}$$

Thus, the probability that expected victims of bias are saved by randomness is non-decreasing in noise variance (decreasing in signal precision) for noise draws ϵ_i drawn from mean-zero distributions where extreme outcomes are less likely. (Q.E.D.)

This shows the existence of a channel through which aggregate discrimination can potentially increase under more precise predictions (if this effect dominates).

Proposition 5. Discrimination is non-decreasing in precision for *expected victims of bias*.

Proof of Proposition 5.

Recall the definition of discrimination (conditional on Y_i) in (11). For normally distributed noise draws, this translates to:

$$\begin{aligned} \text{Discrimination} &= [Pr(Y_i + \epsilon_i < \lambda) - Pr(Y_i + \epsilon_i < \lambda - \mu)|Y_i], \\ &= \Phi(\frac{\lambda - Y_i}{\sigma}) - \Phi(\frac{\lambda - \mu - Y_i}{\sigma}). \end{aligned} \tag{17}$$

Differentiating with respect to σ and $\eta = (\frac{1}{\sigma})$:

$$\frac{d(Discrimination)}{d\sigma} = \phi(\frac{\lambda - \mu - Y_i}{\sigma})(\frac{\lambda - \mu - Y_i}{\sigma^2}) - \phi(\frac{\lambda - Y_i}{\sigma})(\frac{\lambda - Y_i}{\sigma^2}). \tag{18}$$

$$\begin{aligned} \frac{d(Discrimination)}{d\eta} &= \frac{d(Discrimination)}{d(\frac{1}{\sigma})} \\ &= 2[\phi(\frac{\lambda - Y_i}{\sigma})(\lambda - Y_i) - \phi(\frac{\lambda - \mu - Y_i}{\sigma})(\lambda - \mu - Y_i)]. \end{aligned} \tag{19}$$

The sign of the term in (19) and its interaction with the distribution of Y_i determines whether there will be a net increase or decrease in aggregate discrimination under a more-precise prediction technology. Table 1 shows the sign of the term in (19) over the three regions (conditioning sets) of the support of Y_i defined in Figure 1.

	$\phi(\frac{\lambda-Y_i}{\sigma})$	$(\lambda - Y_i)$	$\phi(\frac{\lambda-\mu-Y_i}{\sigma})$	$(\lambda - \mu - Y_i)$	$\frac{d(Discrimination)}{d\eta}$
Region 1 $0 \leq Y_i \leq \lambda - \mu$	$+$	$++$	$++$	$+$	+/-
Region 2 $\boldsymbol{\lambda - \mu < Y_i \leq \lambda}$	$+$	$+$	$+$	$-$	$+$
Region 3 $\lambda < Y_i \leq 1$	$++$	$-$	$+$	$--$	+/-

Note: '++' and '−−' used to denote $>> 0$ and $<< 0$ in a relative sense

Table 1: Sign of $\frac{d(discrimination)}{d\eta}$ over three regions

It is clear from Table 1 that discrimination is always and unambiguously non-decreasing for expected victims of bias (Region 2) as predictions become more precise (Q.E.D)

Note, this is true for this model setup irrespective of the distribution of Y_i and parameter values (bias parameter μ, cutoff λ). The effect of precision gain on discrimination for other regions is ambiguous, going up for some draws of Y_i while going down for others. Note, however, that (19) can be evaluated numerically for any given parameter values. I evaluate (19) holding fixed cutoff $\lambda = 0.5$ and human bias parameter $\mu = 0.1$, and varying starting σ (before AI adoption) over the support of Y_i (range from 0 to 1, spaced at 0.01 interval) in Figure 5.

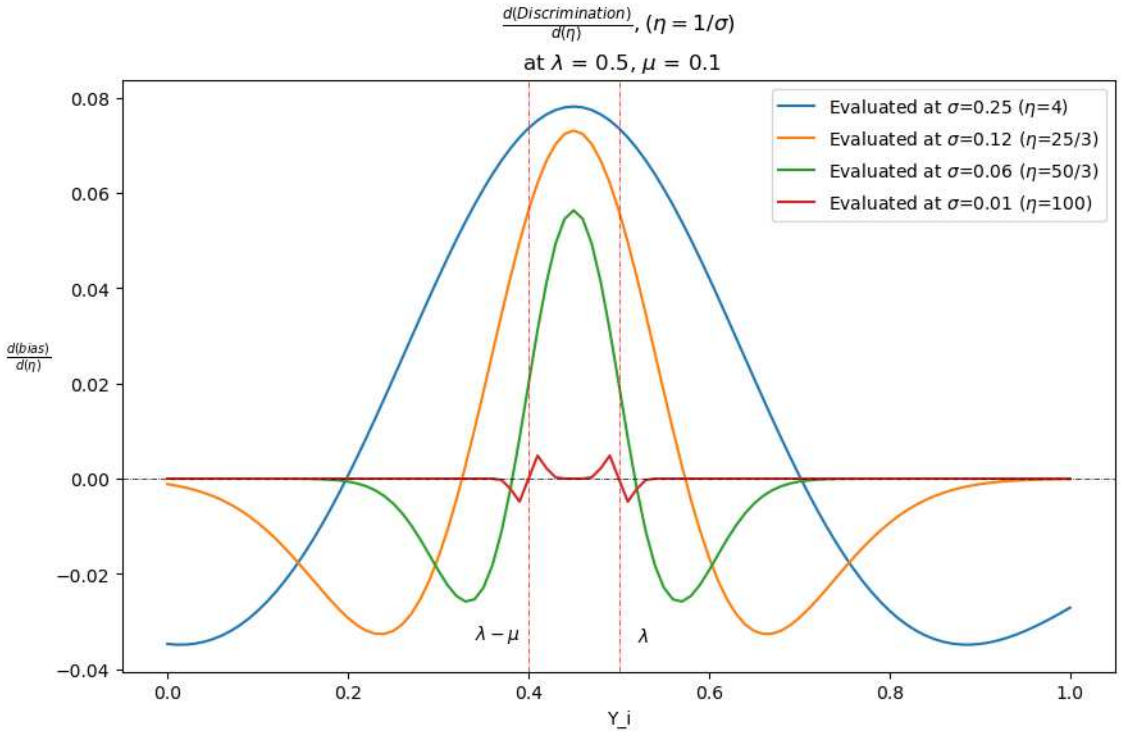

Figure 5: Rate of change in discrimination (conditional on Y_i) for precision gain

Figure 5 shows the rate of change in discrimination (conditional on Y_i) for a small increase in precision gain evaluated over the support of Y_i. Discrimination is increasing with signal precision when the function is above the zero line and vice versa. The two red vertical lines at λ and $\lambda-\mu$ represent the effective cutoffs for whites and minorities, respectively. The area in between them, therefore, indicates the range of Y_i draws of expected victims of bias (Region 2). Note in this region, the function is always non-negative for all parameter values. This is consistent with the closed-form result of Proposition 5. In Regions 1 and 3, the function is positive over some intervals and negative over others depending on parameter values. The role of starting precision level (and noisiness) becomes clear from Figure 5. If existing prediction technology is already very precise (σ small), small precision gains do not lead to a large rate of increase or decrease in discrimination. However, if existing prediction technology is noisy, even a small precision gain (due to the adoption of AI) can lead to a large rate of increase or decrease in discrimination. In Figure 6, I show the sensitivity of the (19) in response to changes in parameter values λ and μ. Figure 6 thus contains versions of Figure 5 with different values of λ and μ, respectively.

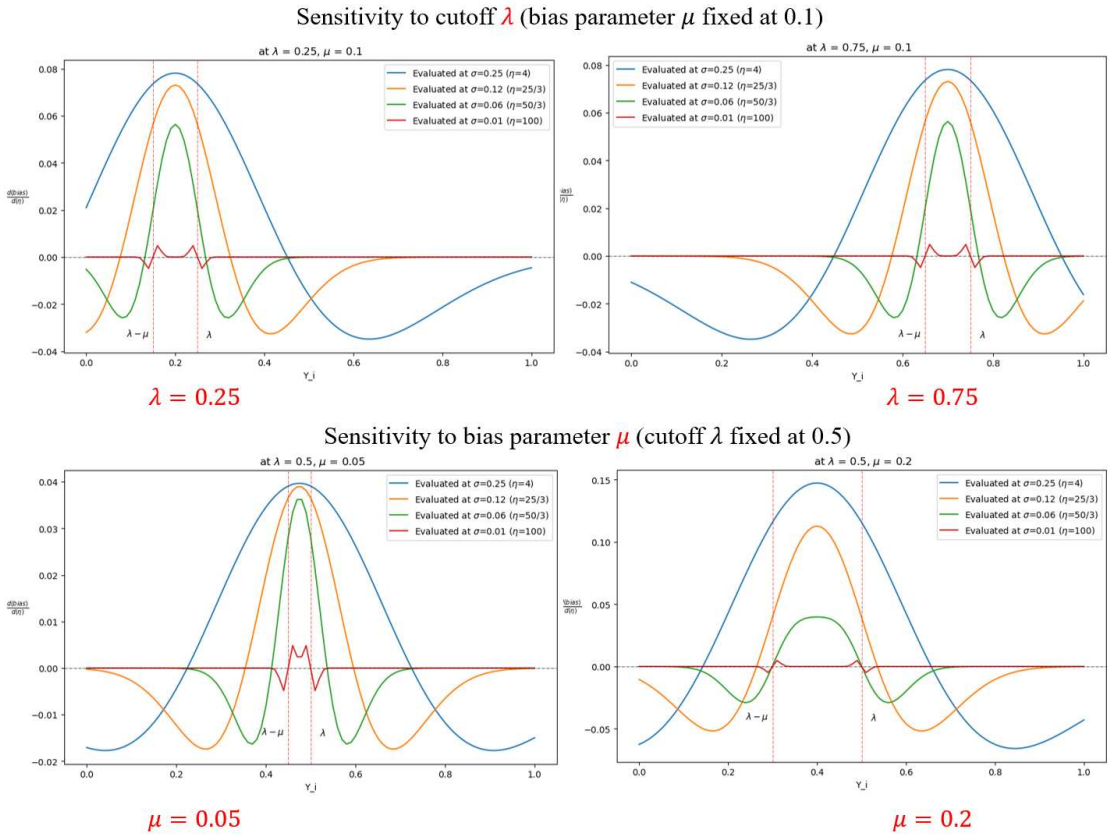

Figure 6: Sensitivity of rate of change of discrimination to parameter values

As can be seen from Figure 6, changes in λ (cutoff) holding everything else constant effectively shift the function to the left or right, whereas changes in μ (bias intensity) holding everything else constant change the size of Region 2 (where the function is always non-decreasing) and the spread of the function. The implication is that, if the mass of Y_i is large in Region 2 relative to the other regions, aggregate discrimination can increase with

precision gain as the effect on Region 2 will dominate (i.e. if Y_i is normally distributed with cutoff λ at 0.5).

The takeaway from this subsection is that, discrimination is always non-decreasing in precision gain for marginal minorities, and aggregate discrimination can increase if their relative mass is large such that this effect dominates. I shift focus to what happens to *aggregate* discrimination under more precise predictions in the next subsection.

3.2 Response of *aggregate* discrimination to increasing signal precision

The net effect on aggregate discrimination depends not only on the function $\frac{d(Discrimination)}{d\eta}$, but also on the distribution of Y_i, and the extent of precision gain. First, in Figure 7, to assess the responsiveness of aggregate discrimination to increasing signal precision, I show the definite integral of the $\frac{d(Discrimination)}{d\eta}$ function for different σ over the support of Y_i (with λ and μ fixed at 0.5 and 0.1 respectively). Note, the functions in Figure 7 are the definitive integrals of the functions in Figure 5. At $Y_i = 1$, if this integral ends above the 0 line, aggregate discrimination is increasing with precision and vice versa.

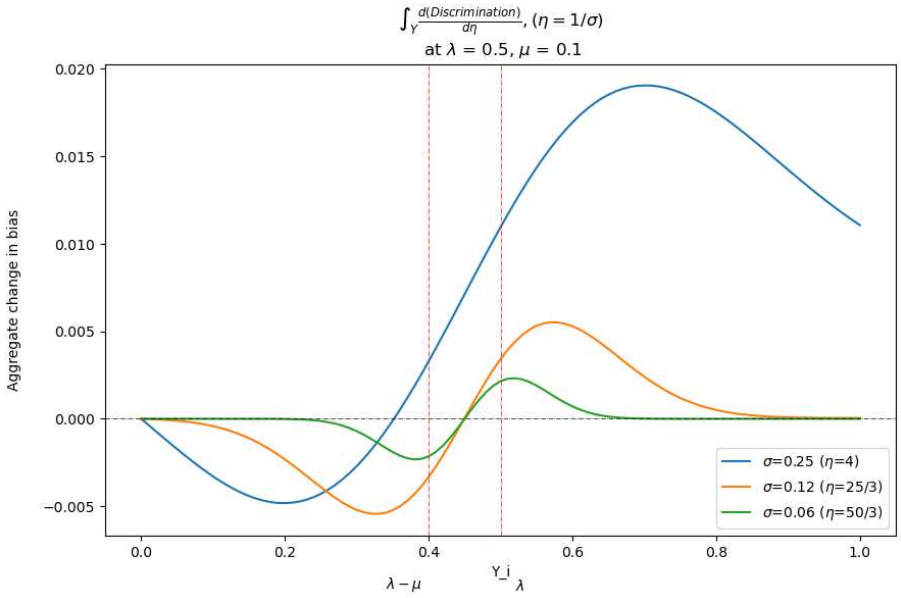

Figure 7: Sensitivity of aggregate rate of change of discrimination to parameter values

From the figure, we see that the net change in aggregate discrimination cancels out if starting noisiness of the prediction technology (before AI adoption) is small, but not when it is large enough. Comparing this figure with Figure 5, it becomes clear that when the underlying $\frac{d(Discrimination)}{d\eta}$ function that is integrated over is not symmetric over the support of Y_i, the aggregate effect on discrimination does not cancel out. From Figure 6, it can be seen that the $\frac{d(Discrimination)}{d\sigma}$ function is more likely to become asymmetric if the starting noisiness level before AI adoption σ is large, if the bias parameter μ is large, and if the cut-

off parameter λ is further away from the center of support of Y_i. Under these conditions, a change in precision is more likely to have a sizable impact on aggregate discrimination. In general, aggregate discrimination can be expected to increase with a precision gain if the mass of Region 2 is large relative to the mass of Regions 1 and 3. The effect is intensified if the starting noisiness level before AI adoption σ is large and if the magnitude of precision gain is large.

I evaluate $\int_Y \frac{d(Discrimination)}{d\eta}$ over a parameter space as follows:

$$\begin{aligned} \mu &\in \{0.01, 0.02, \ldots, 0.2\} && \text{bias parameter.} \\ \sigma &\in \{0.01, 0.02, \ldots, 0.25\} && \text{noise standard deviation.} \\ \lambda &\in \{0.25, 0.5, 0.75\} && \text{cutoff.} \\ Y_i &\sim U(0,1),\ \ N(0.5, \frac{0.5}{3}),\ \ \beta(2,5) && \text{distribution.} \end{aligned}$$

In the parameter space considered, the bias parameter μ is allowed to vary at 0.01 intervals over the range $[0.01, 0.2]$. To put this in perspective, $\mu = 0.1$ means the human evaluator adds a 10% penalty to the machine-predicted default risk for minorities. The noise parameter σ is allowed to vary at 0.01 intervals over the range $[0.01, 0.25]$. $\sigma = 0.01$ and $\sigma = 0.25$ means a signal-to-noise ratio of 50 and 2 respectively for the median applicant[18] with default risk Y_i of 0.5. The three cutoff values considered are 0.5, 0.25, and 0.75. The three different distributions for Y_i considered are uniform, normal, and right-skewed distributions. Note, $N(0.5, \frac{0.5}{3})$ makes the shape of the distribution to be normal while restricting approximately 99.5% of the draws of Y_i within the $[0, 1]$ interval (to maintain a probability interpretation of default risks). The $\beta(2, 5)$ distribution makes the distribution of Y_i right-skewed with the mode at 0.2 while restricting Y_i within the $[0, 1]$ interval. This distribution has some empirical relevance, as the empirical distribution of credit scores is left-skewed, and the default risk Y_i in this setting can be interpreted as negative credit scores.

In Figure 8, I show how $\int_Y \frac{d(Discrimination)}{d\eta}$ changes over the parameter space using a 3D plot, holding fixed $\lambda = 0.5$. The X-axis denotes bias parameter μ, the Y-axis denotes the baseline noise level of the prediction technology before AI adoption σ and Z-axis shows the change in the aggregate rate of change of discrimination in response to a small increase in precision. Different colors are used to denote the three different distributions. The shaded gray surface denotes zero on the Z-axis and is the zero-impact reference line.

[18] assuming a symmetric distribution of Y_i

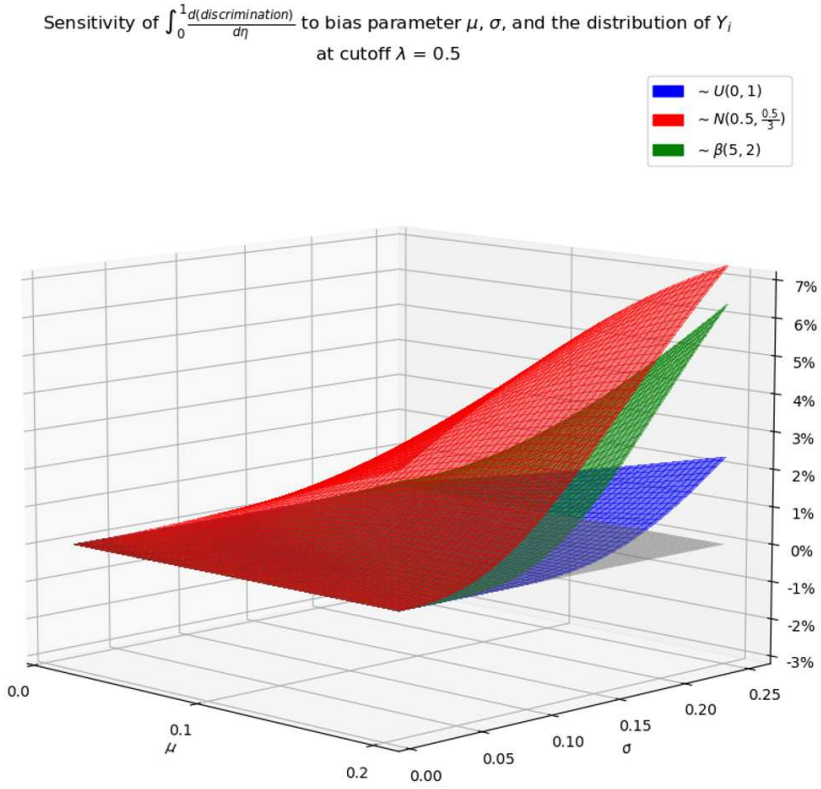

Figure 8: Sensitivity of aggregate rate of change in discrimination to parameter values

It can be seen from Figure 8 that over this parameter space, the aggregate rate of change in discrimination for a small increase in precision is increasing in μ and σ. As expected, the effect is largest for the normal-shaped distribution of Y_i. This is because for the cutoff λ set at 0.5, the relative mass of region 2 is largest under normal $N(0.5, \frac{0.5}{3})$, followed by under right-skewed $\beta(2,5)$ distribution. Also, the effect is negligible for small values of μ and σ, as for these values, the function $\frac{d(Discrimination)}{d\eta}$ is symmetric (Figure 5), so that the net effect cancels out.[19]

Figure 9 shows the sensitivity to the change in the cutoff parameter λ, by setting it to 0.25 (left panel) and 0.75 (right panel), respectively. It is seen that a decrease in the rate of change in aggregate discrimination is also possible at large values of σ and μ.

[19]Recall, the net increase in discrimination in Region 2 gets counteract by the net decreases in discriminations in Regions 1 and 3 when this function is symmetric.

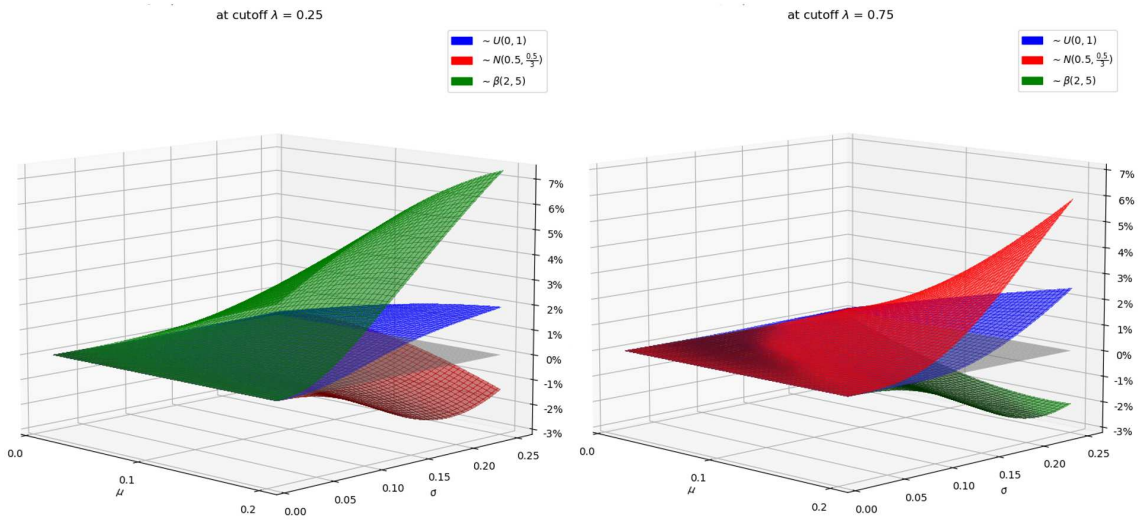

Figure 9: Sensitivity of aggregate rate of change in discrimination to the cutoff parameter

The result of Figure 9 is better understood by comparing the graphs with the corresponding $\frac{d(Discrimination)}{d\eta}$ graph in Figure 6 for $\lambda = 0.25$ and $\lambda = 0.25$ respectively. Figure 10 and Figure 11 place them side by side for convenience. As can be seen in Figure 10, at $\lambda = 0.25$, Region 2, where discrimination increases in response to precision gain, is centered at $\frac{0.25-\mu}{2}$. The mass of right-skewed $\beta(2, 5)$ distribution is concentrated at and near 0.2. As a result, the increase in the aggregate rate of change in discrimination is larger for the $\beta(2, 5)$ distribution. For the normal $N(0.5, \frac{0.5}{3})$ distribution, the mass is concentrated at and near 0.5 where the $\frac{d(Discrimination)}{d\eta}$ function is non-positive for $\lambda = 0.25$. As a result, this negative effect dominates for the normal-shaped distribution, and there is a net decrease in the aggregate rate of change of discrimination with respect to precision when μ and σ become large.

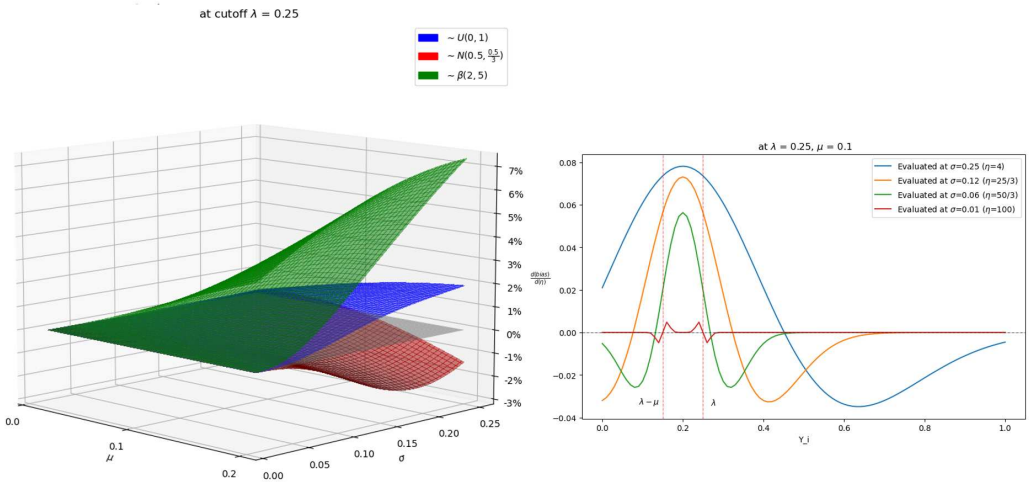

Figure 10: Sensitivity of rate of change in aggregate discrimination at $\lambda = 0.25$

The interpretation of Figure 11 is similar. At $\lambda = 0.75$, Region 2, where discrimination increases in response to precision gain, is centered at $\frac{0.75-\mu}{2}$. The increase in the aggregate rate of change in discrimination is larger for the normal $N(0.5, \frac{0.5}{3})$ distribution as the relative mass near Region 2 is larger for this distribution. In contrast, the right-skewed $\beta(2, 5)$ distribution is concentrated at and near 0.2 where the $\frac{d(Discrimination)}{d\eta}$ is non-positive. As a result, the negative effect dominates for this distribution as σ and μ becomes large.

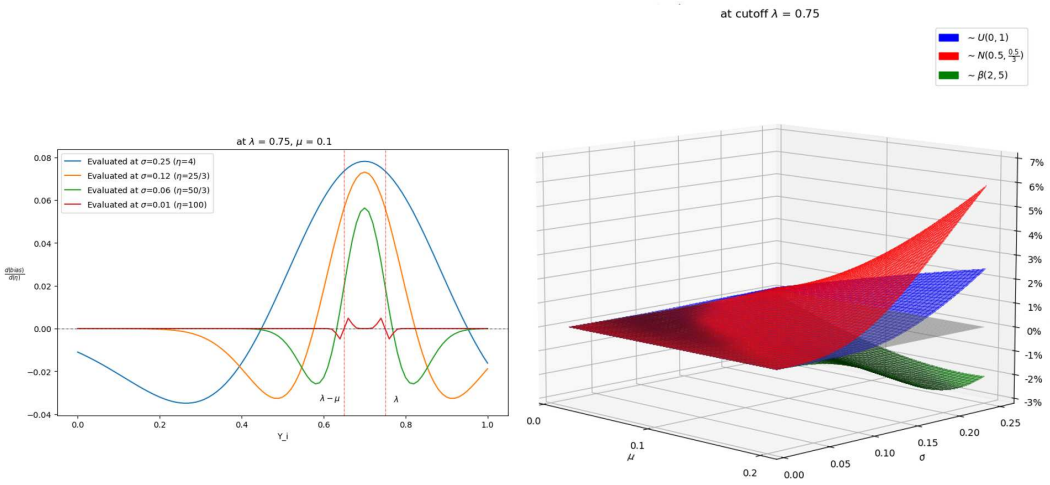

Figure 11: Sensitivity of rate of change in aggregate discrimination at $\lambda = 0.75$

The takeaway from this subsection is that precision gain is neutral to aggregate discrimination when the first derivative of discrimination with respect to precision is symmetric over the support of the qualification distribution. Otherwise, aggregate discrimination can be either increasing or decreasing in precision gain. The aforementioned function becomes more likely to be asymmetric and more asymetric as bias intensity increases, the noisiness of the baseline prediction technology (before the adoption of predictive AI) increases, and the cutoff is further away from the center of the support of the qualification distribution. When the aforementioned function is asymmetric, whether aggregate discrimination is increasing (or decreasing) depends on whether the mass of the distribution of qualifications is concentrated near the regions where discrimination increases (or decreases). In general, aggregate discrimination is increasing in precision gain if the relative mass of *marginal minorities* (or *expected victims of bias*) is large, and this is more likely if the distribution of the qualifications is centered near the cutoff.

4 Conclusion

In this paper, I analyze what happens to discrimination, originating in exogenous bias, as unbiased prediction technology becomes more precise. I study this for settings where the prediction algorithm functions as decision aids and human evaluators who are biased

against minorities but not against whites function as final decision makers. This setting with humans in the loop is quite ubiquitous. However, the role of this human-machine interaction has not been studied extensively in the algorithmic bias literature.

The adoption of predictive AI is conceptualized as a bias-neutral precision gain for the prediction technology. While the model I study applies to a wide range of settings where predictive AI is used, I consider the setting to be in lending for demonstration. Discrimination is defined as the difference in the probability of getting loan approvals between whites and minorities who are otherwise identical in terms of underlying default probability. Aggregate discrimination is defined as the integral of discrimination over the support of default probabilities.

I show that, in absence of accurate and instantaneous belief updating, which is the likely case in short time horizons over which empirical evaluation of predictive AI's impact on discrimination takes place, discrimination is non-decreasing in precision gain for *marginal minorities* near the cutoff. *Marginal minorities* are the *expected victims of bias*, minority applicants who are expected to be denied loans, but would not be expected to be denied loans if they were white. This result arises due to the fact that *marginal minorities* become less likely to be 'saved by randomness' under more precise predictions. This is a strong result as it holds for all parameter values and the choice of underlying distributions of default probabilities.

The relaxation of the assumption of accurate and instantaneous belief updating, at least in short time horizons, is further motivated by several features found in the relevant setting being studied. These features are the black-box, trade-secret, and race-blind nature of predictive algorithms and the limited information and feedback available to human evaluators who make high-volume decisions in fast-paced environments with the possibility of exhaustion.

If the first derivative of discrimination (conditional on default probability) with precision is symmetric or near-symmetric over the support of default risk distribution, the increase in discrimination for the *marginal minorities* mostly cancels out in aggregate, as there are other minority applicants for whom discrimination decreases. Otherwise, aggregate discrimination can either increase or decrease in response to precision gain, and the magnitude of the effect is increasing in bias intensity, noisiness of the prediction technology before the adoption of predictive AI, and the extent of precision gain. In general, an increase in aggregate discrimination is possible if the relative mass of *marginal minorities* is large (compared to other minorities), if the distribution of default probabilities in concentrated near the cutoff, and if the first derivative of discrimination (conditional on default probability) with precision is asymmetric.

The contribution of this paper is twofold. First, it shows a bias-neutral precision gain of an unbiased prediction algorithm can have a disparate impact on discrimination in the short run, decreasing discrimination for some while increasing discrimination for the *marginal minorities* who are worse off or at least the same off, but never better off. Second, it shows the counter-intuitive result that even a bias-neutral precision gain of an unbiased prediction algorithm can potentially amplify exogenous (human) bias, resulting in an aggregate rise in discrimination. It also elucidates the underlying mechanism through which this counter-intuitive result arises. Not accounting for this mechanism may result in AI being unduly blamed for creating bias when the impact of prediction AI is being evaluated. Such evaluation often takes place shortly after the adoption of predictive AI and justifies the modeling choice of a short time horizon. Not understanding this mechanism has pitfalls for society at large, especially when fear and distrust of AI is high, as this may inadvertently thwart or slow down the adoption of a beneficial technology.

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence.* Harvard Business Review Press.

Arnold, David, Will Dobbie, and Crystal S Yang. (2018). *Racial Bias in Bail Decisions.* The Quarterly Journal of Economics 133 (4): 1885–1932.

Avery, Mallony, Andreas Leibbrandt, and Joseph Vecci. (2023). *Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech.* Working Paper.

Barocas, Solon, and Andrew D. Selbst. (2016). *Big Data's Disparate Impact.* California Law Review 104(3): 671-732.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace. (2022). *Consumer-Lending Discrimination in the FinTech Era.* Journal of Financial Economics 143 (1): 30–56.

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

Bohren, J. Aislinn, and Daniel N. Hauser. (2021). *Learning With Heterogeneous Misspecified Models: Characterization and Robustness.* Econometrica 89 (6): 3025–77.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin Pope. (2022). *Inaccurate Statistical Discrimination: An Identification Problem.* Review of Economics and Statistics.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. (2016). *Stereo-*

types. The Quarterly Journal of Economics 131 (4): 1753–94.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. (2018). *What Can Machines Learn, and What Does It Mean for Occupations and the Economy?* AEA Papers and Proceedings 108 (May): 43–47.

Cowgill, Bo. (2019). *Bias and Productivity in Humans and Machines.* Working Paper.

Cowgill, Bo, Fabrizio Dell'Acqua, Sam Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. (2020). *Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics.* Working Paper. https://doi.org/10.2139/ssrn.3615404.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. (2015) *Automated experiments on ad privacy settings.* Proceedings on privacy enhancing technologies, 2015 (1), 92–112.

Hübert, Ryan, and Andrew T Little. (2023). *A Behavioural Theory of Discrimination in Policing.* The Economic Journal, June, uead043.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. (2018). *Human Decisions and Machine Predictions.* The Quarterly Journal of Economics 133 (1): 237–93.

Korinek, Anton. (2023). *Language Models and Cognitive Automation for Economic Research..* NBER Working Paper No. w30957.

Lepage, Louis-Pierre. (2020) *Endogenous Learning and the Persistence of Employer Biases in the Labor Market.* Working Paper.

Monahan, John, Anne Metz, Brandon L. Garrett, and Alexander Jakubow. (2020). *Risk assessment in sentencing and plea bargaining: The roles of prosecutors and defense attorneys.* Behavioral Sciences & the Law 38(1): 1-11.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations.* Science 366(6464): 447-453.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. (2020). *An Economic Perspective on Algorithmic Fairness.* AEA Papers and Proceedings 110 (May): 91–95.