Downstream Effects of Affirmative Action

Sampath Kannan
University of Pennsylvania
kannan@seas.upenn.edu

Aaron Roth
University of Pennsylvania
aaroth@cis.upenn.edu

Juba Ziani
California Institute of Technology
jziani@caltech.edu

ABSTRACT

We study a two-stage model, in which students are 1) admitted to college on the basis of an entrance exam which is a noisy signal about their qualifications (type), and then 2) those students who were admitted to college can be hired by an employer as a function of their college grades, which are an independently drawn noisy signal of their type. Students are drawn from one of two populations, which might have different type distributions. We assume that the employer at the end of the pipeline is rational, in the sense that it computes a posterior distribution on student type conditional on all information that it has available (college admissions, grades, and group membership), and makes a decision based on posterior expectation. We then study what kinds of fairness goals can be achieved by the college by setting its admissions rule and grading policy. For example, the college might have the goal of guaranteeing equal opportunity across populations: that the probability of passing through the pipeline and being hired by the employer should be independent of group membership, conditioned on type. Alternately, the college might have the goal of incentivizing the employer to have a group blind hiring rule. We show that both goals can be achieved when the college does not report grades. On the other hand, we show that under reasonable conditions, these goals are impossible to achieve even in isolation when the college uses an (even minimally) informative grading policy.

CCS CONCEPTS

• **Mathematics of computing** → *Probability and statistics*; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Long-term fairness; affirmative action; college admissions; job market

ACM Reference Format:
Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream Effects of Affirmative Action. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3287560.3287578

Kannan's research was supported in part by NSF grant AF-1763307 and a grant from the Quattrone Center for the Fair Administration of Justice. Roth's research was supported in part by NSF grants CNS-1253345, AF-1763307, and a grant from the Quattrone Center for the Fair Administration of Justice. Ziani's research was supported in part by NSF grants CNS-1331343 and CNS-1518941, and the Linde Graduate Fellowship at Caltech.

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287578

1 INTRODUCTION

For a variety of reasons, including unequal access to primary education, family support, and enrichment activities, different demographic groups can vary widely in their level of preparation by the time they reach their senior year of high school, when they apply for college. In an attempt to correct for this unfortunate reality, many colleges in the United States follow some sort of affirmative action policy in their admissions, which is to say, their admissions decisions explicitly take demographics into account. What is often unstated (and perhaps not even explicitly considered by the colleges) is what exactly the long term goals of these policies are, beyond the short term goal of having a diverse freshman class. In this paper, we consider two explicit goals, and study the extent to which they can be met in a simple two stage model:

(1) **Equal opportunity**: The probability that an individual is accepted to college *and then* ultimately hired by an employer may depend on an individual's type, but conditioned on their type, should not depend on their demographic group.
(2) **Elimination of Downstream Bias**: Rational employers selecting employees from the college population should not make hiring decisions based on group membership.

Neither of these desiderata will necessarily be achieved by admissions rules that ignore demographic information. For example, suppose college admissions is set by a uniform admissions threshold on entrance exam scores. Assuming these scores are equally informative about all groups, this will guarantee that conditioned on a student's type, whether or not she is *admitted to college* will be independent of her group membership, but it does not imply that whether or not she is ultimately hired is independent of her group! This is because exam scores are only a noisy signal about student type. Therefore, if two groups have different prior distributions on type, they will have different posterior distributions on type when conditioned on being admitted to college according to a group-blind admissions rule. The result will be that a Bayesian employer will insist that students from a group with lower mean or higher variance will have to cross a higher threshold on their college grades in order to be hired. In addition to incentivizing explicit group-based discrimination by the employer, this also results in a failure of equal opportunity for the students, because once admitted to college, two individuals of the same type might have to cross different grade thresholds in order to be hired. Thus, a simple "group blind" admissions rule fails to achieve either goal 1 or 2 as laid out above. In this paper, we study the extent to which these goals can be achieved via other means available to the college: in particular, how it admits and grades students.

1.1 Limitations of our Model

When interpreting our results, it is important to understand the scope and limitations of our model. First, this paper considers fairness goals that are limited to preventing inequity from being further propagated — treating opportunities at the high school level and earlier as fixed — and that do not attempt to correct for past inequity. This manifests itself in that our "equal opportunity" goal takes as given that the prospects for employment may "fairly" vary as a function of an individual's type *at the time at which they apply for college*, and does not attempt to address or correct the historical forces that might have resulted in different groups having different type distributions to begin with. Attempting to correct for this kind of historical inequity would require a "value-added model" of education, in which colleges can *change* the type distributions of their student population either through the direct effect of education, or through a second order effect on student behavior before they apply. In our model, colleges do not change student types, they only serve as signaling mechanisms. Similarly, our "equal opportunity" goal aims to equalize the probability that students are hired conditioned on their types — but one might reasonably instead ask for a corrective notion of fairness, in which the probability of passing through the pipeline is *higher* for the historically disadvantaged group conditioned on type. We do not consider this.

Our model also ignores the possibility that exam scores and grades are themselves *biased*. We explicitly assume the opposite — that exam scores and grades are unbiased estimators of student types, for both groups. If instead exam scores were systematically biased downwards for one group, then the response of a rational employer to an admissions policy would be very different — because students who made it through the college pipeline *despite* negative bias would have a higher relative posterior probability of having a high type. There is evidence that effects of this sort are real [2].

Further, the abstraction of one dimensional, stable "types" itself (common in economic models) is clearly an enormous simplification. In reality, the kind of talents valued by employers are multi-dimensional, and dynamically changing. We ignore these complications for simplicity, but believe that studying them are natural and interesting directions for future work.

The two kinds of fairness goals that we study do not speak to the size of the student of employee population coming from each group. For example, in principle, one could satisfy both the equal opportunity and elimination-of-downstream-bias goals that we propose, but at a cost of employing very few individuals from one of the groups. However, we show that even without an additional goal of having large representation from both groups, the fairness goals we set out cannot generally be achieved.

Finally, we assume that employers are single-minded expectation maximizers, with no explicit desire for fairness or diversity. Of course this is often not the case.

Despite these limitations and simplifying assumptions, we find that in the model we study, many natural fairness goals are already impossible. Our paper should be regarded as taking a first step in the study of fairness for pipelines of algorithmic decisions. We think that these negative results are likely to persist in more complex models that attempt to capture additional realism. It is worth exploring whether positive results can be achieved for natural approximations of our fairness notions. If positive results can be found, it would be good to see if they continue to hold as we relax some of the limiting assumptions in our model.

1.2 Our Model and Results

We consider a simple model of admissions, grading and hiring that views the role of colleges only as a means of signaling quality and performing a gatekeeping function, rather than as providing explicit value added[1]. We consider two *groups* representing pre-defined subsets of the population, divided according to socio-economic or other demographic lines. Each student from group i is endowed with a *type* t, which is drawn independently from a Gaussian type distribution P_i that is dependent on the students' group membership. A student's type ultimately measures her value to an employer. We model employers as having a fixed cost C for hiring an individual, and a gain that is proportional to their type. If the employer hires an individual who has type t, they obtain utility $t - C$. A college can choose an admissions rule and a grading policy. Although students types are unobservable, each student has an admissions exam score that is an observable unbiased estimator of their type. We model exam scores as being distributed as a unit variance Gaussian, centered at the student's type. An admissions policy for the school is a mapping between exam scores and admissions probabilities. We allow schools to set different admissions policies for different groups, but for most of our results, we require the natural condition that admissions probabilities within a group be monotonically non-decreasing in exam scores[2]. Deterministic monotone admissions policies simply correspond to setting admissions thresholds based on exam scores. For simplicity, in the body of the paper, we restrict attention to deterministic admissions rules, but in the Appendix, we extend our results to cover probabilistic admissions rules as well.

Schools may also set a grading policy. A grade is also modeled as a Gaussian centered at a student's true type, but the school may choose the variance of the distribution, for example, by controlling the number of conditionally independent evaluations that go into a student's grade. We assume that a student's grade is conditionally independent of her entrance exam score, conditioned on her type. One limiting extreme (infinite variance) corresponds to committing not to report grades at all. This limiting case is actually achievable because schools can simply opt not to share grades — in fact, this practice has been adopted at several top business schools [7]. At the other limiting extreme, types are perfectly observable. This extreme is generally not achievable, and we do not consider it in this paper. In between, the school can modulate the strength of the signal that employers get about student type, beyond the simple indicator that they were admitted to college.

Employers know the prior distributions P_i on student types, as well as the admissions and grading policy of the school. They are

[1] This is consistent with the signaling view of the role of colleges in the economics literature, beginning with [15]

[2] A non-monotone admissions rule would have the property that sometimes a student with a lower exam score would have a higher probability of admission that a student with a higher exam score. Non-monotonicity within a group is highly undesirable, because it would give some students a perverse incentive to intentionally try and lower their exam scores. If such incentives were present, it would no longer be reasonable to model exam scores as unbiased estimators of student types.

rational expectation maximizers. When deciding whether or not to hire a student, they will condition on all information available to them — a student'a group membership, the fact that she was admitted to college under the college's admissions policy, and the grade that she received under the college's grading policy — to form a posterior distribution about the student's type. They will hire exactly those students for whom they have positive expected utility under this posterior distribution.

In order to incentivize a particular employer to use a hiring rule that is independent of group membership, it is necessary to set admissions and grading policies such that for every student admitted to the school, and for every grade g that she may receive, the indicator that the conditional expectation of her type t is above the employer's hiring cost C is independent of the student's group membership. If there is uncertainty about what the employer's hiring cost C is, or if there are multiple employers, then it is necessary to guarantee this property for an interval of hiring costs $C \in [C^-, C^+]$ rather than for just a fixed cost. We distinguish these two cases. We call this property Irrelevance of Group Membership (IGM), in the single threshold and multiple threshold case respectively. A seemingly stronger property that we might desire is that the posterior distribution on student types conditional on admission to college is *identical* for both groups. We call this property *strong* Irrelevance of Group Membership (sIGM). Because it symmetrizes the two groups, it in particular guarantees that members of both groups will be treated identically by rational decision makers at any further stage down the decision making pipeline. It also is a natural goal in and of itself in a competitive market with employers who may offer different wages to different employees (which might here be modelled as differing costs C for employment). In such a market, employees will in equilibrium be offered the posterior expectation of their type as their wage — and so the sIGM property can be viewed as asking that for any two students with the same type, their expected wage conditional on being admitted to college should be identical, independent of their group. We show that in the presence of finite, nonzero variance in both exam scores and grades, IGM in the multiple threshold case implies sIGM. Finally, we say that an admissions rule and grading policy satisfy the *equal opportunity* condition, if a student's probability of making it all the way through the pipeline — i.e. being admitted to college *and then* being hired by the employer, is independent of her group conditioned on her type. Trivially, any group-symmetric admissions policy will satisfy both conditions if the two group type distributions are identical, so for the results that follow, we always assume that the group type distributions are distinct — differing in their mean, their variance, or both.

First, to emphasize that our impossibility results will crucially depend on the fact that exam scores are only a noisy signal of student ability, we consider the noiseless case, in which college admissions can be decided *directly* as a function of student type (this corresponds to the case in which exam scores have no noise). In this case, we can "have it all": there is a simple monotone admissions rule that guarantees both the equal opportunity condition, and satisfies IGM for multiple thresholds — for any grading policy that the school might choose. After establishing this simple result, in the rest of the paper we move on to the more realistic case in which exam scores are only a noisy signal of student type.

Next, we study what is possible if the college chooses to not report grades at all. In this case, we can also "have it all" — simply by setting a sufficiently high, group independent admissions threshold, a school can achieve both equal opportunity and IGM for multiple thresholds. This gives another view of the effects of practicing grade non-disclosure at highly selective schools [7].

Finally, in the bulk of the paper, we study the common case in which the college uses informative grades — i.e. sets the variance of its grade distribution to be some finite value. In this case, we show that it *is* possible to obtain IGM in the single threshold case, but that no monotone admissions rule can obtain sIGM. Because of the equivalence between sIGM and IGM for the multiple threshold case, this implies that no monotone admissions rule can obtain IGM in the multiple threshold case, even in isolation. Next, we consider the equal opportunity condition. One trivial way to obtain it is to simply admit nobody to college. We show that this is in general the only way in the multiple thresholds case: no non-zero monotone admissions rule can satisfy the equal opportunity condition, even in isolation.

1.3 Related Work

Our work fits into two streams of research. Within the recent line of work on algorithmic fairness, the most closely related work is that of Chouldechova [3] and Kleinberg, Mullainathan, and Raghavan [12]. Both of these papers prove the impossibility of simultaneously satisfying certain fairness desiderata in batch classification and regression settings. Broadly speaking, both papers show the impossibility of simultaneously equalizing false positive and false negative rates (related to our equal opportunity goal — see also [8]) and positive predictive value or calibration (related to our IGM goals). Our work is quite different, however: the goals that we study are not direct properties of the classification rule in question (in our case, the college admissions rule), but instead properties of its downstream effects. And while the work of [3, 12] shows the impossibility of simultaneously satisfying these fairness criteria, in our setting, we show that they are often impossible to satisfy even in isolation.

Our paper also fits into an older line of work studying economic models of discrimination and affirmative action, which has its modern roots in [1] and [14]. For example, Coate and Loury [5] and Foster and Vohra [6] study two stage models in which students from two different groups (who are a-priori identical) can in the first stage choose whether or not to make a costly investment in themselves, which will increase their value to employers. In the 2nd stage, employers may set a hiring rule that acts on a noisy signal about student quality. These works show the existence of a self-confirming equilibrium, in which only one group makes investments in themselves and are subsequently given employment opportunities, and consider interventions which can escape these discriminatory equilibria. These works can be viewed as studying the "upstream effects" of affirmative action policies, and explaining the mechanics by which different student populations may end up with different type distributions. The effect of the interventions proposed in these models is very slow, because it requires a new generation of students to recognize the opportunities made available to them via affirmative action policies and make costly investments in

their education in response, well before they enter the job market. In contrast, our work can be viewed as studying the "downstream effects" of these policies and examining shorter term effects which can be realized in a time frame that need not be long enough for type distributions to change.

More recently, the computer science community has begun studying fairness desiderata in dynamic models. Jabbari et al study the costs (measured as their effect on the rate of learning) of imposing fairness constraints on learners in general Markov decision processes [10]. Hu and Chen [9] study a dynamic model of the labor market similar to that of [5, 6] in which two populations are symmetric, but can choose to exert costly effort in order to improve their value to an employer. They study a two stage model of a labor market in which interventions in a "temporary" labor market can lead to high welfare symmetric equilibrium in the long run. Liu et al. [13] study a two round model of lending in which lending decisions in the first round can change the type distribution of applicants in the 2nd round, according to a known, exogenously specified function. They study how statistical constraints on the lending rule can improve or harm outcomes as compared to a myopic (i.e. ignoring dynamic effects) profit maximizing rule, and find that for two kinds of interventions, both improvement and harm are possible, depending on the details of how lending effects the type distribution. Finally, [11] studied the regulator's problem of providing financial incentives for a lender to satisfy fairness constraints in an online classification setting.

2 MODEL

We consider two populations of students, 1 and 2. In population $i \in \{1, 2\}$, each student has a type drawn from a Gaussian distribution $P_i = \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ with mean μ_i and variance σ_i^2. Since our problem is trivial if $P_1 = P_2$, in this paper we assume always that $P_1 \neq P_2$, i.e. the type distributions differ either in their mean, or their variance, or both. We denote by T_i the random variable that represents the type of a student from population i. Throughout the paper, ϕ denotes the probability density function and Φ the cumulative density function of a standard normal random variable with mean 0 and variance 1.

Each student takes a standardized test (SAT, etc.) and obtains a score given by

$$S_i = T_i + X$$

where X follows a normal distribution with mean 0 and variance 1, that does not depend on the population i, i.e., the student's score is a noisy but unbiased estimate of his type.

Additionally, we consider a university that admits students from both populations. The university designs an admission rule $A_i : \mathbb{R} \to [0, 1]$ for each population i, such that a student from population i with score s is accepted with probability $A_i(s)$. We also abuse notation and let A_i denote the binary random variable whose value is 1 if a student is accepted, and 0 otherwise. This admission rule is required to be monotone non-decreasing; i.e. an increase in exam score cannot lead to a *decrease* in admissions probability. We say that an admissions rule is deterministic if $A_i(s) \in \{0, 1\}$. A deterministic monotone admissions rule is characterized by a threshold β_i such that a student is accepted if and only if $S_i \geq \beta_i$. We call such rules"thresholding admissions rules". We focus on thresholding admissions rules in the body of this paper, but extend our results to probabilistic admissions rules to the Appendix. For simplicity of notation, we will often write $x_i(t) = \Pr[A_i = 1 | T_i = t]$ (Note that $x_i(t) = \Pr[S_i \geq \beta_i | T_i = t]$ in the deterministic case).

Every student who is admitted to the university receives a grade, given by:

$$G_i = T_i + Y$$

where Y follows a normal distribution with mean 0 and variance γ^2 that does not depend on the population i. γ can be set by the university, and represents the strength of the signal provided by a grading policy[3]. In our model, the University must commit to a single grading policy to use across groups.

Finally, an employer makes a hiring decision for each student that graduates from the university. The employer knows the priors P_i, the admission rules A_1, A_2 used by the school, the grading policy γ, and observes the grades of the students (as well as the fact that they were admitted to the school). The employer's expected utility for accepting a university graduate from population i with grade g is then given by

$$\mathbb{E}[T_i | G_i = g, A_i = 1] - C$$

where C is the cost for the employer to hire a student. The employer hires a university graduate from population i with grade g if and only if

$$\mathbb{E}[T_i | G_i = g, A_i = 1] \geq C$$

Throughout the paper, we study the feasibility of achieving the following fairness goals:

Definition 1 (Equal opportunity). *Equal opportunity holds if and only if the probability of a student being hired by the employer conditional on his type is independent of the student's group. I.e. if for all types $t \in \mathbb{R}$,*

$$\int_g \Pr[G_1 = g, A_1 = 1 | T_1 = t] \mathbb{1}\{\mathbb{E}[T_1 | G_1 = g, A_1 = 1] \geq C\} dg$$
$$= \int_g \Pr[G_2 = g, A_2 = 1 | T_2 = t] \mathbb{1}\{\mathbb{E}[T_2 | G_2 = g, A_2 = 1] \geq C\} dg$$

Definition 2 (Irrelevance of Group Membership). *Irrelevance of Group Membership (IGM) holds if and only if, conditional on admission by the school and on grade g, the employer's decision on whether to hire a student is independent of the student's group. I.e. if for all grades $g \in \mathbb{R}$,*

$$\mathbb{E}[T_1 | G_1 = g, A_1 = 1] \geq C \Leftrightarrow \mathbb{E}[T_2 | G_2 = g, A_2 = 1] \geq C$$

We further introduce a robust version of IGM, called *strong* Irrelevance of Group Membership, that symmetrizes the two populations and guarantees that members of both populations will be treated identically by rational decision makers at any further stage of the decision making pipeline.

Definition 3 (strong Irrelevance of Group Membership). *Strong Irrelevance of Group Membership (sIGM) holds if and only if, conditional on admission by the school and on grade g, the employer's*

[3] In actuality, of course, students receive many grades, not just one. But note that when one averages two normally distributed random variables, the result is also normally distributed, but with lower variance. Hence, one way to modulate the variance of a grade signal is to modulate the *number* of grades computed. The more assignments and exams that are graded, the lower the variance of the signal. The fewer that are graded, the higher the variance.

posterior on a student's type is independent of the student's population. I.e., for all $g \in \mathbb{R}$, for all $t \in \mathbb{R}$,

$$\Pr[T_1 = t | G_1 = g, A_1 = 1] = \Pr[T_2 = t | G_2 = g, A_2 = 1]$$

We note that sIGM holds if and only if the posterior on students' types conditional on admission by the school are identical:

Claim 1. *sIGM holds if and only if for all $t \in \mathbb{R}$:*

$$\Pr[T_1 = t | A_1 = 1] = \Pr[T_2 = t | A_2 = 1]$$

Proof. See Appendix. □

3 INFERENCE PRELIMINARIES

In this section, we derive some basic properties of the joint distributions on student types, exam scores, admissions rules, and grades that are relevant for reasoning about the employer's Bayesian inference task. We will draw upon these basic results in the coming sections.

3.1 Preliminaries on Gaussians and Multivariate Gaussians

First, we observe that together, student types, exam scores, and grades are distributed according to a multi-variate Gaussian.

Claim 2. *(T_i, S_i, G_i) follows a multivariate normal distribution.*

Proof. A set of random variables is distributed according to a multivariate normal distribution if every linear combination of the variables is distributed as a univariate normal distribution. For all $a, b, c \in \mathbb{R}$, $aT_i + bS_i + cG_i = (a + b + c)T_i + bX_i + cY_i$ follows a normal distribution as the sum of independent normal random variables. □

We now quote a basic fact about the conditional distribution that results when one starts with a multi-variate normal distribution, and conditions on the realization of a subset of its coordinates.

Claim 3. *Let $n \geq 2$ be an integer. Let $Z \in \mathbb{R}^n$ be a random variable following a multi-variate normal distribution. Let $Z = (Z_1, Z_2)$ where $Z_i \in \mathbb{R}^{n_i}$ with $n_1 + n_2 = n$. Suppose Z has mean $m = (m_1, m_2)$ where $m_i \in \mathbb{R}^{n_i}$, and covariance matrix*

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]$$

where $\Sigma_{ij} \in \mathbb{R}^{n_i \times n_j}$. Then $\mathbb{E}[Z_1 | Z_2 = z_2] = m_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - m_2)$ and $\mathrm{Var}[Z_1 | Z_2 = z_2] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Proof. See lecture notes [4]. □

The following technical lemma will also be useful for us.

Claim 4. *The hazard rate $H(x) = \frac{\phi(x)}{1-\Phi(x)}$ of a standard normal random variable is increasing, and satisfies*

$$\lim_{x \to -\infty} H(x) = 0,\ H(x) = x + o_{x \to +\infty}(1)$$

This is a commonly known result in the literature on probability theory and statistics. For completeness, we provide a proof in the Appendix.

3.2 Employer's First Moment Inference

The main lemma of this section characterizes the employer's Bayesian inference task when the college is using a threshold admissions rule: the posterior expectation of a student's type, conditioned on their exam score being sufficiently high to cross the admissions threshold, and on their observed grade. In the appendix, we give the corresponding inference rule for the employer when the college can use an arbitrary monotone admissions rule.

Lemma 1.

$$\begin{aligned}
&\mathbb{E}[T_i | S_i \geq \beta_i, G_i = g] \\
&= \frac{\gamma^2}{\sigma_i^2 + \gamma^2}\mu_i + \frac{\sigma_i^2}{\sigma^2 + \gamma^2} g \\
&+ \frac{\gamma^2 \sigma_i^2}{\sqrt{(\sigma_i^2 + \gamma^2)(\sigma_i^2 + \gamma^2 + \gamma^2\sigma_i^2)}} \cdot H\left(\frac{(\sigma_i^2 + \gamma^2) \cdot \beta_i - \gamma^2 \mu_i - \sigma_i^2 g}{\sqrt{(\sigma_i^2 + \gamma^2)(\sigma_i^2 + \gamma^2 + \gamma^2\sigma_i^2)}}\right)
\end{aligned}$$

where $H(x) = \frac{\phi(x)}{1-\Phi(x)}$ is the Hazard function of a standard normal random variable.

Proof. The proof is given in Appendix. □

A corollary of the previous lemma is that the posterior expectation computed by the employer will satisfy a number of nice regularity conditions which will be useful in proving our impossibility results:

Corollary 1. *$e_i(\mu_i, \sigma_i, \beta_i, g) = \mathbb{E}[T_i | S_i \geq \beta_i, G_i = g]$ is continuous, differentiable, and strictly increasing in each of μ_i, g and β_i. Further,*

$$\lim_{g \to -\infty} e(\mu_i, \sigma_i, \beta_i, g) = -\infty,$$
$$\lim_{g \to +\infty} e_i(\mu_i, \sigma_i, \beta_i, g) = +\infty,$$

and

$$\lim_{\beta_i \to -\infty} e(\mu_i, \sigma_i, \beta_i, g) = \frac{\gamma^2}{\sigma_i^2 + \gamma^2}\mu_i + \frac{\sigma_i^2}{\sigma^2 + \gamma^2} g,$$
$$\lim_{\beta_i \to +\infty} e(\mu_i, \sigma_i, \beta_i, g) = +\infty.$$

Proof. See Appendix. □

Finally, we define a quantity that will be useful to make reference to in a number of our forthcoming arguments: the minimum grade that results in a student from group i being hired by the employer, given a fixed admissions rule.

Definition 4 (Hiring threshold on grades). *We define $g_i^*(C) = \min\{g : \mathbb{E}[T_i | S_i \geq \beta_i, G_i = g] \geq C\}$ the inverse function of $g \to \mathbb{E}[T_i | S_i \geq \beta_i, G_i = g]$.*

By Corollary 1, $g_i^*(.)$ is a well-defined function on domain $\mathbb{R}$, and is continuous, differentiable, and strictly increasing.

3.3 Moments of the posterior distribution for monotone admission rules

The following lemma holds for the general case of monotone, randomized admission rules, and is useful in characterizing the moments of the distribution of types conditional on $A_i = 1$ and $G = g$ in population i:

Lemma 2. *Let $A_i(.)$ be a non-decreasing, non-zero, possibly randomized admission rule. For all $g \in \mathbb{R}$, $\mathbb{E}\left[T_i^k | G_i = g, A_i = 1\right]$ is finite and differentiable in g, and its derivative satisfies the following equation:*

$$\frac{\partial}{\partial g}\mathbb{E}_i\left[T_i^k \middle| G_i = g, A_i = 1\right] = \frac{1}{\gamma^2}\mathbb{E}_i\left[T_i^{k+1} \middle| G_i = g, A_i = 1\right] - \frac{1}{\gamma^2}\mathbb{E}_i\left[T_i^k \middle| G_i = g, A_i = 1\right] \cdot \mathbb{E}_i\left[T_i | G_i = g, A_i = 1\right].$$

Proof. The proof is given in Appendix. □

4 WHEN BOTH CONDITIONS ARE SATISFIABLE

In this section, we observe that there are two settings in which it is possible to "have it all" — satisfying both IGM and equal opportunity even in the multiple threshold case. The first setting is that of noiseless exam scores: when student types are perfectly observable by the school. The second setting is when the school opts not to report grades. We view the first setting as generally unrealisable, since any student evaluation will involve some degree of stochasticity. However the 2nd case — in which a school opts not to report grades — can be realized.

4.1 Noiseless Exam Scores (Observable Types)

First, we observe that if schools can perfectly observe student types (we have noiseless exam scores with $S_i = T_i$), then there is a simple threshold admissions rule that simultaneously achieves IGM and equal opportunity, even in the multiple threshold case. The ideas is simple: Given a range of employer costs $[C^-, C^+]$, the college simply sets an admissions threshold of C^+ or higher, using the same threshold for members of both groups. Because the threshold is the same for both groups, the probability of being admitted to college is a function only of type, and independent of group membership conditioned on type. Because scores were noiseless, admissions to college deterministically certifies that a student's type $t_i \geq C^+$, and so the employer chooses to hire everyone, independently of the grade they receive (and independently of their group membership). Hence, the probability of being hired is the same as the probability of being accepted to college, and is independent of group membership conditioned on type, and the employer's hiring rule is independent of group membership.

Claim 5. *Suppose $S_i = T_i$, i.e. a student's score perfectly reveals his type. Then for any hiring interval of hiring costs $[C^-, C^+] \in \mathbb{R}$, the non-zero admissions rule:*

$$A_i(s) = 1 \Leftrightarrow s \geq C^+$$

for both groups $i \in \{1, 2\}$ satisfies IGM and equal opportunity when paired with any *grading policy.*

Proof. See Appendix. □

Claim 6. *Suppose the school does not assign grades to students. Then for any hiring interval of hiring costs $[C^-, C^+] \in \mathbb{R}$, the non-zero thresholding admissions rule:*

$$A_i(s) = 1 \Leftrightarrow s \geq \beta$$

for both groups $i \in \{1, 2\}$ satisfies IGM and equal opportunity when β is large enough.

Proof. For β big enough, $\mathbb{E}[T_i | S_i \geq \beta] \geq C^+$ as $\lim_{\beta \to +\infty} \mathbb{E}[T_i | S_i \geq \beta] = +\infty$; this can be seen either by following the same steps as in the proof of Lemma 1 to obtain that

$$\mathbb{E}[T_i | S_i \geq \beta] = \mu_i + \frac{\sigma_i^2}{\sqrt{1+\sigma_i^2}} H\left(\frac{\beta_i - \mu_i}{\sqrt{1+\sigma_i^2}}\right)$$

which tends to $+\infty$ when $\beta_i \to +\infty$ by Claim 4. Another way of deriving this expression is by noting that not having a grade is equivalent to having an uninformative grade, i.e. to having $\gamma \to +\infty$. Now, let β be large enough such that in both populations, such that $\mathbb{E}[T_i | S_i \geq \beta] \geq C^+$. IGM immediately holds as every student that is accepted by the school is hired by the employer. Equal opportunity holds because the probability of a student with type t being hired by the employer is exactly the probability that he is admitted by the school (every student admitted by the school is hired by the employer), hence is given by

$$\Pr[S_i \geq \beta | T_i = t] = \int_{s \geq \beta} \phi(s - t) dt,$$

and is independent of the student's population. □

Note that this result is achieved by having the school set a very high admissions threshold (uniformly for both groups), and declining to give grades. Hence, declining to give grades may be a reasonable strategy for promoting our fairness goals in a highly selective school, but does not work when admissions thresholds must be lower. We note that the practice of grade witholding in MBA programs seems to be limited to the very top programs [7].

In the remainder of the paper we consider the case in which exam scores have positive finite variance, and in which the college uses a grading policy with positive finite variance. What will be possible will depend on whether we are in the single or multiple threshold case.

5 THE SINGLE THRESHOLD CASE

In this section, we consider what is possible when there is only a single employer with a hiring cost C that is known to the college. We show that in this case, IGM can always be achieved, but that it is impossible to achieve sIGM.

5.1 IGM can always be achieved

The main idea is as follows: For any grading scheme, and with a single threshold C in mind, the college can separately set different admissions thresholds β_1^* and β_2^* for the two groups respectively such that the posterior expectation for a student type from each group crosses the threshold of C at a grade g^*, which can be made to be the same for both populations. Since the only thing that matters in the employer's hiring decision is whether or not the student's expected type is above or below C, this is enough to cause the employer's hiring decision to be independent of group membership. The next lemma establishes that it is always possible to find such thresholds:

Lemma 3. *For any C in $\mathbb{R}$, there exists thresholds β_1^* and β_2^* and a grade g^* such that*

$$\mathbb{E}\left[T_1|G_1 = g^*, S_1 \geq \beta_1^*\right] = \mathbb{E}\left[T_2|G_2 = g^*, S_2 \geq \beta_2^*\right] = C$$

Proof. It follows by Corollary 1 that

$$\mathbb{E}\left[T_i|G_i = g, S_i \geq \beta_i\right]$$

is continuous in β_i and must reach any value between $\frac{\gamma^2}{\sigma_i^2+\gamma^2}\mu_i + \frac{\sigma_i^2}{\sigma_i^2+\gamma^2}g$ and $+\infty$. For g^* small enough, it must be the case that

$$\frac{\gamma^2}{\sigma_i^2+\gamma^2}\mu_i + \frac{\sigma_i^2}{\sigma_i^2+\gamma^2}g^* \leq C < +\infty,$$

hence there exists β_i^* such that

$$\mathbb{E}\left[T_i|G_i = g^*, S_i \geq \beta_i^*\right] = C.$$

□

Corollary 2. *Fix any C in $\mathbb{R}$. When the school uses thresholding admission rules with thresholds β_1^* and β_2^*, IGM holds for that C.*

Proof. $\mathbb{E}\left[T_i|G_i = g, S_i \geq \beta_i^*\right]$ is a strictly increasing function of g by Corollary 1 , therefore the employer accepts students from any population if and only if $g \geq g^*$ where g^* is population-independent, which proves the results. □

5.2 sIGM is impossible

We now show that strong IGM — making the posterior distributions for both groups identical — is impossible. In addition to its intrinsic interest, this result will be a key ingredient in our impossibility results for the multiple threshold setting.

Lemma 4. *Suppose the priors are distinct. For any two thresholds β_1 and β_2, there must exists $t \in \mathbb{R}$ such that*

$$\Pr\left[T_1 = t|S_1 \geq \beta_1\right] \neq \Pr\left[T_2 = t|S_2 \geq \beta_2\right]$$

I.e., sIGM cannot hold.

Proof. Let $x_i(t) = \Pr\left[S_i \geq \beta_i|T_i = t\right]$. Suppose for all $t \in \mathbb{R}$, sIGM holds, i.e.

$$\Pr\left[T_1 = t|S_1 \geq \beta_1\right] \neq \Pr\left[T_2 = t|S_2 \geq \beta_2\right]$$

by Claim 1. Then

$$\frac{x_1(t)\phi\left(\frac{t-\mu_1}{\sigma_1}\right)}{\Pr\left[S_1 \geq \beta_1\right]} = \frac{x_2(t)\phi\left(\frac{t-\mu_2}{\sigma_2}\right)}{\Pr\left[S_2 \geq \beta_2\right]}$$

hence

$$\frac{x_1(t)}{x_2(t)} = \frac{\sigma_1 \Pr\left[S_1 \geq \beta_1\right]}{\sigma_2 \Pr\left[S_2 \geq \beta_2\right]} \cdot \exp\left(\frac{(t-\mu_2)^2}{2\sigma_2} - \frac{(t-\mu_1)^2}{2\sigma_1^2}\right)$$

$x_1(.)$ and $x_2(.)$ are non-decreasing functions with values in $[0, 1]$, and $x_i(t) = \int_{s \geq \beta_i} \phi(s-t)ds$ is non-zero; therefore, $\lim_{t=+\infty} x_i(t)$ exists and is strictly positive. It must then be the case that $\frac{x_1(t)}{x_2(t)}$ has a finite and strictly positive limit in $+\infty$. On the other hand,

$$\exp\left(\frac{(t-\mu_2)^2}{2\sigma_2} - \frac{(t-\mu_1)^2}{2\sigma_1^2}\right) = K \exp\left(\frac{t^2}{2}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)t\right)$$

for some constant K. It is easy to see that the above quantity tends to either $+\infty$ or 0 as $t \to +\infty$ as long as either $\sigma_1 \neq \sigma_2$ or $\mu_1 \neq \mu_2$ (one of $\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}$ and $\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}$ must be non-zero). This leads to a contradiction. □

5.3 Equal opportunity

We defer the technical results of this section to the Appendix. Lemma 5 shows that for thresholding admission rules, IGM and equal opportunity cannot simultaneously hold for Gaussian priors with the same variance but different mean. This shows that obtaining fairness in the general case is significantly more difficult than in the simple cases in which the types are observable and the school does not assign grades. Lemma 6 shows that arguably stringent conditions on the grade accuracy and the thresholds set by the school must hold for equal opportunity to be possible. We conjecture that these conditions are, in general, impossible to satisfy, making equal opportunity impossible to satisfy even in isolation, in the single threshold case. As we will see in the next section, it is impossible to satisfy in the multiple-threshold case.

6 THE MULTIPLE THRESHOLD CASE

In this section, we turn to the multiple threshold case, which we view as the main setting of interest. In this case, we ask whether we can achieve IGM and equal opportunity not just with respect to a single known hiring cost C, but with respect to an entire interval of hiring costs $C \in [C^-, C^+]$. This will be the case when there are multiple employers, or simply when there is some uncertainty about the hiring threshold used by a single employer.

6.1 IGM is Impossible

In this section, we show that IGM is impossible to achieve even in isolation. The proof proceeds by showing that in the multiple threshold case, IGM must imply sIGM — i.e. that the posterior distributions conditional on admission to college are identical for both groups. Impossibility then follows from the impossibility of achieving sIGM (even for a single threshold), which we proved in the last section.

We first state a technical lemma, showing that if we satisfy IGM for every employer cost C in a continuous interval, we must actually be equalizing the posterior expected type across groups for every grade g in some other continuous interval.

Claim 7. *Let $\underline{C} < \bar{C}$. Suppose that for all $C \in (\underline{C}, \bar{C})$,*

$$\mathbb{E}\left[T_1|G_1 = g, S_1 \geq \beta_1\right] \geq C \Leftrightarrow \mathbb{E}\left[T_2|G_2 = g, S_2 \geq \beta_2\right] \geq C,$$

then it must be the case that for g in some interval (a, b),

$$\mathbb{E}[T_1|G_1 = g, S_1 \geq \beta_1] = \mathbb{E}[T_2|G_2 = g, S_2 \geq \beta_2]$$

Proof. Let $a = g_1^*(\underline{C})$ and $b = g_1^*(\bar{C})$ where $g_1^*(.)$ is the strictly increasing inverse of $g \to \mathbb{E}[T_1|G_1 = g, A_1 = 1]$ as per Claim 1 and Definition 4. Suppose there exists $g \in (a, b)$ such that

$$\mathbb{E}[T_1|A_1 = 1, G_1 = g] > \mathbb{E}[T_2|A_2 = 1, G_2 = g],$$

then for $C = \mathbb{E}[T_1|A_1 = 1, G_1 = g] \in (\underline{C}, \bar{C})$, it must be the case that

$$\mathbb{E}[T_1|A_1 = 1, G_1 = g] \geq C > \mathbb{E}[T_2|A_2 = 1, G_2 = g]$$

which contradicts the assumption of the claim. Now, suppose there exists $g \in (a, b)$ such that

$$\mathbb{E}[T_1|A_1 = 1, G_1 = g] < \mathbb{E}[T_2|A_2 = 1, G_2 = g],$$

then for $\epsilon > 0$ small enough, $C = \mathbb{E}[T_1|A_1 = 1, G_1 = g] + \epsilon \in (\underline{C}, \bar{C})$ and

$$\mathbb{E}[T_1|A_1 = 1, G_1 = g] < C \leq \mathbb{E}[T_2|A_2 = 1, G_2 = g]$$

which also contradicts the assumption of the claim. Therefore, it must be the case that for all $g \in (a, b)$,

$$\mathbb{E}[T_1|A_1 = 1, G_1 = g] = \mathbb{E}[T_2|A_2 = 1, G_2 = g]$$

□

We can now go on to prove the main theorem in this section:

Theorem 1. *Suppose the priors are distinct, then IGM cannot for all hiring costs $C \in (\underline{C}, \bar{C})$.*

Proof. By Claim 7, it must be the case that for all $g \in (a, b)$ for some interval (a, b), $\mathbb{E}[T_1|G_1 = g, S_1 \geq \beta_1] = \mathbb{E}[T_2|G_2 = g, S_2 \geq \beta_2]$. For all g in (a, b), by Lemma 2

$$\frac{\partial}{\partial g}\mathbb{E}[T_1|G_1 = g, S_1 \geq \beta_1] = \frac{\partial}{\partial g}\mathbb{E}[T_2|G_2 = g, S_2 \geq \beta_2]$$

and hence $E\left[T_1^2|G_1 = g, S_1 \geq \beta_1\right] = \mathbb{E}\left[T_2^2|G_2 = g, S_2 \geq \beta_2\right]$. It is easy to see that using the same argument by induction yields that for all integers k,

$$\mathbb{E}\left[T_1^k|G_1 = g, S_1 \geq \beta_1\right] = \mathbb{E}\left[T_2^k|G_2 = g, S_1 \geq \beta_1\right]$$

Since the distributions of types for the two populations conditional on $G_i = g, S_i \geq \beta_i$ admit a moment generating function (this follows immediately from the fact that P_i admits a moment generating function) and have identical moments, it must be that the distributions are the same for all $g \in (a, b)$. I.e., for all $g \in (a, b)$, we have

$$\Pr[T_1 = t|G_1 = g, S_1 \geq \beta_1] = \Pr[T_2 = t|G_2 = g, S_1 \geq \beta_1]$$

We have that in population i,

$$\Pr[T_i = t|G_i = g, S_i \geq \beta_i] = \frac{\Pr[T_i = t|S_i \geq \beta_i]\,\phi\left(\frac{g-t}{\gamma}\right)}{\int_t \Pr[T_i = t|S_i \geq \beta_i]\,\phi\left(\frac{g-t}{\gamma}\right)dt}$$

Note that $\int_t \Pr[T_i = t|S_i \geq \beta_i]\,\phi\left(\frac{g-t}{\gamma}\right)dt$ is a function of g only, that we will denote $p_i(g)$ from now on.

$$\Pr[T_1 = t|G_1 = g, S_1 \geq \beta_i] = \Pr[T_2 = t|G_2 = g, S_2 \geq \beta_2]$$

implies

$$\frac{\Pr[T_1 = t|S_1 \geq \beta_1]}{\Pr[T_2 = t|S_2 \geq \beta_2]} = \frac{p_1(g)}{p_2(g)}$$

for all $g \in (a, b)$ and $t \in \mathbb{R}$. $\Pr[T_1 = t|S_1 \geq \beta_1]$ and $\Pr[T_2 = t|S_2 \geq \beta_2]$ are both probability density functions that integrate to 1, so it must be the case that $\frac{p_1(g)}{p_2(g)} = 1$ and $\Pr[T_1 = t|S_1 \geq \beta_1] = \Pr[T_2 = t|S_2 \geq \beta_2]$. Therefore, sIGM must hold, which we have shown is impossible in Lemma 4. □

6.2 Equal opportunity cannot hold

Finally, we show that in the multiple threshold case, it is also impossible to satisfy the equal opportunity condition.

Theorem 2. *Suppose the priors are distinct. There exist no thresholding admission rules such that equal opportunity is guaranteed for all $C \in (\underline{C}, \bar{C})$, for any $\underline{C} < \bar{C}$.*

Proof. It is easy to see $x_i(t) = \int_s A_i(s)\phi(s-t)ds = \int_u A_i(u+t)\phi(u)du$ is monotone non-decreasing in t and non-zero. Remember

$$e_i(g) = \mathbb{E}[T_i|G_i = g, S_i \geq \beta_i]$$

has a strictly increasing and differentiable inverse $g_i^*(.)$ on $(-\infty, +\infty)$ by Corollary 1, and a student is hired by the employer if and only if $g \geq g_i^*(C)$. A student with type t in population i gets therefore hired with probability

$$\int_{g \geq g^*(C)} x_i(t)\phi\left(\frac{g-t}{\gamma}\right)dt = x_i(t)\left(1 - \Phi\left(\frac{g_i^*(C) - t}{\gamma}\right)\right)$$

equal opportunity then imply that $\forall t \in \mathbb{R}, C \in (\underline{C}, \bar{C})$,

$$\frac{x_1(t)}{x_2(t)} \cdot \left(1 - \Phi\left(\frac{g_1^*(C) - t}{\gamma}\right)\right) = \left(1 - \Phi\left(\frac{g_2^*(C) - t}{\gamma}\right)\right)$$

Taking the first order derivative in C of both sides of the above equation, we have that for all $C \in (\underline{C}, \bar{C})$, for all $t \in \mathbb{R}$,

$$\frac{x_1(t)}{x_2(t)} \cdot \frac{\frac{\partial g_1^*}{\partial C}(C)}{\frac{\partial g_2^*}{\partial C}(C)} = \frac{\phi\left(\frac{g_2^*(C)-t}{\gamma}\right)}{\phi\left(\frac{g_1^*(C)-t}{\gamma}\right)}$$

Suppose for some $C \in (\underline{C}, \bar{C})$, $g_1^*(C) \neq g_2^*(C)$. Without loss of generality, renumber the populations such that $g_2^*(C) > g_1^*(C)$. We have that

$$\frac{\phi\left(\frac{g_2^*(C)-t}{\gamma}\right)}{\phi\left(\frac{g_1^*(C)-t}{\gamma}\right)} = \exp\left(\frac{2\left(g_2^*(C) - g_1^*(C)\right)t + g_1^*(C)^2 - g_2^*(C)^2}{2\gamma^2}\right)$$

and we know that $g_i^*(.)$ is a strictly increasing function so $\frac{\partial g_2^*}{\partial C}(C) > 0$ so it must be the case that

$$\lim_{t \to +\infty} \frac{x_1(t)}{x_2(t)} = +\infty.$$

Since $x_1(t)$ is upper-bounded by 1, this implies in particular that $x_2(t) \to 0$ as $t \to +\infty$, which contradicts $x_2(.)$ being a non-zero, non-decreasing function. Hence, it must be the case that for all $C \in (\underline{C}, \bar{C})$, $g_1^*(C) = g_2^*(C)$, i.e. IGM holds. By Lemma 1, this is impossible. □

7 CONCLUSION

We consider two natural fairness goals that a college might have for its affirmative action policies: granting equal opportunity to individuals with the same type when graduating from high school, independent of their group membership, and incentivizing downstream employers to make hiring decisions that are independent of group membership. We show that these goals can be simultaneously achieved by highly selective colleges (i.e. those with very high admissions thresholds) — *but only if they do not report grades to employers*. This provides another view on this practice, which is followed by several highly selective MBA programs. On the other hand, we find that these goals are generally unachievable even in isolation if schools report informative grades. These impossibility results crucially hinge on the fact that exam scores and grades provide only *noisy* signals about student types, and hence require rational expectation maximizers to reason about prior type distributions, which can vary by group.

Our paper leaves open a natural technical question: can a college set admissions and informative grading policies to realize the equal opportunity condition, in the *single threshold* case? We conjecture that the answer to this question is *no*, and in the Appendix, we give a theorem supporting this conjecture — ruling out the possibility for deterministic admissions rules in every case except when the grading variance is exactly 1.

Finally, a natural question left open by our work is quantifying the extent to which *approximate* notions of our fairness goals are achievable, and *at what cost*. For example, can one guarantee that the ratio of the probabilities of a positive outcome between two students with the same type, but from different populations is close to 1? For students with the same grade? Given a constraint on the ratio, what is the most equal representation in the college class that we can guarantee for the two populations?

Acknowledgements

We thank Mallesh Pai and Jonathan Ullman for helpful discussions at an early stage of this work. We thank Jonathan Roth for pointing out an economic interpretation of our sIGM condition.

REFERENCES

[1] Kenneth Arrow et al. 1973. The theory of discrimination. *Discrimination in labor markets* 3, 10 (1973), 3–33.
[2] J Aislinn Bohren, Alex Imas, and Michael Rosenberg. 2017. The Dynamics of Discrimination: Theory and Evidence. (2017).
[3] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[4] Chuong. 2008. The Multivariate Gaussian Distribution.
[5] Stephen Coate and Glenn C Loury. 1993. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review* (1993), 1220–1240.
[6] Dean P Foster and Rakesh V Vohra. 1992. An economic argument for affirmative action. *Rationality and Society* 4, 2 (1992), 176–188.
[7] Daniel Gottlieb and Kent Smetters. 2011. *Grade non-disclosure.* Technical Report. National Bureau of Economic Research.
[8] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
[9] Lily Hu and Yiling Chen. 2018. A Short-term Intervention for Long-term Fairness in the Labor Market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW*, Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 1389–1398.
[10] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in Reinforcement Learning. In *International Conference on Machine Learning*. 1617–1626.
[11] Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. 2017. Fairness Incentives for Myopic Agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 369–386.
[12] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*.
[13] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *International Conference on Machine Learning*.
[14] Edmund S Phelps. 1972. The statistical theory of racism and sexism. *The american economic review* (1972), 659–661.
[15] Andrew Michael Spence. 1974. *Market signaling: Informational transfer in hiring and related screening processes*. Vol. 143. Harvard Univ Press.