# Access to Population-Level Signaling as a Source of Inequality

Nicole Immorlica
Microsoft Research
nicimm@gmail.com

Katrina Ligett
Hebrew University of Jerusalem
katrina@cs.huji.ac.il

Juba Ziani
California Institute of Technology
jziani@caltech.edu

## ABSTRACT

We identify and explore differential access to population-level *signaling* (also known as *information design*) as a source of unequal access to opportunity. A population-level signaler has potentially noisy observations of a binary type for each member of a population and, based on this, produces a signal about each member. A decision-maker infers types from signals and accepts those individuals whose type is high in expectation. We assume the signaler of the disadvantaged population reveals her observations to the decision-maker, whereas the signaler of the advantaged population forms signals strategically. We study the expected utility of the populations as measured by the fraction of accepted members, as well as the false positive rates (FPR) and false negative rates (FNR).

We first show the intuitive results that for a fixed environment, the advantaged population has higher expected utility, higher FPR, and lower FNR, than the disadvantaged one (despite having identical population quality), and that more accurate observations improve the expected utility of the advantaged population while harming that of the disadvantaged one. We next explore the introduction of a publicly-observable signal, such as a test score, as a potential intervention. Our main finding is that this natural intervention, intended to *reduce* the inequality between the populations' utilities, may actually *exacerbate* it in settings where observations and test scores are noisy.

## CCS CONCEPTS

• **Theory of computation** → *Algorithmic game theory*;

## KEYWORDS

Fairness; strategic signaling; information design; university admissions

**ACM Reference Format:**
Nicole Immorlica, Katrina Ligett, and Juba Ziani. 2019. Access to Population-Level Signaling as a Source of Inequality. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287579

## 1 INTRODUCTION

Settings where personal data drive consequential decisions, at large scale, abound—financial data determine loan decisions, personal history affects bail and sentencing, academic records feed into admissions and hiring. Data-driven decision-making is not reserved for major life events, of course; on a minute-by-minute basis, our digital trails are used to determine the news we see, the job ads we are shown, and the behaviors we are nudged towards.

There has been an explosion of interest recently in the ways in which such data-driven decision-making can reinforce and amplify injustices. One goal of the literature has been to identify the points in the decision-making pipeline that can contribute to unfairness. For example, are *data more noisy or less plentiful* for a disadvantaged population than for an advantaged one? Are the available *data less relevant* to the decision-making task with respect to the disadvantaged population? Has the disadvantaged population historically been *prevented or discouraged from acquiring good data profiles* that would lead to favorable decisions? Is the decision-maker simply *making worse decisions* about the disadvantaged population, despite access to data that could prevent it?

In this paper, we study *access to population-level signaling* as a source of inequity that, to our knowledge, has not received attention in the literature. We consider settings where the data of individuals in a population passes to a *population-level signaler*, and the signaler determines what function of the data is provided as a signal to a decision-maker. The signaler can serve as an advocate for the population by filtering or noising its individuals' data, but cannot outright lie to the decision-maker; whatever function the signaler chooses to map from individuals' data to signals must be fixed and known to the decision-maker.

Examples of population-level strategic signalers include high schools, who, in order to increase the chances that their students will be admitted to prestigious universities, inflate their grades, refuse to release class rankings [35], and provide glowing recommendation letters for more than just the best students. Likewise, law firms advocate on behalf of their client populations by selectively revealing information or advocating for trial vs. plea bargains. Even the choice of advertisements we see online is based on signals about us sold by exchanges, who wish to make their ad-viewing population seem as valuable as possible.

Our interest in asymmetric information in general and in population level strategic signaling in particular are inspired by the recent wave of interest in these issues in the economics literature (see Section 2 for an overview). In particular, the model we adopt to study these issues in the context of inequity parallels the highly influential work on Bayesian persuasion [28] and information design [5].

In order to explore the role that population-level strategic signaling can play in reinforcing inequity, we investigate its impact in a stylized model of university admissions.

Ligett's research was funded in part by the HUJI Cyber Security Research Center in conjunction with the Israel National Cyber Directorate (INCD) in the Prime Minister's Office, the Israeli Science Foundation, and by a DARPA Brandeis subcontract. Ziani's research was supported in part by NSF grants CNS-1331343 and CNS-1518941, the Linde Graduate Fellowship at Caltech, and the inaugural PIMCO Graduate Fellowship at Caltech.

*FAT* '19, January 29–31, 2019, Atlanta, GA, USA*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287579

We consider a setting in which a high school's information about its students is noisy but unbiased. Throughout, we call this noisy information *grades*, but emphasize that it may incorporate additional sources of information such as observations of personality and effort, that are also indicative of student quality. Importantly, all relevant information about student quality is observed directly by the school alone.

The school then aggregates each student's information into a signal about that student that is transmitted to the university. This aggregation method is called a *signaling scheme*, or informally, a (randomized) mapping from a student's information to a recommendation. A school could, for instance, choose to give the same recommendation for all its students, effectively aggregating the information about all students into one statement about average quality. Or, for example, the school could choose to provide positive recommendations to only those students that it believes, based on its information, to have high ability.

The university makes admission decisions based on these recommendations, with the goal of admitting qualified students and rejecting unqualified ones.[1] A school might make recommendations designed to maximize the number of their students admitted by the university. We call such a school *strategic*. Alternatively, a school might simply report the information it has collected on its students to the university directly. We call such a school *revealing*. As is common in economics, we assume that the university knows the signaling scheme chosen by the school (but does not know the realization of any randomness the school uses in its mapping). One justification typically given for such an assumption is that the university could learn this mapping over time, as it observes student quality from past years.

As expected, we find that strategic schools with accurate information about their students have a significant advantage over revealing schools, and, in the absence of intervention, strategic schools get more of their students (including unqualified ones) admitted by the university.

A common intervention in this setting is the standardized test. The university could require students to take a standardized test before being considered for admission, and use test scores in addition to the school's recommendations in an effort to enable more-informed admissions decisions. Intuitively, the role of the standardized test is that it "adds information back in" that was obfuscated by a strategic school in its recommendations, and so one might naturally expect the test to reduce inequity in the admissions process. While such a standardized test does increase the accuracy of admissions decisions, we show that when the test is a noisy estimate of student quality, it may in fact exacerbate the impact of disparities in signaling between schools.

*Summary of contributions.* We highlight access to strategic population level signaling, as studied in the economics literature, as a potential source of inequity. We derive the optimal signaling scheme for a school in Section 4.1 and compute the resulting school utility and false positive and negative rates in Section 4.2. We then show in Section 4.3 that disparities in abilities to signal strategically can constitute a non-negligible source of inequity. In Section 5, we study the effect of a standardized test that students must take before applying to the university, and highlight its limitations in addressing signaling-based inequity.

## 2 RELATED WORK

There is a large literature on individual-level signaling in economics, following on the Nobel-prize-winning work of Spence [38]. The general model there is quite different from our population-level signaling model; in the Spence model, *individuals* (not populations) invest in *costly* (in terms of money or effort) signals whose costs correlate with the individual's type. In that model, equilibria can emerge where high-type individuals are more likely to invest in the signal than low-types, which can result in the signal being useful for admissions or hiring.

Closer to our setting, Ostrovsky and Schwarz [35] study a model in which schools provide noisy information about their students to potential employers. Their focus is on understanding properties of the equilibria of the system; they do not fully characterize the equilibria, they do not consider the role of signaling in compounding inequity, and they do not investigate the impact of interventions like our standardized test. Unlike us, they do not consider the case where the schools have imperfect observations of the students' types. Such work falls into a broader literature on optimal information structures (e.g., [37]).

The impact of information asymmetries is a common theme in economics today, with key early work including Brocas and Carrillo [7]. Our model of signaling is inspired by the influential work on Bayesian Persuasion [28], where a persuader (played, in our model, by the school) commits to revealing some fixed function of the types of the population it serves; this revelation is used as the basis of a decision that impacts the welfare of both the decider and the persuader (and the persuader's constituents). The Bayesian Persuasion model has been applied to a variety of domains, e.g. [2, 4, 6, 16, 26, 34, 36, 39], and generalizations and alternatives to this model have been studied in [1, 3, 20, 21, 33, 37]. Recent work [10–14, 16, 22] has explored algorithmic aspects of persuasion settings. To our knowledge, ours is the first work to consider access to population-level signaling, Bayesian Persuasion, or information design as a source of inequity.

Recent work on fairness has highlighted a number of objectives that one might wish to enforce when allocating resources to or making decisions about large numbers of individuals. At a high level, these objectives tend to focus either on ensuring group-level fairness [8, 17, 19, 23, 24, 29, 31, 32, 40] or individual-level fairness [15, 27, 30]. The metrics we study—expected utility, false positive rates and false negative rates—are generally considered to be metrics of group fairness, but they also (coarsely) compare the extent to which similar individuals are being treated similarly.

One very interesting recent paper on fairness [25] does incorporate Spence-style individual-level signaling; in their model, a worker can choose whether and how much to invest in human capital, and this acts as an imperfect signal on whether the worker is qualified. Although their model and its implications are very different from ours, they similarly investigate the impact of upstream interventions on downstream group-level unfairness. Similar notions of individual-level signaling can also be found in [9, 18].

[1] In our simple model, the university does not have a fixed capacity, nor does it consider complementarities between students.

## 3 MODEL

We consider a setting with high schools (henceforth, "schools"), and a single university. A school has a population of students. Each student $i$ has a binary type $t_i \in \{0,1\}$ that represents the quality of the student. The students' types are drawn i.i.d. from a Bernoulli distribution with mean $p$; that is, a student has type 1 w.p. $p$ and 0 w.p. $1-p$. A student's type is private, that is, known to the student but unknown to both the school and the university. The prior $p$ is public and common knowledge to all agents.

A school observes noisy information about the types of each of its students. To formally model this, we assume student $i$ has a grade $g_i \in \{0,1\}$, which is observed by the school but is unknown to the university.

The grade $g_i$ for student $i$ is drawn as follows: $\Pr[g_i = 0|t_i = 0] = \Pr[g_i = 1|t_i = 1] = q$, for $q \in [1/2, 1]$.[2] That is, the student's type is flipped with some probability $1-q$. As $q$ increases, the grade $g_i$ becomes a more accurate estimate of the student's type $t_i$. The grade $g_i$ is known to the school but *not* the university. The distribution $q$ of the grade, however, is public, i.e., common knowledge to all parties.

A school has access to a (possibly trivial or uncountably infinite) set of signals $\Sigma$, and commits to a signaling scheme mapping grades $g$ to probability distributions over signals in $\Sigma$. For each student $i$, the university makes an accept/reject decision based on the distribution of the types $p$, the distribution of the grades $q$, and the realization of the signal chosen by the school. The goal of the university is to maximize the quality of the students it accepts.[3] In particular, we model the university as having additive utility over the set of students it accepts, with utility 1 for accepting a student of high type ($t_i = 1$), and utility $-1$ for a student with low type ($t_i = 0$). We assume that the university has unlimited capacity; therefore, the university accepts exactly those students who induce non-negative expected utility given the common priors and the signal.[4] We measure a school's utility by the expected fraction of its students who are admitted to the university. We note that this choice of utility measures the *access to opportunity* (defined as admittance to university) of the school's students. We refer to a school as *revealing* if it simply transmits the grade to the university as the signal. We refer to a school as *strategic* if it employs the optimal strategic signaling scheme, as examined in Section 4.1. A strategic school thus maximizes its expected utility.

In several places, we will discuss the distribution of students accepted by the university. To do so, it is useful to introduce the notions of *false positive* and *false negative* rates. The *false positive rate* of a school is the (expected) probability that a student with type 0 is accepted by the university. The *false negative rate* of a school is the (expected) probability that a student with type 1 is rejected by the university.

We introduce several assumptions that restrict our attention to settings of interest. First, we assume the expected quality of a student is negative, such that the university would reject students without any signal from the school.

Assumption 1. *The university's expected utility for accepting any given student, absent any auxiliary information, is negative, i.e., $p-(1-p) < 0$, and therefore $p < 1/2$.*

Next we assume the university's expected utility of accepting a student with a high (resp. low) grade is positive (resp. negative).

Assumption 2. *The university has non-negative expected utility for accepting a student with a high grade, and negative expected utility for accepting a student with a low grade:*

$$\Pr[t=1|g=1] - \Pr[t=0|g=1] \geq 0;$$
$$\Pr[t=1|g=0] - \Pr[t=0|g=0] < 0.$$

*These can be rewritten as:*

$$pq - (1-p)(1-q) \geq 0;$$
$$p(1-q) - (1-p)q < 0.$$

We note that if the expected utility of accepting a student with a high grade were negative, then none of the school's students would be admitted by the university under any signaling scheme. On the other hand, if the expected utility of accepting a student with a low grade were positive, then the university would always accept every student.[5] Thus, this assumption restricts our analysis to the regime in which the utilities of revealing and strategic schools may differ.

The following easy consequence of these assumptions will be useful in our analysis.

Observation 3. *Under Assumption 1, Assumption 2 implies $q \geq 1-p$.*

We conclude with the following well-known result (see, e.g., Kamenica and Gentzkow [28]) that an optimal signaling scheme contains, without loss of generality, at most as many signals as there are actions available to the decision-maker. In our setting, this corresponds to restricting $|\Sigma| = 2$ as the university makes an accept/reject decision for each student.

The result, reproduced below for our setting, follows from a revelation-principle type argument. The idea is to replicate the utilities of a signaling scheme with many signals by first producing a signal according to the original scheme and then simply reporting to the university, as a signal in the simplified scheme, the action $\sigma^+ =$ *accept* or $\sigma^- =$ *reject* that it would choose to take as a result of seeing the original signal.

Theorem 4 (Kamenica and Getzkow [28]). *Suppose $\Sigma$ is a measurable (but potentially uncountable) set with at least two elements. Let $\Sigma'$ be such that $|\Sigma'| = 2$. Given any original signaling scheme mapping to $\Delta(\Sigma)$, there exists a new signaling scheme mapping to $\Delta(\Sigma')$ that induces the same utilities for the school and the university as those induced by the original scheme. Further, one can write $\Sigma' = \{\sigma^-, \sigma^+\}$ such that a student with signal $\sigma^+$ is accepted by the university with probability 1, and a student with signal $\sigma^-$ is rejected with probability 1.*

When $|\Sigma| = 1$, signals carry no information, making mute the question of access to signaling schemes. Therefore, throughout the paper, we make the assumption that $|\Sigma| = 2$ and denote its elements by $\Sigma = \{\sigma^+, \sigma^-\}$. This is without loss of generality, by the argument above.

[2] The assumption that $q \geq 1/2$ is without loss of generality; when $q < 1/2$, one can set $q = 1-q$, $g_i = 1-g_i$ and all results carry through by symmetry.

[3] There is no notion here of students "applying" to the university or not; the university considers *all* students for admission.

[4] When indifferent, the university accepts the student.

[5] In fact, this condition is already ruled out by Assumption 1.

# 4 THE IMPACTS OF SIGNALING SCHEMES

The goal of this paper is to highlight the role of access to strategic signaling in creating unequal access to opportunity and explore the intervention of a standardized test as a way to combat this inequity. In order to do so, we first formulate optimal signaling schemes, and then we study their impact on students and their relationship to noisy grades.

## 4.1 Optimal signaling scheme

We first derive the optimal signaling scheme. The idea is to pack low-quality students together with high quality students by giving both the *accept* signal $\sigma^+$. A school is limited in the extent to which it can do so, as it must ensure the university obtains non-negative expected utility by accepting all the students who have signal $\sigma^+$. The following theorem provides the right balance.

Theorem 5. *The optimal signaling scheme for a school is*

$$\Pr\left[\sigma^+ |\, g = 0\right] = \frac{p+q-1}{q-p}$$
$$\Pr\left[\sigma^+ |\, g = 1\right] = 1.$$

Proof. As per the revelation principle in Theorem 4, we can let $\sigma^+$ be a signal such that all students with that signal are accepted by the university, and $\sigma^-$ a signal such that all students with that signal are rejected. Conditional on $\sigma^+$, we can write the probabilities that a student is of each type as

$$\begin{aligned}\Pr[t=1|\sigma^+] =& \frac{\Pr\left[t=1,\sigma^+\right]}{\Pr\left[\sigma^+\right]}\\ =& \Pr[t=1]\cdot\frac{\Pr[\sigma^+|t=1]}{\Pr[\sigma^+]}\\ =& \Pr[t=1]\cdot\frac{\Pr[\sigma^+|g=1]\Pr[g=1|t=1]}{\Pr[\sigma^+]}\\ &+\Pr[t=1]\cdot\frac{\Pr[\sigma^+|g=0]\Pr[g=0|t=1]}{\Pr[\sigma^+]}\\ =& p\cdot\frac{q\Pr[\sigma^+|g=1]+(1-q)\Pr[\sigma^+|g=0]}{\Pr\left[\sigma^+\right]}\end{aligned}$$

and, similarly,

$$\Pr[t=0|\sigma^+] = (1-p)\cdot\frac{(1-q)\Pr[\sigma^+|g=1]+q\Pr[\sigma^+|g=0]}{\Pr\left[\sigma^+\right]}.$$

The university's expected utility when accepting all those students with signal $\sigma^+$ is non-negative if and only if such a student is at least as likely to be of type 1 as of type 0, that is, $\Pr[t=0|\sigma^+] \leq \Pr[t=1|\sigma^+]$. Plugging in and rearranging, this gives the constraint

$$\begin{aligned}&\Pr[\sigma^+|g=0]\cdot(q(1-p)-p(1-q))\\ &\qquad\leq \Pr[\sigma^+|g=1]\cdot(pq-(1-q)(1-p)).\end{aligned}$$

Recall that $q(1-p)-p(1-q) > 0$ by Assumption 2, and thus the constraint can be rewritten as

$$\begin{aligned}\Pr[\sigma^+|g=0] &\leq \frac{pq-(1-q)(1-p)}{q(1-p)-p(1-q)}\cdot\Pr[\sigma^+|g=1]\\ &= \frac{p+q-1}{q-p}\cdot\Pr[\sigma^+|g=1].\end{aligned}$$

The school's expected utility is

$$\Pr[\sigma^+] = \Pr[\sigma^+|g=0]\Pr[g=0]+\Pr[\sigma^+|g=1]\Pr[g=1].$$

Since $\Pr[\sigma^+|g=1]$ is unconstrained, the school's utility is maximized by setting it to 1. The school's utility is, similarly, maximized by maximizing the value of $\Pr[\sigma^+|g=0]$, which, given the constraint, occurs by setting

$$\begin{aligned}\Pr[\sigma^+|g=0] &= \frac{p+q-1}{q-p}\cdot\Pr[\sigma^+|g=1]\\ &= \frac{p+q-1}{q-p}.\end{aligned}$$

□

## 4.2 School's utility, false positive and false negative rates

In this section, we calculate the expected utility, false positive, and false negative rate achieved by a school, depending on the accuracy of its grades and whether it uses the optimal strategic signaling scheme when transmitting information about its students to the university. These lemmas will form the basis of our evaluation of the impacts of strategic signaling, in Section 4.3. Recall that we refer to a school that does not strategically signal and instead transmits its raw grades to the university as *revealing*.

The proofs of the following Lemmas follow by direct calculations. We provide an exposition of the more involved calculations of Lemmas 7 and 9 in the Appendix.

Lemma 6 (Revealing school's utility). *The expected utility $U_r(p,q)$ of a revealing school is*

$$U_r(p,q) = pq+(1-p)(1-q).$$

*For the special case of a revealing school with accurate grades (when $q=1$), we have*

$$U_r(p,1) = p.$$

A revealing school gets exactly the students with high grades accepted, as per Assumption 2; in particular, a $q$ fraction of high-type students will have a high grade and be accepted, while a $(1-q)$ fraction of the low-type students will be accepted.

Lemma 7 (Strategic school's utility). *A school's expected utility $U_s(p,q)$ when it signals strategically is given by*

$$U_s(p,q) = 1+(p+q-2pq)\cdot\frac{2p-1}{q-p}.$$

*For the special case of a strategic school with accurate grades (when $q=1$), we have*

$$U_s(p,1) = 2p.$$

A school that signals strategically gets exactly those students with a signal of $\sigma^+$ accepted, as per the revelation principle argument of Theorem 4; a student with a high grade will be accepted with probability $\Pr\left[\sigma^+|\, g=1\right]$ and a student with a low grade with probability $\Pr\left[\sigma^+|\, g=0\right]$, with the probabilities chosen according to Theorem 5.

Lemma 8 (Revealing school's FPR/FNR). *When a school is revealing, the false positive rate is given by*

$$FPR_r(p,q) = 1-q$$

*and the false negative rate by*

$$FNR_r(p, q) = 1 - q.$$

*For the special case of a revealing school with accurate grades (when $q = 1$), we have $FPR_r(p, 1) = FNR_r(p, 1) = 0$.*

In the case of a revealing school, a low-type (resp. high-type) student obtains a low (resp. high) grade and gets rejected (resp. accepted) with probability $1 - q$, i.e., if the grade does not match the type.

Lemma 9 (Strategic school's FPR/FNR). *When a school signals strategically, the false positive rate is given by*

$$FPR_s(p, q) = 1 - q + q \cdot \frac{p + q - 1}{q - p}$$

*and the false negative rate by*

$$FNR_s(p, q) = (1 - q)\frac{1 - 2p}{q - p}.$$

*For the special case of a strategic school with accurate grades (when $q = 1$), we have $FPR_s(p, 1) = \frac{p}{1-p}$ and $FNR_s(p, 1) = 0$.*

In the case of a school that signals strategically according to Theorem 5, a low-type student gets accepted with probability $\Pr\left[\sigma^+ | \, g = 1\right] = 1$ if his grade is 1 (which occurs with probability $1 - q$), and probability $\Pr\left[\sigma^+ | \, g = 0\right]$ if his grade is $g = 0$ (which occurs with probability $q$). On the other hand, a high-type student gets rejected when his signal is $\sigma^-$; because $\Pr\left[\sigma^+ | \, g = 1\right] = 1$, this happens only when $g = 0$ and the signal is $\sigma^-$, i.e. with probability $\Pr\left[\sigma^- | \, g = 0\right] \Pr[g = 0 | t = 1]$.

*Remark.* While we chose to focus on average population (i.e., school) utility in this paper, because of space constraints, one can use these derivations of FRP and FNP to calculate the welfare of subpopulations, such as low-type students at a revealing school, which then implies population-level utility comparisons as well. One interesting observation is that, using the above Lemmas and Assumptions 1 and 2, one can see that the FPR of a strategic school is *larger* and the FNR *smaller* than that of a revealing school. Thus, while it is intuitively obvious that low-type students prefer a strategic school, these calculations show that high-type students also prefer a strategic school (and the preference is strict unless the assumptions hold with equality).

### 4.3 Consequences of strategic signaling for access to opportunity

In this section, we quantify the impact of access to strategic signaling and its interaction with accuracy of the information (grades) on which the signals are based. We study both the resulting expected utility of a school as well as the resulting acceptance rates of both types of students. We find that the ability to strategically signal always has a positive (although bounded) impact, increasing students' acceptance rates and the school's expected utility. The benefit of strategic signaling for both students and the school improves (boundedly so) with the accuracy of the grades, whereas a revealing school and its students receive (potentially dramatically) higher expected utility from noisy grades. The following theorem is a direct consequence of Lemmas 18, 19, 20, 21, 22 in the Appendix.

Theorem 10. *For all $p < 1/2$ and $q > q' \geq 1 - p$, the following hold:*

- *accuracy in grades benefits strategic schools,*

$$\frac{1}{1-p} U_s(p, q') \geq U_s(p, q) \geq U_s(p, q');$$

- *strategic schools have higher expected utility than revealing schools,*

$$2U_r(p, q) \geq U_s(p, q) \geq U_r(p, q);$$

- *and accuracy in grades harms revealing schools,*

$$2(1 - p)U_r(p, q) \geq U_r(p, q') \geq U_r(p, q).$$

*Further, all above bounds are tight for some $q, q'$.*

We see that, perhaps counter-intuitively, adding noise to the grades can help a revealing school get more students admitted, up to a point.[6] This follows from the fact that adding noise to the grade increases the number of students with a high grade overall, by Assumption 1, as there are more low-type students (whose representation increases as grade accuracy decreases) than high-type students (whose representation decreases as grade accuracy decreases). Adding noise to grades is, however, a blunt instrument, in that it drives up both false negatives and false positives (see Lemma 8), which limits its utility benefits. The ability to signal strategically is more subtle, driving up false positives (and expected utility), at no cost of false negatives. The power of strategic signaling is maximized when schools have access to highly accurate grades. Accurate information, the ability to control the noise level of that information, and, most notably, the ability to strategically signal about that information, therefore constitute powerful drivers of unequal access to opportunity in settings where key information is transmitted to a decision-maker on behalf of a population.

We can derive comparisons resulting in similar insights for the false positive and false negative rates of revealing and strategic schools (see Appendix).

## 5 INTERVENTION: STANDARDIZED TEST

The prior sections show that unequal access to strategic signaling can result in unequal access to opportunity. This is driven by high error rates for students accepted from schools with signaling technologies and/or noisy grades. The university has a vested interested in decreasing this error rate as it harms the university's utility. In addition, an outside body or the university itself might be concerned about the resulting unequal access to opportunity. In this section, we explore the impact of a common intervention: the standardized test. While availability of a test score certainly can only improve the expected utility of the university,[7] we find that it has an ambiguous effect on the inequity. In particular, for a large range of parameter settings, the introduction of a test can *increase* the inequality in access to opportunity.

[6] A similar observation in a somewhat different setting was made in work of Ostrovsky and Schwarz [35].

[7] This is because the expected utility of the university from strategic schools without test scores is zero, and so can only increase. For revealing schools, the university gets strictly more information with test scores and hence more utility.

### 5.1 Augmented model

Throughout this section, we augment the model of Section 3 to add the requirement that each student must take a test, and the results of that test are visible both to the student's school and to the university. (The school may then incorporate the test results into its subsequent strategic behavior.)

We model the test score $s_i \in \{0,1\}$ of student $i$ as a noisy estimate of $t_i$, conditionally independent from the grade $g_i$, obtained as follows: $\Pr[s_i = 0|t_i = 0] = \Pr[s_i = 1|t_i = 1] = \delta$, for $\delta \in [1/2, 1]$.[8] The score $s_i$ is public, i.e., the school and the university both observe it.

A school has access to a set of signals $\Sigma$ as before, but now can design a signaling scheme $\sigma : \{0,1\} \times \{0,1\} \to \Delta(\Sigma)$ that is a function of both the student's grade and his test score; i.e., the school designs $\Pr\left[\sigma|\, g_i, s_i\right]$ for $\sigma \in \Sigma$. The university again makes accept/reject decisions that maximize its expected utility, but now the university has access to the test score $s_i$ and its distribution $\delta$ as well as the signal and the distributions $p$ and $q$. As before, a *strategic* school chooses a signaling scheme that maximizes the fraction of students accepted whereas a *revealing* school simply transmits the grade to the university as the signal.

As in Section 3, we introduce an assumption controlling the noise $\delta$ of the test.

Assumption 11. *The university has non-negative expected utility for accepting a student with a high test score, and negative expected utility for accepting a student with a low test score:*

$$0 \le p\delta - (1-p)(1-\delta)$$
$$0 > p(1-\delta) - (1-p)\delta.$$

We note that if the expected utility of accepting a student with a high test score were negative, or the expected utility of accepting a student with a low test score were positive, then in the absence of signals, the university would always accept either none or all of the students. Note that regimes when the standardized test is uninformative on its own but becomes informative when coupled with grades may still be interesting. However, even under Assumption 11, which excludes certain parameter ranges from consideration, we have a rich enough model to illustrate our main findings. In the Appendix, we show how to relax this assumption, and how doing so affects the optimal signaling scheme.

The following consequence will be useful in our analysis.

Observation 12. *Under Assumption 1, Assumption 11 implies*

$$\delta \ge 1 - p.$$

Fixing $p$, we denote by $u_{q,\delta}(g,s)$ the expected utility the university derives from admitting a student with score $s$ and grade $g$:

$$u_{q,\delta}(g,s) := \Pr[t_i = 1|\, g,s] - \Pr[t_i = 0|\, g,s].$$

When $\delta = q = 1$, $u_{q,\delta}(s,g)$ is not defined for $s \neq g$ as in this case $s$ and $g$ are perfectly correlated. For notational convenience, we define $u_{q,\delta}(s,g) = -1$ in these cases.

Lemma 13. *Assumptions 2 and 11 together imply that the university receives non-negative expected utility from accepting a student with both a high grade and a high score, and negative expected utility from a student with both a low grade and a low score:*

$$u_{q,\delta}(1,1) \ge 0 > u_{q,\delta}(0,0).$$

*This can be rewritten as*

$$pq\delta - (1-p)(1-q)(1-\delta) \ge 0;$$
$$p(1-q)(1-\delta) - (1-p)q\delta < 0.$$

Theorem 4 (the revelation principle) also holds in this setting, and so we assume for the remainder of this section that $\Sigma = \{\sigma^-, \sigma^+\}$, without loss of generality.

### 5.2 Optimal signaling

We first derive the optimal strategic signaling scheme. Again, a school would like to pack low-quality students together with high quality students, but is now limited in its ability to do so by their test scores. If the expected utility the university receives from a student with a high grade but low test score is negative ($u_{q,\delta}(1,0) < 0$), then this student (and in fact any student with a low test score) will be rejected regardless of the signal from the school. Otherwise ($u_{q,\delta}(1,0) \ge 0$), the school can signal to the university to accept such a student, and can additionally pack in some low-grade-low-score students, subject to maintaining non-negative expected utility for the university.

Theorem 14. *The optimal signaling scheme for a school with access to grades and a test score, under Assumption 11, is*

$$\Pr\left[\sigma^+|\, g = 1, s = 1\right] = 1$$
$$\Pr\left[\sigma^+|\, g = 0, s = 1\right] = 1$$
$$\Pr\left[\sigma^+|\, g = 1, s = 0\right] = \begin{cases} 1, & \textit{if}\, u_{q,\delta}(1,0) \ge 0 \\ 0, & \textit{if}\, u_{q,\delta}(1,0) < 0 \end{cases}$$
$$\Pr\left[\sigma^+|\, g = 0, s = 0\right] = \begin{cases} \frac{pq(1-\delta)-(1-p)(1-q)\delta}{(1-p)q\delta-p(1-q)(1-\delta)}, & \textit{if}\, u_{q,\delta}(1,0) \ge 0 \\ 0, & \textit{if}\, u_{q,\delta}(1,0) < 0 \end{cases}$$

We defer the proof to the Appendix.

### 5.3 School's utility, false positive and false negative rates

In this section, we calculate the expected utility achieved by both a strategic school and a revealing school as a function of the type distribution, the accuracy of its grades, and the accuracy of the standardized test score. We defer all proofs to the Appendix.

For a revealing school, the university always accepts high-grade high-score students. If high grades are more informative than low test scores (that is, if $u_{q,\delta}(1,0) \ge 0$, which depends on $p$ as well as $q$ and $\delta$ and happens, for instance, if $p = 1/4$, $q = 9/10$, and $\delta = 7/10$), then the university also accepts students with low test scores, benefiting the school. Alternatively, if high test scores are more informative than low grades (i.e., $u_{q,\delta}(0,1) \ge 0$), then the university also accepts students with low grades. These conditions provide additional boosts to the utility of a revealing school.

Lemma 15 (Revealing school's utility). *The expected utility $U_r(p,q,\delta)$ of a revealing school with access to grades and a test score*

[8] The assumption that $\delta \ge 1/2$ is, as with our analogous assumption about the grades, without loss of generality.

*is*

$$\begin{aligned} U_r(p,q,\delta) &= pq\delta + (1-p)(1-q)(1-\delta) \\ &\quad + \mathbb{1}\left[u_{q,\delta}(1,0) \geq 0\right] \left(pq(1-\delta) + (1-p)(1-q)\delta\right) \\ &\quad + \mathbb{1}\left[u_{q,\delta}(0,1) \geq 0\right] \left(p(1-q)\delta + (1-p)q(1-\delta)\right). \end{aligned}$$

*For the special case of a revealing school with accurate grades (when* $q = 1$*), we have*

$$U_r(p,1,\delta) = p.$$

As illustrated in Figure 1, for fixed $p$ and $\delta$, $U_r(p,q,\delta)$ may not be a decreasing function of $q$. In fact, when $q$ is small enough, the grades are completely uninformative and the university only admits students with a test score of 1. In that regime, the expected utility for a revealing school is therefore constant in $q$. For intermediate values of $q$, the grades are still uninformative on their own but are informative coupled with a high standardized test score; at this point, only students with both a score and a grade of 1 get admitted by the university, and the school's expected utility suddenly drops when compared to smaller $q$. The school's expected utility in that regime is increasing in $q$ as, under Assumption 11, increasing the value of $q$ increases the fraction of students with both high scores and high grades. Finally, when $q$ is large enough, the grades are significant enough on their own that only students with high grades are admitted; this leads to a jump in expected utility compared to the intermediate regime. In this regime for high values of $q$, the school's expected utility is decreasing as a result of the fact that increasing the value of $q$ now decreases the number of students with a high grade by Assumption 1, as seen in Section 4.3.

Lemma 16 (Strategic school's utility). *The expected utility* $U_s(p,q)$ *when a school signals strategically and* $u_{q,\delta}(1,0) < 0$ *is*

$$U_s(p,q,\delta) = p\delta + (1-p)(1-\delta);$$

*when* $u_{q,\delta}(1,0) \geq 0$*, the expected utility is*

$$\begin{aligned} U_s(p,q,\delta) &= (1 - p(1-q)(1-\delta) - (1-p)q\delta) \\ &\quad + (p(1-q)(1-\delta) + (1-p)q\delta)\,\frac{pq(1-\delta) - (1-p)(1-q)\delta}{(1-p)q\delta - p(1-q)(1-\delta)}. \end{aligned}$$

*For the special case of a strategic school with accurate grades (when* $q = 1$*), we have*

$$U_s(p,1,\delta) = 1 - \delta + p.$$

The expected utility of a strategic school is, unsurprisingly, monotone in $q$ (as illustrated in Figure 1), as higher-quality information about its students' types allows the school to signal more effectively. For small and intermediate values of $q$ (i.e., insignificant grades), the university bases admission decisions solely on the standardized test score and only admits students with a score of 1 (it has positive expected utility from doing so, by Assumption 11); in this regime, a strategic school's expected utility is hence constant. When $q$ becomes large enough, i.e., when the grades are significant enough, the university starts having positive expected utility from admitting students with a high grade even if they have a low score, and the school can start bundling these students together with the high score students, leading to a jump in its expected utility. The plotted parameters for the figures are chosen to satisfy Assumptions 1, 2 and 11; the discontinuities occur at $q$ such that $u_{q,\delta}(0,1) = 0$ and $u_{q,\delta}(1,0) = 0$.

We also calculate the false positive and false negative rates of strategic and revealing schools; we defer this derivation to the Appendix.

## 5.4 Impact of Standardized Test

With a perfect standardized test or, in fact, a sufficiently good one, (i.e., high enough $\delta$), it is not hard to see that the university accepts exactly those students with a high test score from strategic as well as revealing schools. Thus, no matter the accuracy of the grades or distribution of types, the standardized test results in equal expected utility, and hence equal access to opportunity, for revealing and strategic schools (see Appendix for details). Similarly, if grades are accurate (i.e., $q = 1$), then a revealing school's expected utility is fixed at $p$ whereas a strategic school's expected utility is only diminished (from $2p$ without the test) by the extra constraints introduced by a standardized test. Thus, in this case as well, a standardized test decreases the inequality between the utilities of a strategic and a revealing school, making the ratio of utilities less than 2 (see Appendix for details).

Figure 2 plots $U_s(p,q)/U_r(p,q)$, with and without test scores, as a function of $q$, for $p = 0.35$ and different values of $\delta$. The form of the utility ratio between a strategic and a revealing school in the absence of a test score follows from the fact that both utilities are continuous, and that the expected utility of a strategic school increases while that of a revealing school decreases in $q$, as we have seen in Section 4.3. The form in the presence of a test score can be explained as follows. First, when in the regime of small values of $q$, only students with a high standardized test score are admitted by the university, in which case admission decisions do not depend on how the schools act and both the strategic school and the revealing school have the same expected utility, leading to a ratio of 1. For intermediate values of $q$, we have previously discussed that the utility for a strategic school remains constant (the university still has positive utility for students with a score of 1 and the strategic school can bundle all such students together, regardless of grade), while the utility for a revealing school suddenly drops (only students with both a high grade and a high score are admitted) and is increasing in $q$, explaining the sudden drop in ratio of utilities at the change of regime, and the decreasing monotonicity of the ratio in $q$ within the intermediate regime. When $q$ becomes large enough, we have seen that both the revealing and the strategic school experience a jump in utilities, which explains the second discontinuity in the ratio of utilities. Because the revealing school has significantly lower utility than the strategic school for intermediate values of $q$, the relative jump in the utility of a revealing school is higher than the relative jump in utility of a strategic school. Because in the regime with high values of $q$, the utility of a strategic school is increasing and that of a revealing school is decreasing, the ratio of utilities is increasing.

Interestingly, we observe that the introduction of a standardized test does not always decrease inequity. For noisy grades, when the test score is also sufficiently noisy, the test may have the effect of increasing the ratio of utilities between a strategic school and a revealing school. This is clearly illustrated in Figure 2, where the curve with test scores sometimes lies above that without a test.

Some intuition for this result is as follows. In the regime for intermediate values of $q$, as $q$ becomes more and more inaccurate, the ratio of utilities in the presence of a standardized test increases and eventually overtakes the ratio in the absence of a standardized test (which decreases to 1 as the grades become more inaccurate). In the regime for high values of $q$, the university admits students with a high grade only, independently of what their standardized test scores are; therefore, the utility of a revealing school is the same with or without a standardized test. On the other hand, when the standardized test score becomes more inaccurate, the strategic school can take advantage of the noise in said score to bundle in more students than if there was no standardized test: the university loses in utility from accepting unqualified students with high scores, but at the same time gains in utility from accepting qualified students with low scores, allowing a strategic school to bundle more students when compared to the case with no standardized test. As $\delta$ decreases and the standardized test becomes less and less accurate, a strategic school starts losing fewer high-score students to rejection than it gains in admitted low-score students, and its utility increases.

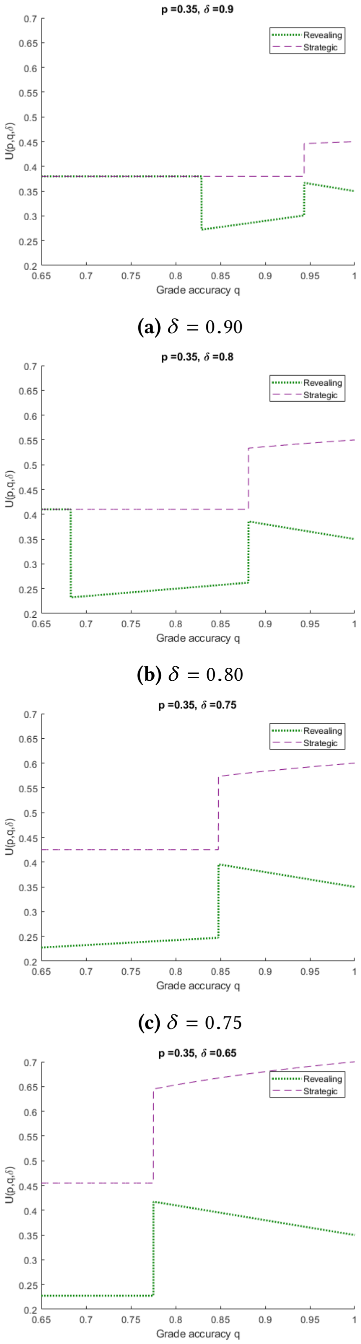
## 6 FURTHER DISCUSSION AND FUTURE DIRECTIONS

Our paper, in introducing the study of inequity induced by population-level signaling, raises a number of directions for future work. We discuss a few of them here.

First, one might be interested in enriching the model of the standardized test intervention. For example, there could be asymmetries in how students from different schools perform on the standardized test. One might imagine students at an advantaged school might be better prepared for the test (e.g., by investment in expensive test-prep courses), giving them an edge in the form of an increased probability of performing well on the test. Suppose, for example, that high-type students in an advantaged school had a higher probability of passing the test than high-type students at a disadvantaged school. In such a situation, more high-type students from the advantaged school would be admitted by the university, and, as the utility for the university to accept high score students increased, the advantaged school could also bundle a larger number of low-type students with its high-type students. That is, a jump in high-types' exam performance increases students' utilities at that school, even for low types; this effect could further exacerbate disparities between and an advantaged and a disadvantaged school. An interesting question could be to quantify how much disparities between schools would increase in such a setting. One could also analyze other variants of advantage on the exam, such as an increased probability of passing both for high-types and for low-types.

One might also imagine that students in an advantaged school might have access to more resources and could take the standardized test several times, while students in a disadvantaged school could only take the test once. When only the highest test score is reported to the university (as is common in practice for university admissions in the United States), it can be seen that this reduces to the situation described above, in which students in each school take the standardized test exactly once, but students in the advantaged school have a higher probability of passing. An extreme case of such a situation would be when the advantaged school's students could take the test enough times that they would pass with a probability approaching 1; in such a case, a test score from the advantaged school would be meaningless to the university. On the other hand, the test would still be significant for the disadvantaged school, and could have the effect of reducing the number of its students that are accepted, further increasing disparities between schools. A natural question would be to quantify such disparities for intermediate values of the number of times that an advantaged-school student can take the standardized test.

Finally, throughout the paper, we assume that the university has unlimited capacity and is willing to accept every student that provides it with non-negative expected utility. One might ask what would happen if the university had a limited capacity. The university might then rank students as a function of their school of origin, their signal, and their test score (in the presence of a standardized test), and only accepted the highest-ranked students. If an advantaged school had the ability to make its students look better than a disadvantaged school (for example, an advantaged school might have more accurate grades and have a higher ability to strategically signal), then the advantaged school could guarantee that some of its students would get first pick by the university, to the detriment of a disadvantaged school—which would only have access to the (possibly small) remaining capacity. A natural direction would be to understand how much of an effect this limited capacity setting can have on inequity.



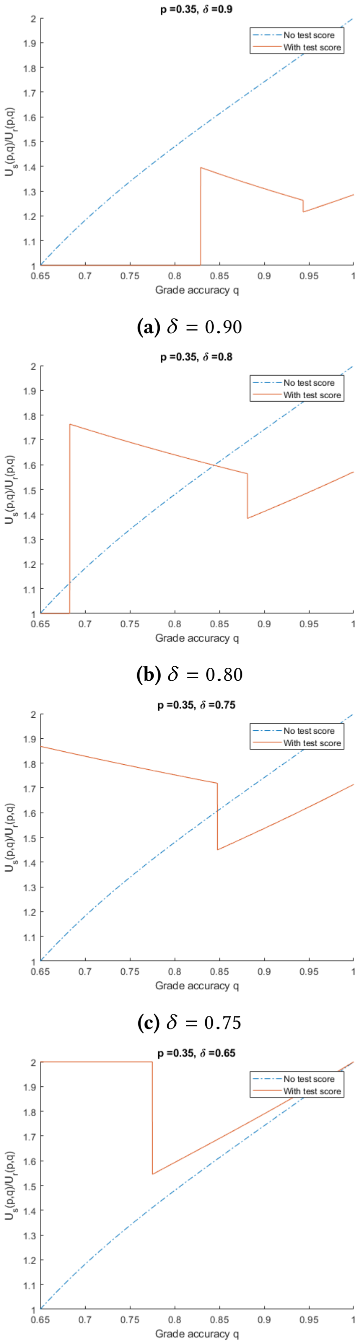
**(a)** $\delta = 0.90$

**(b)** $\delta = 0.80$

**(c)** $\delta = 0.75$

**(d)** $\delta = 0.65$

**Figure 1: Strategic school utility $U_s(p, q, \delta)$ and revealing school utility $U_r(p, q, \delta)$ as a function of the grade accuracy $q$, for average student type $p = 0.35$. We observe that the expected utility may be non-monotone in $q$.**



**(a)** $\delta = 0.90$

**(b)** $\delta = 0.80$

**(c)** $\delta = 0.75$

**(d)** $\delta = 0.65$

**Figure 2: The ratio $U_s(p, q)/U_r(p, q)$ of utilities of a strategic school vs. a revealing school, as a function of the grade accuracy $q$, with and without test score. We observe that the test score intervention may increase inequality.**

## REFERENCES

[1] Ricardo Alonso and Odilon Camara. 2016. Persuading Voters. *American Economic Review* 106, 11 (November 2016), 3590–3605. https://doi.org/10.1257/aer.20140737

[2] Simon P. Anderson and Regis Renault. 2006. Advertising Content. *American Economic Review* 96, 1 (March 2006), 93–113. https://doi.org/10.1257/000282806776157632

[3] Itai Arieli and Yakov Babichenko. 2016. Private Bayesian Persuasion. *Available at SSRN* (September 2016).

[4] Dirk Bergemann, Benjamin Brooks, and Stephen Morris. 2015. The Limits of Price Discrimination. *American Economic Review* 105, 3 (March 2015), 921–57. https://doi.org/10.1257/aer.20130848

[5] Dirk Bergemann and Stephen Morris. 2017. Information design: A unified perspective. (2017).

[6] Dirk Bergemann and Martin Pesendorfer. 2007. Information structures in optimal auctions. *Journal of Economic Theory* 137, 1 (2007), 580–609. https://EconPapers.repec.org/RePEc:eee:jetheo:v:137:y:2007:i:1:p:580-609

[7] Isabelle Brocas and Juan D. Carrillo. 2007. Influence through ignorance. *RAND Journal of Economics* 38, 4 (2007), 931–947. https://doi.org/10.1111/j.0741-6261.2007.00119.x

[8] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[9] Stephen Coate and Glenn C Loury. 1993. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review* (1993), 1220–1240.

[10] Shaddin Dughmi. 2014. On the Hardness of Signaling. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS '14)*. IEEE Computer Society, Washington, DC, USA, 354–363. https://doi.org/10.1109/FOCS.2014.45

[11] Shaddin Dughmi. 2017. Algorithmic information structure design: a survey. *ACM SIGecom Exchanges* 15, 2 (2017), 2–24.

[12] Shaddin Dughmi, Nicole Immorlica, and Aaron Roth. 2014. Constrained Signaling in Auction Design. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '14)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1341–1357. http://dl.acm.org/citation.cfm?id=2634074.2634173

[13] Shaddin Dughmi and Haifeng Xu. 2016. Algorithmic Bayesian Persuasion. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing (STOC '16)*. ACM, New York, NY, USA, 412–425. https://doi.org/10.1145/2897518.2897583

[14] Shaddin Dughmi and Haifeng Xu. 2017. Algorithmic Persuasion with No Externalities. In *Proceedings of the 2017 ACM Conference on Economics and Computation (EC '17)*. ACM, New York, NY, USA, 351–368. https://doi.org/10.1145/3033274.3085152

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[16] Yuval Emek, Michal Feldman, Iftah Gamzu, Renato Paes Leme, and Moshe Tennenholtz. 2012. Signaling Schemes for Revenue Maximization. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. ACM, New York, NY, USA, 514–531. https://doi.org/10.1145/2229012.2229051

[17] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[18] Dean P Foster and Rakesh V Vohra. 1992. An economic argument for affirmative action. *Rationality and Society* 4, 2 (1992), 176–188.

[19] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *CoRR* abs/1609.07236 (2016). arXiv:1609.07236 http://arxiv.org/abs/1609.07236

[20] Matthew Gentzkow and Emir Kamenica. 2014. Costly persuasion. *American Economic Review* 104, 5 (2014), 457–62.

[21] Matthew Gentzkow and Emir Kamenica. 2017. Competition in Persuasion. *The Review of Economic Studies* 84, 1 (2017), 300–322. https://doi.org/10.1093/restud/rdw052

[22] Mingyu Guo and Argyrios Deligkas. 2013. Revenue Maximization via Hiding Item Attributes. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*. AAAI Press, 157–163. http://dl.acm.org/citation.cfm?id=2540128.2540153

[23] S. Hajian and J. Domingo-Ferrer. 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (July 2013), 1445–1459. https://doi.org/10.1109/TKDE.2012.72

[24] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 3323–3331. http://dl.acm.org/citation.cfm?id=3157382.3157469

[25] Lily Hu and Yiling Chen. 2017. Fairness at Equilibrium in the Labor Market. *CoRR* abs/1707.01590 (2017). arXiv:1707.01590 http://arxiv.org/abs/1707.01590

[26] Justin P. Johnson and David Myatt. 2006. On the Simple Economics of Advertising, Marketing, and Product Design. *American Economic Review* 96, 3 (2006), 756–784. https://EconPapers.repec.org/RePEc:aea:aecrev:v:96:y:2006:i:3:p:756-784

[27] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 325–333. http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.pdf

[28] Emir Kamenica and Matthew Gentzkow. 2011. Bayesian Persuasion. *American Economic Review* 101, 6 (2011), 2590–2615. https://EconPapers.repec.org/RePEc:aea:aecrev:v:101:y:2011:i:6:p:2590-2615

[29] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (01 Oct 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[30] 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 (EC '17)*. ACM, New York, NY, USA, 369–386. https://doi.org/10.1145/3033274.3085154

[31] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*.

[32] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*.

[33] Anton Kolotilin, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li. 2017. Persuasion of a privately informed receiver. *Econometrica* 85, 6 (2017), 1949–1964.

[34] Ilan Kremer, Yishay Mansour, and Motty Perry. 2014. Implementing the "Wisdom of the Crowd". *Journal of Political Economy* 122, 5 (2014), 988–1012. http://www.jstor.org/stable/10.1086/676597

[35] Michael Ostrovsky and Michael Schwarz. 2010. Information Disclosure and Unraveling in Matching Markets. *American Economic Journal: Microeconomics* 2, 2 (May 2010), 34–63. https://doi.org/10.1257/mic.2.2.34

[36] Zinovi Rabinovich, Albert Xin Jiang, Manish Jain, and Haifeng Xu. 2015. Information Disclosure As a Means to Security. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 645–653. http://dl.acm.org/citation.cfm?id=2772879.2773237

[37] Luis Rayo and Ilya Segal. 2010. Optimal information disclosure. *Journal of political Economy* 118, 5 (2010), 949–987.

[38] Michael Spence. 1973. Job Market Signaling. *The Quarterly Journal of Economics* 87, 3 (1973), 355–374. https://doi.org/10.2307/1882010

[39] Haifeng Xu, Zinovi Rabinovich, Shaddin Dughmi, and Milind Tambe. 2015. Exploring Information Asymmetry in Two-stage Security Games. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 1057–1063. http://dl.acm.org/citation.cfm?id=2887007.2887154

[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1171–1180. https://doi.org/10.1145/3038912.3052660