ARTICLE

WILEY

Algorithmic Fairness and Statistical Discrimination

John W. Patty [ID] | Elizabeth Maggie Penn [ID]

Political Science and Quantitative Theory &
Methods, Emory University, Atlanta, Georgia

Correspondence

John W. Patty, Political Science and
Quantitative Theory & Methods, Emory
University, Atlanta, GA, USA.
Email: jwpatty@gmail.com

Abstract

Algorithmic fairness is a new interdisciplinary field of study
focused on how to measure whether a process, or **algorithm**,
may unintentionally produce unfair outcomes, as well as
whether or how the potential unfairness of such processes
can be mitigated. Statistical discrimination describes a set
of informational issues that can induce rational (*i.e.*, Bayes-
ian) decision-making to lead to unfair outcomes even in the
absence of discriminatory intent. In this article, we provide
overviews of these two related literatures and draw connec-
tions between them. The comparison illustrates both the
conflict between rationality and fairness and the impor-
tance of endogeneity (*e.g.*, "rational expectations" and
"self-fulfilling prophecies") in defining and pursuing fairness.
Taken in concert, we argue that the two traditions suggest
a value for considering new fairness notions that explicitly
account for how the individual characteristics an algorithm
intends to measure may change in response to the algorithm.

1 | INTRODUCTION

Algorithms are increasingly important to many, if not most, facets of everyday life. Algorithms determine not only
how resources and opportunities—such as employment, housing, credit, and education—are distributed, but also
dictate both the options available to choose from and the information available about those options. Furthermore,
aided by the availability of vast individual-level data and cheap computational resources, these increasingly ubiqui-
tous algorithms are also increasingly opaque and complex. This combination of pervasiveness and inscrutability has
led to concerns about the social impacts of algorithms.[1]

    Algorithmic Fairness. Algorithmic fairness (*AF*) is a new term describing the study of how to evaluate rule-based
procedures for making decisions about diverse individuals. At the heart of this study is the presumption that certain
ways of discriminating between two or more individuals are undesirable (*i.e.*, "unfair"), whereas others are less suspect,
or even desirable (*i.e.*, "permissible"). This field has quickly emerged as an active, important, and multidisciplinary

research agenda over the past 20 years.[2] Driving this emergence is the increased use of sophisticated algorithms in various settings that affect people on an everyday basis.

The principles underlying measures of AF are typically statistical, with a focus on different notions of an algorithm's **accuracy**. As we discuss in more detail below, an algorithm A is typically defined to be more "fair" than another, B, if A is more "equally accurate" with respect to people from different groups (Sharifi-Malvajerdi et al. (2019)), frequently defined by one or more "protected traits," such as race, gender, or sexual orientation. In general, regardless of the fairness metric one chooses, AF seeks to design an algorithm such that its errors do not systematically disadvantage one group of people relative to another.

Statistical Discrimination. The literature on statistical discrimination (*SD*) is more established than that on AF. Rather than measuring and classifying disparities in algorithmic performance across groups, this literature squarely aims to identify the root causes of discrimination, and to disentangle disparate outcomes due to discrimination (*i.e.*, *disparate treatment*) from those due to exogenous disparities across groups (Lang and Kahn-Lang Spitzer (2020)). Theories of SD typically assume that a decision-maker, such as an employer, makes decisions based on observable traits of individuals (*e.g.* gender or race) that may be correlated with unobservable but outcome-relevant traits, such as skill.[3]

Our goal in this article is to introduce the reader to these two, methodologically distinct, approaches to the study of discrimination and to highlight the theoretical connections between the fields. The AF literature takes an exogenous collection of individuals and traits as a primitive and aims to design a sorting mechanism that balances predictive accuracy with fairness. Theoretical work on SD takes a strategic decision-maker and strategic agents as primitives, and aims to understand why inequality can emerge in various decision-making environments. Clearly the approaches differ in the agency afforded to the individuals being evaluated, and in how decisions about individuals are conceptualized (via a mathematical formula versus the profit-maximizing actions of an employer, for example). The most significant difference between the two approaches is in how they conceive of the outcome-relevant traits of individuals that the algorithm or decision-maker seeks to learn (again, such as the skill of an applicant). In the AF literature these individual-level traits are typically taken as exogenous. In theories of SD they are often the product of investment by the people being evaluated, and therefore endogenous to the actions of the decision-maker.

1.1 | Distinguishing Between Algorithmic Fairness & Statistical Discrimination

While the principal goal of this article is to clarify the connections between AF and SD, it is important to note up front that the two are neither equivalent nor "logically nested": there are questions in SD that have nothing to do with AF, and vice-versa. To make this clear, we first walk through a contextualized hypothetical "employment" example and then turn to a recent, real world example of a policy question related to AF, but not SD in a non-employment context.

Suppose that applicants for a job are from two different groups, "male" and "female." Every applicant is either qualified or not, but this is not directly observable. Rather, each applicant has taken a **test**, and the result of this test for applicant is positively correlated with whether he or she is qualified. To make things concrete, suppose that the test is scored on a 0–100 point scale. The employer can observe both the applicant's test score and his or her group membership, and suppose that the employer hires any applicant from group $g \in \{$male, female$\}$ if and only if his or her test score is greater than or equal to the employer's **threshold** for group g, denoted by $t(g) \in \{0, ..., 100, 101\}$.[4] Both AF and SD are interested in the pair of thresholds used by the employer, t(male) and t(female). This stylized setting allows us to clearly identify **discrimination** between the two groups: whenever t(male) $\neq t$(female), the employer would hire applicants from one group with a test score that would **not** lead to hiring an applicant from the other group. Thus, if t(male) < t(female), then the employer applies a "less demanding threshold" when deciding whether to hire a male applicant than when deciding whether to hire a female applicant.

The principal distinction between AF and SD is essentially whose welfare—the applicants' or the employer's—the algorithm should prioritize, and the effect of this prioritization on outcomes. The two approaches can be compared with respect to the types of questions they tend to tackle as follows.

- Studies of **Algorithmic Fairness** tend to focus on questions like:
  1. How do the thresholds affect the applicants' welfares?
  2. What does it mean to treat applicants from both groups of applicants **fairly**?
  3. Which pair(s) of thresholds (if any) treat both groups of applicants fairly?
- Studies of **Statistical Discrimination** tend to focus on questions like:
  1. How do the thresholds affect the employer's welfare?
  2. Which pair(s) of thresholds maximize the employer's welfare?
  3. What factors might justify the employer using different thresholds for the two groups?
  4. How do these thresholds affect individual and group behavior?

We now discuss a recent study of algorithmic fairness in the real world—traffic cameras in Chicago—as a way of illustrating the difference between these related analytical frameworks.

Traffic Cameras, Fairness, & Discrimination. A recent concrete example of AF is provided by a recent ProPublica study of traffic cameras in Chicago.[5] The study found that, in Chicago in 2020, "the ticketing rate for households in majority-Black ZIP codes jumped to more than three times that of households in majority-white areas. For households in majority-Hispanic ZIP codes, there was an increase, but it was much smaller." An AF perspective on this situation essentially asks why this disparity emerges and, more provocatively, how one might reduce or eliminate it. This perspective is particularly helpful in this type of setting because, while this disparity has widened over the two decades since the cameras were introduced in Chicago, there is little reason to suspect that traffic cameras themselves are distinguishing between drivers based on their race or home neighborhood, per se. In this specific case, this perspective allows one to see that the disparity is at least arguably due to speed limits and driving conditions being distributed in a "non-race blind" fashion across Chicago.

Enforcement, Algorithmic Fairness, & Welfare. The analysis also demonstrates the complexities of AF in the real world. Chicago Mayor Lori Lightfoot's administration described traffic cameras as a tool to "a tool in the toolkit to help alleviate" traffic fatalities and, from an empirical standpoint, Black Chicagoans were twice as likely to die in a traffic accident as white Chicagoans in 2017. Accordingly, Black Chicagoans are differentially treated by both traffic accidents *and* traffic tickets. As we return to below in Section 2, this is a classic conundrum in AF settings — an algorithm intended to protect certain individuals might be more error-prone exactly when it is "interacting with," or "treating," those individuals.

Motives, Outcomes, & Statistical Discrimination. From a SD standpoint, on the other hand, one might ask *why* Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets. As the ProPublica article describes, Chicago Mayor Lori Lightfoot—despite arguing as a candidate that Chicago should reduce its dependence on traffic tickets and other fines as a revenue source—proposed lowering the minimum speed at which a speeding ticket would be issued. This proposal, which was adopted by the Chicago City Council in 2021, prompted some to question how much Mayor Lightfoot cared about racial disparities, as opposed to the City of Chicago's serious structural deficit. The question of "is Mayor Lightfoot more interested in racial equality or city revenue?" is directly analogous to the seminal question in statistical discrimination, "is that employer simply maximizing profits or are they racist?" Accordingly, in the SD literature, theoretical work (some of which is summarized in Section 6) has established a partial typology of "non-racist" explanations for "apparently racist" behavior.

In terms of the traffic camera example, we can also distinguish the AF and SD viewpoints as

- **Algorithmic Fairness**: Can we make Chicago's traffic enforcement more fair? If so, how?
- **Statistical Discrimination**: Why did Chicago use an unfair traffic enforcement algorithm?

We believe both of these questions are important. Unfortunately, answering one typically requires some assumptions that are either "outside" or, in some circumstances, at odds with the intents and purposes of the other question. SD has a longer intellectual history and, partially as a result, provides an excellent example of such a conflict. Specifically, as we discuss in Section 6, one strand of work assumes that applicants from the male and female groups have different rates of qualification for exogenous reasons (such as historical discrimination, differences in income, *etc.*)[6] whereas another important strand of work explicitly presumes that the two groups are identical in *ex ante* terms and exhibit different behaviors due to differences in actors' *beliefs* about each other.[7]

1.2 | Rationality Versus Fairness

A theme running throughout this article is that a key contrast between the AF & SD approaches revolves around the question of **rationality** or, in slightly different terms, **efficiency**. Many SD theories are focused on how the pursuit of efficiency (*e.g.*, by an employer, job applicant, government, or other individuals) can generate behavior that is discriminatory. On the other hand, AF is less concerned with efficiency (partly because the basic framework does not presume anything about individuals' motives/goals).

This disjuncture is itself informative: it partially isolates a fundamental tension between rationality/efficiency and fairness in many settings. By taking this contrast seriously, one can leverage the various layers of conflict between various notions of fairness to link them with various actors' instrumental motivations. Taken as a whole, the SD literature is less focused on any specific notion of discrimination or fairness and instead focuses on how the pursuit of instrumental goals might have spillover effects on fairness, broadly written. Similarly, the younger AF literature is less concerned with any specific notion of efficiency, implying that its conclusions are robust to the specific goals of individuals within the algorithmic process in question.

1.3 | Outline of the Article

We structure the article as follows. We first describe a few real-world examples where issues of AF and SD loom large. We will return to one of these examples in later sections in the hopes of making the various fairness concepts (and the stakes associated with them) more concrete. In Section 3 we introduce notation and in Section 4 we provide a brief introduction to some of the more well-known concepts in AF. Section 5 describes how well-known concepts in AF may be at odds with each other. Section 6 lays out a simple model of SD in order to illustrate some connections with questions of AF.

2 | FAIRNESS & DISCRIMINATION IN EMPLOYMENT

The AF and SD literatures are both relevant to a wide array of common economic, social, and political decisions.[8] In this section we briefly explore these connections within the most widely studied setting: employment decisions.

Before discussing these, however, it is important to note that AF is relevant for *any* decision-making process in which decisions depend on individual- and/or group-level traits. Accordingly, the questions examined in this literature are not tied to the use of an explicitly described algorithm (much less a complicated one). Similarly, the issues raised here are relevant to SD in a wide variety of situations, including many in which the term "discrimination" is not widely used.

Employment. A firm is faced with the possibility of hiring an individual. The applicant has two observable traits: gender (male or female) and education level. In addition, the applicant has an unobserved trait (skilled or not). Though the firm can't observe the applicant's skill it wants to hire the individual if and only if he or she is skilled.

A key question for the employer, of course, is how to figure out whether the applicant is skilled or not. Skill may be correlated with either or both gender and education. In such cases, it is permissible for the firm to base its hiring decision on education, but not on gender. Hiring on the basis of an observed trait is known as *disparate treatment*. Disparate treatment on the basis of education is legal, but on the basis of gender is illegal.

Disparate impact occurs when the outcomes experienced by one gender are better than the other.[9] For example, if the employer hires 50% of male applicants, but only 10% of female applicants, then its hiring process exhibits disparate impact, regardless of whether the process also exhibits disparate treatment. In practice, illegal disparate impact in employment typically occurs when the employer uses one or more indicators that are not a "reasonable measure of job performance."

Disparate impact is the primary focus of AF in hiring and promotion decisions. Conversely, the literature on SD—with its focus on the equilibrium actions of goal-oriented agents—is more often concerned with questions of disparate treatment. Disparate treatment is both more straightforward and illegal, *per se*, whereas disparate impact is not illegal if it is caused by some factor that is sufficiently correlated with job performance.[10] Moreover, remedying disparate impact may necessitate disparate treatment. Two central conclusions are:

1. Mitigating disparate treatment can exacerbate disparate impact (and vice-versa), and
2. It is generally impossible to eliminate both disparate treatment and disparate impact.

The Potential Conflict Between Disparate Treatment and Disparate Impact. The conflict between minimization of disparate treatment and disparate impact can be seen in practice in *Ricci v. DeStefano*, 557 U.S. 557 (2009). This case, brought under the Civil Rights Act of 1964, focused on the use of an exam for promotion within the City of New Haven, CT's Fire Department. Twenty firefighters—19 of whom were white and one of whom was Hispanic—passed the exam, but were not promoted because no Black firefighters passed the exam. The City of New Haven ignored the test results because they worried about a disparate impact claim with respect to the exam's outcomes. The twenty non-Black firefighters argued that ignoring the exam results constituted disparate treatment under Title VII of the Civil Rights Act of 1964.

The Supreme Court ruled that the City of New Haven's decision to ignore the test results was illegal because the Court did not agree that New Haven would have been subject to a charge of disparate impact if, after administering the test, the fire department had failed to promote any Black firefighters. Indeed, the legal reasoning's complexity is at least in part due to the fact that minimizing disparate treatment while also minimizing disparate impact was impossible in this specific case. Perhaps obviously, this type of situation is not uncommon. For the remainder of this article we simplify matters and use the term "discrimination" to refer to either disparate treatment or disparate impact.

3 | THEORETICAL FUNDAMENTALS

We now define the standard primitives joining the AF and SD frameworks. For simplicity, we focus on a common setting for this type of work, motivated by the employment setting briefly discussed above. In this setting, there is an **employer** and a pool of **applicants**. The employer will decide whether to hire each applicant on the basis of the applicant's observable characteristics (*e.g.*, education, prior work history, entrance exams, credit history, *etc.*).

Each applicant has a single, *unobserved* characteristic that is of interest to the decision-maker (*e.g.*, is the individual "qualified" for the job or not). For any applicant, the "hiring algorithm" (which might "represent a strategic employer" or not) makes a binary choice (*e.g.* to hire or not). Hiring a qualified individual or not hiring an unqualified individual are each considered a *success*, while hiring an unqualified applicant or not hiring a qualified applicant are each considered *failures* of the algorithm.

Thus, the "employer's objective" for the algorithm is essentially to correctly predict whether each applicant is qualified or not. While we attempt to discuss this material in as transparent a fashion as possible, a little notation will greatly aid our presentation and, more importantly, comparability of our arguments with those in the AF and SD literatures.[11]

3.1 | Basic Building Blocks

As described above, we consider a setting with an employer, E, and a pool of applicants, $N = \{1, 2, ..., n\}$. Each applicant $i \in N$ is described by the following:

1. A profile of *permissible* traits, $x_i = (x_i^1, \ldots, x_i^m)$.
   Examples: Education, technical skills, test scores, credit history.
2. A profile of *sensitive* traits, $a_i = (a_i^{m+1}, \ldots, a_i^M)$.
   Examples: Gender, race, ethnicity, marital status.
3. An *outcome*, $y_i \in \{0, 1\}$,
   Examples: Qualification for the job, profitability of investment, efficacy of treatment.
4. A *decision*, $\delta_i \in \{0, 1\}$.
   Examples: Did i get the job? Did i get admitted? Did i get the loan?

Unless otherwise stated, x_i, a_i, and δ_i are assumed to be observable to the employer, but y_i is not. We briefly describe each of these terms below. Note that for the moment we assume that traits and outcomes are exogenous and fixed. We will relax this assumption later when describing paths for future research, by allowing individuals to potentially exert some control over both.

Traits. We will write $v_i = (a_i, x_i)$ to denote individual i's sensitive and permissible traits and V to denote the set of all profiles of individual traits. The key distinction between the types of traits is that it is "okay" to discriminate between individuals on the basis of differences in their permissible traits, but potentially "not okay" to discriminate between them on the basis of one or more sensitive traits. In the employment example, the applicant's education may be a permissible trait, while his or her gender may be a sensitive trait.

Outcomes. An individual's outcome is a binary measure of a characteristic that is of interest to the employer. In the hiring context, y_i might represent whether the individual would succeed in the job if hired. In the education context, y_i might represent whether i will graduate from college. In the lending example, y_i might represent whether i will repay the loan if it is granted to i.

Decisions. Finally, the goal of the algorithm is to ultimately generate decision(s) that will possibly affect the individuals in question. These decisions are a function of each individual's traits, with $\delta_i \in \{0, 1\}$ being the *decision* regarding person i. In our employment example, $\delta_i = 1$ could represent the decision to hire individual i, and $\delta_i = 0$ would be the decision to not hire i. In light of these definitions, our *algorithm* is the collection of decisions, δ.

3.2 | The Confusion Matrix

We conclude our discussion of "theoretical fundamentals" with an illustration of a **confusion matrix** for this setting of binary outcomes and binary decisions. The simplified setting is useful because it focuses attention on exactly two kinds of mistakes: Type I and Type II errors. This makes it easy to consider and compare fairness goals that focus on the algorithm's relative **predictive performance** across sensitive traits, as discussed below in Section 4.2. That said, this convenience is not without some loss of generality because the 2 × 2 case sets aside a range of

TABLE 1 Predictive Performance with Binary Outcomes and Decisions

	Predicted Outcome		
True Outcome	Positive ($\delta_i = 1$)	Negative ($\delta_i = 0$)	
Positive ($y_i = 1$)	True Positives (TP)	False Negatives (FN)	**True Positive Rate (TPR)** $\frac{TP}{TP+FN}$
Negative ($y_i = 0$)	False Positives (FP)	True Negatives (TN)	**True Negative Rate (TNR)** $\frac{TN}{FP+TN}$
	Positive Predictive Value (PPV) $\frac{TP}{TP+FP}$	**Negative Predictive Value (NPV)** $\frac{TN}{TN+FN}$	

considerations that are potentially important, such as the possibility of "small" versus "large" misclassifications by a decision-maker or algorithm.

Table 1 describes several well-known measures of the predictive performance of an algorithm in this 2 × 2 case. These measures are used to inform a number of the fairness goals that we will shortly define in Section 4. Before proceeding, we will simply note that different stakeholders may prioritize these measures very differently (Narayanan (2018)). Suppose, for example, that outcomes represent whether a worker is qualified ($y_i = 1$) or not ($y_i = 0$), and decisions represent whether the worker is hired ($\delta_i = 1$) or not ($\delta_i = 0$). A qualified applicant may care most about the *true positive rate* of the hiring algorithm, or the probability that they get the job. A firm may care most about the *positive predictive value* of the hiring algorithm, or the probability that a worker that is hired is actually qualified for the job. In general, the relevant actors may care most about measures that condition on their own traits or decisions (outcome y_i for the worker; decision δ_i for the employer).

4 | ALGORITHMIC FAIRNESS GOALS

Corbett-Davies and Goel (2018) define the following two (non-exclusive) families of fairness goals that can be thought of as "fairness from the data alone." As Mitchell et al. (2021) observe, these definitions equate fairness with equalities that can be derived solely from the distribution of traits, outcomes, and decisions.

1. **Anti-classification.** Sensitive traits are not directly used to make decisions.
2. **Classification Parity.** Predictive performance is independent of sensitive traits.

Corbett-Davies and Goel (2018) illustrate some pathological consequences that can follow from achieving anti-classification and/or classification parity, as well as a general incompatibility between measures of classification parity. We now define and briefly discuss these categories of fairness goals.

4.1 | Anti-classification

In many contexts it is inadmissible or illegal to condition one's decision on an individual's sensitive traits. Doing so constitutes disparate treatment under US federal employment law. There has consequently been a clear focus on ensuring that sensitive traits do not factor into how an algorithm or employer evaluates individuals. An algorithm satisfies anti-classification if two individuals with the same permissible traits receive the same decision, or:

$$x_i = x_j \Rightarrow \delta_i = \delta_j.$$

Anti-classification restricts the information that decisions can be responsive to. In this sense, of the three categories of fairness goals, it is the most clearly associated with **process**: what factors can directly affect the algorithm's decision for any given individual? In addition, it is also trivially satisfiable. For example, anti-classification is satisfied simply by having the algorithm assign every individual the same decision ($\delta_i = \delta_j$ for all i, j).

In spite of this pathology, the criterion has important appeal as well. For example, as a direct reflection of a desire to not discriminate against individuals *solely* on the basis of a sensitive trait such as race or gender, it represents a minimal but clear attempt to reduce or eliminate *taste-based* discrimination. This is a large part of the reason that anti-classification is the focus of (and defends against) disparate treatment claims. An important foundational implication of both AF and SD theories is that anti-classification alone will not only not necessarily eliminate discrimination in practice–because of the possibility of disparate impact–but that satisfying it can in some situations *exacerbate* discrimination.

4.2 | Measures of Predictive Performance

Predictive performance is a somewhat loosely defined concept describing various approaches to evaluating how well an algorithm "works." There are many measures of predictive performance, and which measure(s) are appropriate depend on the goals/outputs of the model in question. A classic distinction, for example, divides prediction problems into **classification**—in which the model is attempting to assign individuals/data points to discrete (often "unordered") categories—or **regression** problems, in which the model is associating each individual/data point with a real number. Given our central focus in this article, we consider only classification measures of predictive performance.

Any measure of predictive performance can be used to define a related notion of fairness, often referred to as a form of *parity* or *balance*. Satisfaction of such parity/balance requirements typically requires that the chosen measure(s) of predictive performance be independent of an individual's sensitive traits. Recently, scholars have noticed that many of these measures of parity/balance can be inconsistent with each other.[12] These **impossibility results** imply that, in many cases, *any* algorithm must be "unfair" by at least one measure of predictive performance.

Comparing Predictive Performance Across Sensitive Traits. The predictive performance of an algorithm in the 2×2 baseline case is described by its confusion matrix (Table 1, above). The table contains a wide array of information. A key point is the fact that the positive and negative predictive values, PPV and NPV, are conditioned on the decision, δ_i, whereas the true positive and true negative rates, TPR and TNR, are conditioned on the outcome, y_i. They may look similar but, as is often the case with conditional probabilities, the relationship between the two can

be counterintuitive and depends upon the latent distribution of outcomes, y_i. A key finding in this literature is that different fairness goals will tend to suggest different algorithms whenever the distribution of outcomes, y_i, depends non-trivially on the individuals' sensitive traits, a_i. We now discuss three particularly well-known measures of classification parity: *predictive parity*, *error rate balance*, and *demographic parity*.

Predictive Parity. Predictive parity captures the idea that, conditional on the decision δ, individuals with different sensitive traits should be equally likely to have the same outcome y_i. This boils down to equality of positive predictive values (PPV) across groups, or negative predictive values (NPV), or both. Formally, predictive parity is defined as follows.

Definition 1 (Predictive Parity) *An algorithm δ satisfies*

1. ***Positive predictive parity*** *if each pair of groups, a_i, a_j, have the same positive predictive value:*

$$PPV(a_i) \equiv \Pr[y_i = 1 | \delta_i = 1, a_i] = \Pr[y_i = 1 | \delta_i = 1, a_j] \equiv PPV(a_j). \tag{1}$$

2. ***Negative predictive parity*** *if each pair of groups, a_i, a_j, have the same negative predictive value:*

$$NPV(a_i) \equiv \Pr[y_i = 0 | \delta_i = 0, a_i] = \Pr[y_i = 0 | \delta_i = 0, a_j] \equiv NPV(a_j).$$

Finally, δ satisfies **predictive parity** if δ satisfies both positive & negative predictive parity.

Error Rate Balance. Error rate balance compares true and false positive rates and requires that individuals differing only with respect to sensitive traits are equally likely to be misclassified by the algorithm. Formally, error rate balance is defined as follows.

Definition 2 (Error Rate Balance) *An algorithm δ satisfies **positive error balance** if each pair of groups, a_i, a_j, have the same true positive rate:*

$$TPR(a_i) \equiv \Pr[\delta_i = 1 | y_i = 1, a_i] = \Pr[\delta_i = 1 | y_i = 1, a_j] \equiv TPR(a_j).$$

*Similarly, an algorithm δ satisfies **negative error balance** if each pair of groups, a_i, a_j, have the same true negative rate:*

$$TNR(a_i) \equiv \Pr[\delta_i = 0 | y_i = 0, a_i] = \Pr[\delta_i = 0 | y_i = 0, a_j] \equiv TNR(a_j).$$

*Finally, δ satisfies **error rate balance** (ERB) if δ satisfies both positive & negative error balance.*

Demographic Parity. Demographic parity (sometimes referred to as statistical parity or group fairness) is a widely employed fairness criterion. Substantively, demographic parity is satisfied when sensitive traits do not affect the distribution of decisions for a randomly drawn individual. Demographic parity has several virtues: it is transparent and simple to calculate, and it cleanly captures one classic notion of equality of outcomes. Of course, one downside of demographic parity is that it can be satisfied by an algorithm that doesn't respond to permissible traits, or any data at all. Formally, demographic parity is defined as follows.

Definition 3 *An algorithm δ satisfies **demographic parity** if, for any pair of profiles of sensitive traits, a_i, a_j,*

$$\Pr[\delta_i | a_i] = \Pr[\delta_i | a_j].$$

5 | THE INHERENT CONFLICT BETWEEN FAIRNESS NOTIONS

With several notions of fairness at hand, one might naturally ask under what conditions they will be in agreement. For better or worse, recent theoretical work demonstrates that predictive parity and error rate balance are generally not consistent with each other.

Theorem 4 (*Kleinberg et al. (2016)*) *If an algorithm satisfies Predictive Parity and Error Rate Balance, then one or both of the following must be satisfied:*

$$\Pr[y_i = 1 | a_i, x_i] \in \{0, 1\} \text{ for all } x_i, a_i \text{ Perfect Predictor,} \tag{2}$$

$$\Pr[y_i = 1 | a_i] = \Pr[y_i = 1 | a_i'] \text{ for all } a_i, a_i' \text{ Equal Base Rates.} \tag{3}$$

Theorem 4 establishes that simultaneously satisfying predictive parity and error rate balance is possible *only if* (1) individual outcomes (y_i) can be predicted *perfectly* ("perfect predictor," Equation (2)) and/or (2) individual outcomes are uncorrelated with sensitive traits ("equal base rates," Equation (3)). Note that if perfect predictor is satisfied, then there is no uncertainty in the employer's classification problem, implying that "the optimal algorithm" is trivial. If, in addition or on the other hand, equal base rates is satisfied, then the two groups are statistically indistinguishable with respect to outcomes.

The Implications of Impossibility. At this point, it is worthwhile to briefly think about what impossibility results tell us.[13] The basic idea of an impossibility result is as follows: "nothing can satisfy criteria *A* and *B* simultaneously unless condition *C* is satisfied." In this terminology, *A* and *B* are typically each desiderata that we hope a decision or measure might satisfy, and *C* is a (typically) "special" set of circumstances delimiting the impossibility claim. In Theorem 4, criterion *A* is "equalizes PPV," criterion *B* is "satisfies error rate balance," and condition *C* is "perfect predictor or equal base rates" is/are satisfied. A common misunderstanding of such results (such as Arrow's Theorem Arrow (1951), the Gibbard-Satterthwaite theorem (Gibbard (1973), Satterthwaite (1975), Penn et al. (2011)), or Sen's "Paretian Liberal" (Sen (1970))) is that they describe situations in which it is *very difficult* (or, perhaps, it is "rarely possible") to satisfy criteria *A* and *B* if condition *C* is not satisfied. This is a fundamental and important misunderstanding: impossibility theorems tell us that unless condition *C* is satisfied, criteria *A* and *B* **cannot** be satisfied simultaneously.

Incompatibility of Predictive Parity and Error Rate Balance. Viewed from a different perspective, Theorem 4 establishes that the notions of predictive parity and error rate balance are not only generally incompatible, but also intertwined. This is intuitive in some ways: they consider two different, but related, conditional probabilities. Predictive parity concerns the fairness of the algorithm with respect to applicants' outcomes *conditional on the algorithm's decision*. Error rate balance concerns the fairness of the algorithm with respect to the algorithm's decisions *conditional on the applicants' outcomes*. One reason that Theorem 4 is so surprising at first essentially flows from a "common" mistake with respect to Bayes's rule:

$$PPV = \frac{\Pr[y_i = 1 \,\&\, \delta_i = 1]}{\Pr[\delta_i = 1]} = \frac{\Pr[y_i = 1 \,\&\, \delta_i = 1]}{\underbrace{\Pr[\delta_i = 1 \,\&\, y_i = 1]}_{\text{True Positive}} + \underbrace{\Pr[\delta_i = 1 \,\&\, y_i = 0]}_{\text{False Positive}}}, \tag{4}$$

$$TPR = \frac{\Pr[y_i = 1 \,\&\, \delta_i = 1]}{\Pr[y_i = 1]} = \frac{\Pr[y_i = 1 \,\&\, \delta_i = 1]}{\underbrace{\Pr[y_i = 1 \,\&\, \delta_i = 1]}_{\text{True Positive}} + \underbrace{\Pr[y_i = 1 \,\&\, \delta_i = 0]}_{\text{False Negative}}}, \tag{5}$$

$$NPV = \frac{\Pr[y_i = 0 \,\&\, \delta_i = 0]}{\Pr[\delta_i = 0]} = \frac{\Pr[y_i = 0 \,\&\, \delta_i = 0]}{\underbrace{\Pr[\delta_i = 0 \,\&\, y_i = 0]}_{\text{True Negative}} + \underbrace{\Pr[\delta_i = 0 \,\&\, y_i = 1]}_{\text{False Negative}}}, \tag{6}$$

$$TNR = \frac{\Pr[y_i = 0 \ \& \ \delta_i = 0]}{\Pr[y_i = 0]} = \frac{\Pr[y_i = 0 \ \& \ \delta_i = 0]}{\underbrace{\Pr[y_i = 0 \ \& \ \delta_i = 0]}_{\text{True Negative}} + \underbrace{\Pr[y_i = 0 \ \& \ \delta_i = 1]}_{\text{False Positive}}}, \tag{7}$$

Equations (4)–(7) identify the interlocking connections between PPV, NPV, TPR, and TNR. Comparing the pairs of labels in the denominators of the right hand sides of (for example) Equations (4) and (5), notice that they have one additive term in common (the total frequency of true positives), but differ with respect to the second term of the corresponding sums (total frequency of false positives for PPV and total frequency of false negatives for TPR). A fundamental feature of ("most") non-trivial algorithmic problems (captured by the perfect predictor condition, Equation (2), in Theorem 4) is that these "unmatched" terms of the sums in the two denominators are each non-zero. A foundational fact for Theorem 4 and other impossibility theorems about measures of predictive performance is that the three terms contained in one or both of the denominators on the right hand sides of Equations (4) and (5) represent all but one of four nonnegative numbers that must sum to one. Accordingly, *once the denominators of* Equations (4) and (5) *are fully specified, the denominators of both* Equations (6) and (7) *are fully specified as well.*

Predictive Parity & Error **Rate Balance in Conflict, in Practice.** COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a commercially marketed tool designed to assess the **recidivism risk** of criminal defendants. It has been shown that COMPAS violates error rate balance: an investigation published in ProPublica by Angwin et al. (2016) showed that, while the tool satisfies positive predictive parity, it violates error rate balance in that Black defendants were twice as likely to be mis-classified as "high risk" as white defendants, and white defendants were twice as likely to be mis-classified as "low risk" as Black defendants. Chouldechova (2017) showed, via an admirably simple identity, that positive predictive parity is incompatible with error rate balance unless there are equal base rates (outcomes are independent of sensitive traits, which, in the COMPAS debate, would necessitate equal rates of recidivism for Blacks and whites), or the algorithm disallows certain types of errors (specifically false positives). This result, proved independently of the impossibility theorem of Kleinberg et al. (2016), demonstrates that satisfaction of positive predictive parity generally requires that error rate balance be violated.

Positive Predictive Parity in Hiring. In terms of employment decisions, positive predictive parity requires that the outcomes among those who are hired are distributed identically across individuals possessing different sensitive traits. Considering gender as the sensitive trait, for example, positive predictive parity in employment requires that men who are hired are equally as likely to be skilled as women who are hired. The distribution of outcomes among the hired will typically tend to differ from those not hired, but the distribution of outcomes in either case should not depend on the individuals' sensitive traits. (Conversely, negative predictive parity requires that the distribution of skills among those *not* hired be identical for men and women.) If positive predictive parity is violated, then the employer should differentiate between hired individuals on the basis of gender for subsequent promotion and retention decisions (see Fryer Jr (2007) for an excellent discussion of this issue that is directly relevant to the focus of our article).

Error rate balance focuses on fairness conditional upon individual outcome, rather than the algorithm's decision. Just as predictive parity tends to weigh more heavily in the decision-maker's subsequent decision-making, error rate balance may tend to weigh more heavily in applicants' decision-making, particularly when outcomes are endogenous and/or the algorithm is sensitive to individual outcomes.[14] This is because the error rates associated with an algorithm will affect the marginal costs and benefits of investing in a positive outcome to an applicant motivated by a desire for a positive decision from the algorithm.

To see this, consider an employment setting in which the individual outcome corresponds to the skill of the applicant, and this skill is the product of costly investment by the applicant. Violation of error rate balance may lead to a correlation between an individual's sensitive traits and his or her (1) application for the job and/or (2) investment in skill acquisition. To the degree that the employer is truly indifferent about sensitive traits, per se, this may not have an impact on the employer.

5.1 | Demographic Parity: Motivations and Challenges

As mentioned above, demographic parity is appealing for a variety of reasons, including its simplicity and transparency. However, it is a "group-level" measure of equality, meaning that it does not account for the possibility that the distribution of permissible traits and/or outcomes may differ across (sensitive trait-defined) groups. For example, if individuals with one value of a sensitive trait are more likely than those with a different value of that trait to have an outcome of $y_i = 1$, then demographic parity is not necessarily a reasonable fairness goal from an equity and/or social welfare standpoint.

Less provocatively, it might be the case that outcomes are independent of sensitive traits, but correlated with one or more permissible traits that are themselves correlated with sensitive traits. For example, gender might not be correlated with skill once one "controls for" permissible traits such as education, physical ability, etc., but is correlated with skill if one ignores these factors. A straightforward extension of demographic parity to allow for such inter-group differences with respect to permissible traits can be defined as follows:

Definition 5 *An algorithm δ satisfies **conditional demographic parity** if, for any pair of profiles of sensitive traits, a_i, a_j,*

$$x_i = x_j \Rightarrow \Pr\left[\delta_i | a_i, x_i\right] = \Pr\left[\delta_i | a_j, x_j\right].$$

This generalized notion of demographic parity is an "individual-level" measure of equality, because it accounts for differences between individuals within a given group. Referred to by several names, including *conditional demographic parity* (Mitchell et al. (2021)), *unawareness* (Kusner et al. (2017)), or *treatment parity* (Lipton et al. (2018)) these notions are each equivalent to anti-classification in the settings considered here (Mitchell et al. (2021), p. 153). While not our main point in this article, it is useful to notice that these equivalence results provide further evidence of the interconnectedness of many of the fairness notions currently being analyzed and used in practice. Somewhat less opaquely, many notions of algorithmic fairness explicitly and/or implicitly combine "various components of each other." When this is done explicitly, we view this aggregation of notions as a constructive and sometimes illuminating exercise. However, impossibility results like Theorem 4 imply that things are much murkier when the overlap is less transparent, because one can easily create two or more aggregated notions of fairness that are inconsistent with each other.

In any event, conditional demographic parity has at least one strength relative to absolute demographic parity (it is sensitive to permissible traits) but also at least one weakness (satisfaction of conditional demographic parity *requires* satisfaction of anti-classification). Accordingly, we set demographic parity to the side for the majority of the discussion below.

6 | STATISTICAL DISCRIMINATION

Models of SD aim to explain unequal outcomes across groups game theoretically, as the product of rational decision-making by individuals. As in the literature on AF, these models typically assume that a decision-maker is imperfectly informed about some important characteristic held by an individual. If group membership is correlated with the characteristic then it is rational for the decision-maker to use group statistics as a proxy for the unobserved individual-level characteristic.

Fang and Moro (2011) provide a comprehensive review of the literature on SD, and note that, broadly speaking, there are two strands of models that correspond to two different explanations for inter-group inequality.

Phelpsian Models of Statistical Discrimination. The first class of models discussed by Fang and Moro (2011) build on work by Phelps (1972) and assumes that there may be intrinsic differences between the groups. These differences may be in outcomes (*e.g.* women are, on average, more skilled than men) or in how information about outcomes is conveyed to the decision-maker (*e.g.* men and women are equally skilled on average, but education is a

noisier/less informative signal of skill for men than for women). In both cases, it can be rational for the decision-maker to treat members of the two groups differently. That said, the nature of the distinction between the groups (different prevalence of outcomes vs. differently reliable measures of these outcomes) is relevant for the possibility of achieving different fairness goals, as we return to in Section 6.2, below.

Arrovian Models of Statistical Discrimination. The second class of models discussed by Fang and Moro (2011) build on work by Arrow (1973) and assume that the groups are ex-ante identical. These models yield multiple equilibria, with inequality being characterized by coordination on different equilibria by different groups. The decision-maker ultimately arrives at different beliefs about the two groups, and these beliefs are correct ("self-confirming") and derived in equilibrium. Importantly, these models assume endogeneity of the outcome variable, y_i. Suppose that y_i corresponds to whether i is skilled or not, but that skill is the product of some costly investment by the worker. If, for example, an employer believes that $y_i = 0$ for all individuals in a group then those individuals will have no incentive to invest in becoming skilled, because they have no hope of being hired. This leads to the employer having correct beliefs about the group, in equilibrium.

In the following sections we present a simple model of SD that can accommodate features of both the Phelpsian and Arrovian frameworks. We then use this baseline model to examine questions of AF within a straightforward setting of strategic interaction. We use much of the same notation presented in Section 3 in order to draw connections between these two literatures.

6.1 | A Simple Model of Statistical Discrimination

Consider a situation in which an employer E is deciding whether to hire a worker W. W is either qualified for the job or not. W's outcome variable, $y_W \in \{0, 1\}$, is referred to as W's *qualification*. If $y_W = 1$ then W is qualified and if $y_W = 0$ then W is not qualified. It is assumed that y_W is private information, known only by W.[15]

In addition to being either qualified or not, W also has two potentially observable traits.[16] Gender, $a_W \in \{m, f\}$ is a sensitive trait. A test score $x_W \in \mathbb{Z}_+$ is a permissible trait. The information about W that is available to E is $v_W \in \{m, f\} \times \mathbb{Z}_+$. The fraction of males that are qualified for the job is denoted by $p_m \in [0, 1]$ and the fraction of females that are qualified for the job is similarly denoted by p_f. The terms p_m and p_f represent the "prevalence of qualification" within each respective group, or the base rates of qualification, and E is assumed to have correct beliefs about these values.

The Employer's Payoffs. We assume that E receives a payoff of $B > 0$ for hiring a qualified worker, a payoff of $\omega < 0$ for hiring an unqualified worker, and a payoff equal to 0 for not hiring the worker (regardless of W's qualification). Note that none of these payoffs depend on the worker's gender, a_W. This implies that any employment discrimination within this setting is *not* taste-based (Becker (1971)).

The Employer's Beliefs. Working within a game theoretic tradition, we assume that E always forms correct beliefs about W's qualification, y_W, conditional on the information E has (*i.e.*, conditional on v_W). These beliefs represent a **calibrated assessor**, in the language of AF. Consequently, E's beliefs about W's qualification conditional on observing W's gender and test score will coincide with the probabilities described in Equation (10) or (12), as appropriate, to be defined below.

Optimal Hiring. In this setting, it is well-known (*e.g.*, Corbett-Davies et al. (2017)) that the employer's optimal rule (from the standpoint of expected payoff maximization) balances benefits (correct choice) against costs (incorrect choice) and is described by a threshold rule of the following form:

$$\delta(a_W, x_W) = \begin{cases} 1 & \text{if } \Pr[y_W = 1 \mid a_W, x_W] \geq \bar{s}(B, \omega), \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where, given the assumptions about the employer's payoffs, the threshold $\bar{s}(B, \omega)$ is defined by the following:

$$\bar{s}(B, \omega) \equiv \frac{-\omega}{B - \omega}. \tag{9}$$

Equation (8) implicitly characterizes an assessor used by E to calculate the probability that W is qualified, and this assessor is $s = \Pr[y_W = 1 | a_W, x_W]$. However, independence of $\bar{s}(B, \omega)$ from the worker's traits (v_W) will generally lead optimal hiring by the employer to violate of one or more of the fairness goals discussed above in Section 4. To see this, consider the distribution of traits displayed in Table 2, where the test score x_i can take values in $\{0, 1\}$, and assume that $\omega = -2$ and $B = 1$. Then, Equation (9) yields the threshold $\bar{s} = 2/3$.

Given the distribution of traits in Table 2, the probability a male with $x_i = 1$ is qualified is $\frac{.4}{.4 + .1} = \frac{4}{5}$ and the probability a female with $x_i = 1$ is qualified is $\frac{.3}{.3 + .2} = \frac{3}{5}$. At hiring threshold $\bar{s} = \frac{2}{3}$ the employer's optimal hiring algorithm will yield conditional probabilities of the worker being hired given x_i and a_i displayed in Table 3. While the threshold $\bar{s}$ is independent of a_i, the inferences drawn from x_i about y_i are *not* independent of a_i. The optimal hiring algorithm in this example will violate both error rate balance and demographic parity. As discussed in Section 1.2, the example in Tables 2 and 3 illustrates the tension between rationality and fairness: rational decision-making by the employer may lead to apparently discriminatory hiring behavior *even if the employer does not care about the worker's sensitive traits*.[17]

6.2 | Phelpsian Models of Statistical Discrimination

As noted earlier, Phelpsian models of statistical discrimination explicitly allow for differences between the groups, and there are at least two approaches of this form. The first is based on differences in base rates (*i.e.*, prevalence of qualification) between the groups. The second is based on the data being more reliable for one group than the other. As we will show below, these two approaches present different challenges for the attainment of fairness goals.

Differences in Prevalences Across Groups

To focus squarely on the impact of gender-varying base rates of qualification, we assume for now that $1 > p_f > p_m > 0$ and that the test result x takes values in $\{1, 2, 3\}$. The test result x is a noisy signal of qualification that, conditional on y, is distributed as described in Table 4.

As noted above, perfect Bayesian equilibrium in this model requires that the employer form correct beliefs about the worker's outcome, y_W, conditional on the worker's gender, a_W, and test score, x_W. Given the testing technology defined in Table 4, the probability that W is qualified conditional on observing W's gender a_W and test score x_W is:

$$\Pr[y_W = 1 | a_W, x_W] = \begin{cases} 0 & \text{if } x_W = 1, \\ \frac{p_{a_W}(1 - \phi)}{p_{a_W}(1 - \phi) + (1 - p_{a_W})(1 - \phi)} = p_{a_W} & \text{if } x_W = 2, \\ 1 & \text{if } x_W = 3. \end{cases} \tag{10}$$

In words, if E observes a test result of $x = 1$ he knows with certainty that W is unqualified, and if E observes $x = 3$ he knows with certainty that W is qualified. We therefore restrict attention to E's hiring decision in the event that he observes the "muddled" test result of $x = 2$. Letting $d(m)$ and $d(f)$ represent the respective probabilities that a male or

TABLE 2 An Example of Optimal Hiring Being Unfair

| y_i | $a_i = M$ | | $a_i = F$ | |
	$x_i = 0$	$x_i = 1$	$x_i = 0$	$x_i = 1$
0	0.4	0.1	0.3	0.2
1	0.1	0.4	0.2	0.3

TABLE 3 An Example of Optimal Hiring Being Unfair, Continued

	$a_i = M$		$a_i = F$	
	$x_i = 0$	$x_i = 1$	$x_i = 0$	$x_i = 1$
Pr[hired$|x_i, a_i$]	0	1	0	0

TABLE 4 A Gender-Blind Testing Technology

	$y_W = 0$	$y_W = 1$
Pr[$x_W = 1	y_W$]	ϕ	0
Pr[$x_W = 2	y_W$]	$1 - \phi$	$1 - \phi$
Pr[$x_W = 3	y_W$]	0	ϕ

a female worker receiving a test score of $x = 2$ is hired by E, we'll specifically consider decision rules of the following form for any pair of probabilities $(d(m), d(f)) \in [0, 1] \times [0, 1]$[18]:

$$\delta(a_W, x_W) = \begin{cases} 1 & \text{if } x_W = 3, \\ 1 & \text{with probability } d(a_W) \text{ if } x_W = 2, \\ 0 & \text{with probability } 1 - d(a_W) \text{ if } x_W = 2, \\ 0 & \text{if } x_W = 1. \end{cases} \quad (11)$$

Consequently, E hires those receiving an $x = 3$ with certainty and does not hire those receiving an $x = 1$. He hires those receiving an $x = 2$ with probability $d(a_W)$. Before continuing, we can immediately address when the employer's hiring will satisfy anti-classification. A hiring rule as described by Equation (11) satisfies anti-classification if and only if

$$d(m) = d(f),$$

meaning that the probability of being hired conditional on test score is independent of gender.

Fairness in the Phelpsian Model with Different Prevalences. We are now in a position to summarize under what conditions hiring by the employer will satisfy predictive parity and/or error rate balance. We'll begin by characterizing the optimal rule, from the employer's perspective.

- *Optimality.* By Equations (9) and (10), the optimal decision rule for the employer conditional on observing a score of $x_w = 2$ is:

$$d(a_W) = \begin{cases} 1 & \text{if } p_{a_W} \geq \dfrac{-\omega}{B - \omega}, \\ 0 & \text{otherwise.} \end{cases}$$

Since the optimal rule is a cutpoint rule, it will satisfy anti-classification whenever the two groups either both do or both don't satisfy the above cutpoint rule. As we will show below, if anti-classification is met, then, for this gender-blind testing technology, error rate balance is also satisfied. In the case that the cutpoint is only met for the gender with a higher prevalence, then anti-classification and error rate balance are violated at the optimal rule. Predictive parity, on the other hand, requires hiring individuals receiving a score of $x_W = 2$ probabilistically and at different rates, and is consequently incompatible with a cutpoint rule.

- *Predictive Parity.* If the employer is using a hiring rule of the form described by Equation (11), then employment outcomes cannot satisfy predictive parity. For example, the positive predictive value of hiring individuals with sensitive trait a_W is equal to the following:

$$\Pr[y_W = 1|\delta_W = 1, a_W] = \frac{p_{a_W}\left(\phi + d(a_W)(1-\phi)\right)}{p_{a_W}\left(\phi + d(a_W)(1-\phi)\right) + \left(1 - p_{a_W}\right)d(a_W)(1-\phi)}.$$

Positive predictive value is increasing in p_{a_W} and decreasing in $d(a_W)$; consequently, equalizing PPV across genders when $p_f > p_m$ requires $\delta(f) > \delta(m)$. In other words, males receiving the test score of $x_W = 2$ must be hired at a lower rate than females, and so satisfaction of PPV in this environment necessarily entails a violation of anti-classification unless $d(m) = d(f) = 0$, or no false positives are admitted. Moreover, this relationship between prevalence and hiring rates ($d(f)$, $d(m)$) persists when considering satisfaction of negative predictive value, again requiring that the group with lower prevalence be hired at a lower rate conditional on a score of $x_W = 2$. Consequently, unless certain types of errors are disallowed, predictive parity is incompatible with anti-classification, optimality, and (as we now show) error rate balance.

- *Error Rate Balance.* While predictive parity is incompatible with anti-classification in this setting, the structure of the testing technology implies that any hiring rule that satisfies anti-classification (*i.e.*, $d(m) = d(f) = \tilde{d} \in [0, 1]$) necessarily satisfies error rate balance, and vice versa. This is because, unlike positive and negative predictive values, the true positive and negative rates are independent of group prevalence:

$$TPR = \Pr\left[\delta_W = 1|y_W = 1, a_W, \tilde{d}\right] = \frac{\phi + \tilde{d} \cdot (1-\phi)}{\phi + \tilde{d} \cdot (1-\phi) + (1-\phi)\left(1 - \tilde{d}\right)} = \tilde{d} + \phi\left(1 - \tilde{d}\right),$$

$$TNR = \Pr\left[\delta_W = 0|y_W = 0, a_W, \tilde{d}\right] = \frac{\phi + (1-\phi)\left(1 - \tilde{d}\right)}{\phi + (1-\phi)\left(1 - \tilde{d}\right) + (1-\phi)\tilde{d}} = 1 - \tilde{d}(1-\phi).$$

Importantly, the relevant feature of the testing technology that enables satisfaction of error rate balance is the fact that the probability of receiving the "muddled" test score of $x_W = 2$ is independent of both gender and qualification. We now turn to an environment where this is no longer the case.

Differences in Information Precision Across Groups

In the previous paragraphs we highlighted a setting in which the testing technology was equally informative about male and female outcomes, but discrimination in the form of a violation of predictive parity emerged when prevalence differed between the two genders and the rule was "gender-blind," satisfying anti-classification. An observationally similar form of discrimination can occur even if prevalence is identical across the two genders because of gender-based differences in the precision of the testing technology. Specifically, suppose that $p_m = p_f = \tilde{p} \in [0, 1]$, but the quality of the test outcome depends on the worker's sensitive trait, a_W, as described in Table 5.

TABLE 5 A Gender-Sensitive Testing Technology

	$y_W = 0$	$y_W = 1$
$\Pr[x_W = 1|y_W, a_W]$	ϕ^{a_W}	0
$\Pr[x_W = 2|y_W, a_W]$	$1 - \phi^{a_W}$	$1 - \phi^{a_W}$
$\Pr[x_W = 3|y_W, a_W]$	0	ϕ^{a_W}

For this analysis, we assume that $\phi^m \neq \phi^f$. Given the testing technology defined in Table 5, the probability that W is qualified conditional on observing W's gender a_W and test score x_W is:

$$\Pr[y_W = 1 | a_W, x_W] = \begin{cases} 0 & \text{if } x_W = 1, \\ \dfrac{\tilde{p} \cdot (1 - \phi^{a_W})}{\tilde{p} \cdot (1 - \phi^{a_W}) + (1 - \tilde{p})(1 - \phi^{a_W})} = \tilde{p} & \text{if } x_W = 2, \\ 1 & \text{if } x_W = 3. \end{cases} \tag{12}$$

Fairness in the Phelpsian Model with Different Information Precision. Again, we are in a position to summarize under what conditions hiring by the employer will satisfy optimality, predictive parity and/or error rate balance.

- *Optimality.* By Equations (9) and (12), the optimal decision rule for the employer conditional on observing a score of $x_W = 2$ is:

$$d(a_W) = \begin{cases} 1 & \text{if } \tilde{p} \geq \dfrac{-\omega}{B - \omega}, \\ 0 & \text{otherwise.} \end{cases}$$

In this case the optimal rule will always satisfy anti-classification.[19] However, satisfaction of either error rate balance or predictive parity will require a violation of anti-classification, and error rate balance may not be attainable at all in this setting. We begin with a discussion of predictive parity.

- *Predictive Parity.* The positive predictive value of hiring individuals with sensitive trait a_W is equal to the following:

$$\Pr[y_W = 1 | \delta_W = 1, a_W] = \frac{\tilde{p} \cdot (\phi^{a_W} + d(a_W)(1 - \phi^{a_W}))}{\tilde{p} \cdot (\phi^{a_W} + d(a_W)(1 - \phi^{a_W})) + (1 - \tilde{p}) d(a_W)(1 - \phi^{a_W})}.$$

It is straightforward to show that both positive predictive value and negative predictive value are increasing in ϕ^{a_W}, that PPV is decreasing in $d(a_W)$ and that NPV is increasing in $d(a_W)$. Consequently, predictive parity cannot be satisfied via any hiring strategy $(d(m), d(f))$, and PPV can only be equalized if $d(m) = d(f) = 0$.

- *Error Rate Balance.* Similar to predictive parity, it is straightforward to show that both the true positive rate and the true negative rate are increasing in ϕ^{a_W}. At the same time, the TPR is decreasing in $d(a_W)$ while the TNR is increasing in $d(a_W)$. Consequently, satisfaction of error rate balance is not possible in this setting.

Fairness & Rationality in the Phelpsian Model of Statistical Discrimination. The previous sections have shown that the two Phelpsian approaches to modeling group differences (i.e. whether groups differ in their base rates of qualified individuals versus whether the data/testing technology is differentially noisy for the two groups) have real implications for the possibility of achieving fairness goals at all, much less achieving goals that are compatible with optimal decision-making by the employer. In the former case where groups differ only in prevalence, any decision-rule satisfying anti-classification will necessarily balance error rates for the two groups, and it is possible for optimal-decision making to satisfy anti-classification. In the latter case, it is generically impossible to satisfy either error rate balance or predictive parity at all.

Setting optimality of the decision rule aside, these points suggest that arguments favoring a "group-blind" decision rule may be fundamentally misguided if the data describing members of one group are systematically noisier than

the data describing another. And as Chouldechova and Roth (2020) note, these environments may be ubiquitous. Consider, for example, predicting college success from SAT scores. Chouldechova and Roth (p. 18) write,

> The majority population employs SAT tutors and takes the exam multiple times, reporting only the highest score. The minority population does not. We should naturally expect both that SAT scores are higher amongst the majority population, and that their relationship to college performance is differently calibrated compared to the minority population. But if we train a group-blind classifier to minimize overall error, if it cannot simultaneously fit both populations optimally, it will fit the majority population. This is because—simply by virtue of their numbers—the fit to the majority population is more important to overall error than the fit to the minority population. This leads to a different (and higher) distribution of errors in the minority population.

6.3 | Arrovian Models of Statistical Discrimination

The main distinction between Arrovian and Phelpsian models of statistical discrimination is that, while Phelpsian models treat the outcomes as exogenously determined, Arrovian models allow for outcomes to be endogenous in the sense that the worker may invest in his or her own outcome, y_w.[20] This decision is motivated (at least in part) by a desire to influence the test score, x_w, in pursuit of getting hired ($\delta_w = 1$). This endogeneity opens the possibility of *multiple equilibria*. The existence of multiple equilibria complicates the pursuit of fairness in at least one important way: the hiring rule used by the employer in one equilibrium may satisfy some fairness goals that are violated by the hiring rule used in a different equilibrium.

The existence of multiple equilibria occurs in situations in which SD is "self-confirming." Perfect Bayesian equilibria are based on "rational expectations" by the employer, in which he or she has correct beliefs about any given worker's outcome, y_w, based on the information available to the employer (v_w) at the time he or she makes the hiring decision. If the employer believes that workers from a given group (a_w) are less likely to be qualified ($\Pr[y_w = 1 | a_w]$ is lower), then the employer will require a higher test score to justify hiring individuals from that group. This can, in some situations, reduce the degree to which investing in effort will increase the probability that a worker from that group will be hired, thereby reducing the probability that a worker from that group will find it in his or her interest to actually invest in becoming qualified.

Coate and Loury (1993). An excellent analysis of an Arrovian model of statistical discrimination is provided by Coate and Loury (1993). One of several key points of their analysis is that, when the workers' sensitive trait is observed by the employer at the time of making the hiring decision, individuals with different sensitive traits may be treated by the employer differently in the sense that the hiring rule for one group is different from the hiring rule applied to a different group. This, in turn, leads to each worker's incentive to invest in obtaining qualification endogenously depending on the worker's sensitive trait. Accordingly, discriminatory behavior by the employer may emerge as a result of the equilibrium played by the employer and worker depending on the worker's sensitive trait (in game theoretic terms, this is referred to as *equilibrium selection*). For example, it can be the case that the employer believes that women invest in qualification with some positive probability, but that men do not. In this case, the employer may (correctly) be willing to hire women whose test scores are high enough but (correctly) never hire a male applicant regardless of his or her test score. *This type of discriminatory equilibrium can emerge even if men and women are otherwise identical.*

The Foundations of Arrovian Models of Statistical Discrimination. Arrow's work built upon Phelps's model and, accordingly, Arrovian models of statistical discrimination tend to include the same basic components as Phelpsian models, but also a few more. For example, in order to incorporate endogeneity of workers' outcomes within an employment setting, one must consider the workers' individual incentives. In general, incorporating these incentives leads to the possibility of multiple equilibria. As mentioned above (Section 6), one classic example of equilibrium multiplicity emerges from the possibility of moral hazard: the employer cannot directly observe outcomes, so to

the extent that the employer believes that workers with certain characteristics are unlikely to invest in their own outcomes, the employer will be less likely to hire them, *ceteris paribus*, leading in many settings to there being an equilibrium (or multiple equilibria) in which workers with those characteristics do not invest in outcomes, and do not get hired. In general, when multiple equilibria exist, they differ in the hiring rules used by the employer for workers from different groups, so that, *if the employer can observe individuals' sensitive traits, the employer's hiring behavior discriminates between the two groups in some or all of the equilibria.*

Informational Foundations of Arrovian Discrimination. As discussed above, Phelpsian models of statistical discrimination can be consistent with discrimination emerging "simply" as a result of informational differences between the two groups. These differences, as portrayed in Section 6.2, emanate from exogenous trait-sensitive differences in the testing technology. In Arrovian models of statistical discrimination, discrimination is also the result of informational differences. As opposed to Phelpsian models, the informational differences between groups when discrimination arises in equilibrium in an Arrovian model are *endogenous*. In a sense, this means that the Arrovian explanation for informationally-induced discrimination is closer to a "general equilibrium" explanation than that offered by similar Phelpsian models, because the mechanism of causality is contained within the model. However, this advantage is mitigated by the lack of a mechanism to explain why the employer and worker (correctly and jointly) believe that their equilibrium behaviors should be conditioned on an exogenous trait that is (by assumption) otherwise *per se* irrelevant to both players.[21]

7 | CONCLUSIONS

Studies of AF and SD are similarly concerned with ensuring that the decisions affecting individuals are free of discriminatory bias. However, these fields take very different approaches to the study of fairness. We have argued that the most significant difference between the two approaches is in how they conceive of the outcome-relevant traits of individuals that an algorithm seeks to learn. Studies of AF often conceive of these traits as exogenous to the algorithm itself, with the algorithm simply seeking to identify whether the individuals being classified possess the trait or not. Models of SD often conceive of these traits as the product of choices made by the individuals being evaluated, and therefore endogenous to the algorithm by which individuals are classified. In this article we've attempted to draw some connections between these two literatures, demonstrating how and when various well-known notions of AF may be satisfied within a simple model of SD. We particularly hope that readers will be inspired to pursue the study of how classification algorithms can positively and negatively affect the incentives of individuals to invest in their own defining characteristics.

There are many directions to pursue in this line of work, particularly when considering how our fairness principles might accommodate how people respond to algorithms. An especially timely topic concerns the informational environment that an algorithm faces, which is often the product of policymaking and debate. In recent work, Patty and Penn (2022) consider the effect of so-called "ban the box" policy proposals that require employers to ignore (or at least delay) considering whether a job applicant has a criminal record when making hiring decisions. Using a model that blends both the Phelpsian and Arrovian traditions of SD, they show that this reduction of information can (in equilibrium) sometimes benefit not only job applicants, but also the employer. However, it may also be the case that the information loss associated with banning the box can harm both the employer and the applicants in equilibrium. The possibility of both beneficial and harmful outcomes stemming from ban the box policies may be one explanation for why these policies have met with mixed success in practice.

Jung et al. (2020) consider a related model of statistical discrimination in the context of criminal sentencing. The authors show that a commitment to disregard group information when issuing an incarceration rule of guilt or innocence produces a distribution of decisions that equalizes false positive and false negative rates across demographic groups. Moreover, this "group blind" decision rule serves to minimize overall crime, even though it disregards information that is statistically informative to the decision-maker. The beneficial effect of this information loss occurs because demographic information is not within any agent's control and the decision-maker is assumed to be indifferent with respect to the accuracy of the algorithm, per se. Therefore, using demographic information would simply distort individual incentives to commit crime in ways that cannot help the decision-maker.

A different approach to the question of the information available to an algorithm would allow individuals to reveal or conceal their own personal information. If everyone has the opportunity to choose the information available to an algorithm, when will it be (or can it *ever* be) strictly beneficial for a person to conceal information about themselves? Moreover, there are many environments in which people can manipulate information about themselves. People can lie, and the more important the information is to an algorithm, the stronger the incentive people will face to manipulate it, and the less informative that information will consequently become. In an example of recent work along these lines, Frankel and Kartik (2022) demonstrate that decision-makers may have incentives to under-utilize information for reasons that are entirely distinct from those we've previously discussed (e.g. incentivizing skill acquisition, or promoting fair outcomes). In their model, underutilizing information can actually improve an algorithm's accuracy by reducing peoples' incentives to manipulate data. It is our hope that this article serves to inspire future research into these, and other, important questions concerning algorithmic design.

ORCID

John W. Patty https://orcid.org/0000-0002-1142-9334
Elizabeth Maggie Penn https://orcid.org/0000-0002-2018-4367

ENDNOTES

[1] Kleinberg and Mullainathan (2019) offer an insightful and powerful analysis of the potential costs of pursuing simplicity in these algorithms. As with all of the works cited herein, this work has greatly facilitated our thinking through many issues in this new literature.

[2] Issues of AF are being raised and studied in numerous academic fields, including medicine, criminology, philosophy, law, economics, and computer science.

[3] Theories of *statistical* discrimination differ from those of *taste-based* discrimination (*e.g.*, Becker (1971)) in that the discrimination arises from valid statistical inference and not from any animus toward particular groups. That said, the fact that a decision-maker may have no preferential bias toward or against any group does not mean that the outcomes stemming from a statistically discriminatory process are any less harmful than those stemming from taste-based discrimination.

[4] Note that $t(g)$ = 101 is equivalent to "never hire any applicant from group g."

[5] Hopkins, Emily and Sanchez, Melissa, "Chicago's 'Race-Neutral' Traffic Cameras Ticket Black and Latino Drivers the Most," *ProPublica*, January 5, 2022: https://www.propublica.org/article/chicagos-race-neutral-traffic-cameras-ticket-black-and-latino-drivers-the-most.

[6] These are referred to as *Phelpsian models of statistical discrimination*, as discussed in Section 6.2.

[7] These are referred to as *Arrovian models of statistical discrimination*, as discussed in Section 6.3.

[8] Some other common decisions in which algorithms, and their fairness, have attracted attention recently include college admissions, lending (*e.g.*, Munnell et al. (1996), Ladd (1998)), housing (*e.g.*, Foggo and Villasenor (2021)), criminal sentencing (*e.g.*, Washington (2018)), and advertising (Miller and Hosanagar (2019)).

[9] In practice, the difference in outcomes must exceed some positive threshold, the level of which varies across different contexts.

[10] The question of whether intent is relevant to establishing disparate impact is a complicated one (*Texas Dept. of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, 576 U.S. 519 (2015)).

[11] We generally follow the notation and terminology of Mitchell et al. (2021).

[12] For example, Kleinberg et al. (2016), Chouldechova (2017), and Berk et al. (2018).

[13] For a more general take on what impossibility theorems do (and do not) imply, see Patty and Penn (2014).

[14] We consider both of these possibilities in Section 6, below.

[15] Knowledge of y_W by W is not necessary for our analysis but clarifies some of the main theoretical points.

[16] Unless stated otherwise, we assume that both traits are observable by the employer.

[17] In other words, we have assumed away the possibility of taste-based discrimination. Allowing for this possibility can be accomplished by allowing B and/or ω to depend on the worker's permissible and/or sensitive traits (*i.e.*, B_{v_W} and/or ω_{v_W}).

[18] The optimal rule defined in Equation (8), above, is contained in this family of hiring rules.

[19] Anti-classification is satisfied in this example only because we assume that the probability of receiving a $x = 2$ is the same for both qualified and unqualified types.

[20] Lundberg and Startz (1983) first extended Phelps's model to also allow for endogeneity of the outcome variable via costly investment in human capital.

[21] Put another way, the equilibrium selection in Arrovian models is effectively playing the same role as the exogenous sensitive trait-sensitivity of the testing technology in the Phelpsian models.

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arrow, K. J. (1951). *Social Choice and Individual Values*. John Wiley and Sons.

Arrow, K. J. (1973). Some Ordinalist-Utilitarian Notes on Rawls?s Theory of Justice. *Journal of Philosophy*, 70(9), 245–263. https://doi.org/10.2307/2025006

Becker, G. S. (1971). *The Economics of Discrimination* (2nd ed.). University of Chicago Press.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), 3–44. https://doi.org/10.1177/0049124118782533

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5), 82–89. https://doi.org/10.1145/3376898

Coate, S., & Loury, G. C. (1993). Will Affirmative-action Policies Eliminate Negative Stereotypes? *The American Economic Review*, 1220–1240.

Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint 1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797–806).

Fang, H., & Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. *Handbook of social economics*, 1, 133–200. https://doi.org/10.1016/b978-0-444-53187-2.00005-x

Foggo, V., & Villasenor, J. (2021). Algorithms, Housing Discrimination, and the New Disparate Impact Rule. *Science and Technology Law Review*, 22(1), 1–62.

Frankel, A., & Kartik, N. (2022). Improving Information from Manipulable Data. *Journal of the European Economic Association*, 20(1), 79–115. https://doi.org/10.1093/jeea/jvab017

Fryer, R. G., Jr. (2007). Belief Flipping in a Dynamic Model of Statistical Discrimination. *Journal of Public Economics*, 91(5–6), 1151–1166. https://doi.org/10.1016/j.jpubeco.2006.05.015

Gibbard, A. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4), 587–601. https://doi.org/10.2307/1914083

Jung, C., Kannan, S., Lee, C., Pai, M., Roth, A., & Vohra, R. (2020). Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation* (pp. 677–678).

Kleinberg, J., & Mullainathan, S. (2019). Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability. Technical report National Bureau of Economic Research.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint 1609.05807*.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *arXiv preprint arXiv:1703.06856*.

Ladd, H. F. (1998). Evidence on Discrimination in Mortgage Lending. *The Journal of Economic Perspectives*, 12(2), 41–62. https://doi.org/10.1257/jep.12.2.41

Lang, K., & Spitzer, A. K. L. (2020). Race discrimination: An economic perspective. *The Journal of Economic Perspectives*, 34(2), 68–89. https://doi.org/10.1257/jep.34.2.68

Lipton, Z. C., Chouldechova, A., & McAuley, J. (2018). Does Mitigating ML's Impact Disparity Require Treatment Disparity? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 8136–8146).

Lundberg, S. J., & Startz, R. (1983). Private discrimination and social intervention in competitive labor market. *The American Economic Review*, *73*(3), 340–347.

Miller, A. P., & Hosanagar, K. (2019). How Targeted Ads and Dynamic Pricing Can Perpetuate Bias. *Harvard Business Review*.

Mitchell, S., Potash, E., Barocas, S., Alexander, D.' A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Munnell, A. H., Tootell, G. M. B., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, 25–53.

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp.*, New York, USA (Vol. *1170*).

Patty, J. W., & Penn, E. M. (2014). *Social Choice and Legitimacy: The Possibilities of Impossibility*. Cambridge University Press.

Patty, J. W., & Penn, E. M. (2022). Banning the Box: Information, Incentives, and Statistical Discrimination." Forthcoming. *Quarterly Journal of Political Science*.

Penn, E. M., Patty, J. W., & Gailmard, S. (2011). Manipulation and Single-Peakedness: A General Result. *American Journal of Political Science*, *55*(2), 436–449. https://doi.org/10.1111/j.1540-5907.2010.00502.x

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review*, *62*(4), 659–661.

Satterthwaite, M. A. (1975). Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, *10*(2), 187–217. https://doi.org/10.1016/0022-0531(75)90050-2

Sen, A. K. (1970). The Impossibility of a Paretian Liberal. *Journal of Political Economy*, *78*(1), 152–157. https://doi.org/10.1086/259614

Sharifi-Malvajerdi, S., Kearns, M., & Roth, A. (2019). Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems*, *32*, 8242–8251.

Washington, A. L. (2018). How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate. *Colorado Technology Law Journal*, *17*, 131.

AUTHOR BIOGRAPHIES

John W. Patty is Professor of Political Science and Quantitative Theory & Methods at Emory University. He is currently co-editor of the Journal of Theoretical Politics and the Political Economy of Institutions and Decisions book series at Cambridge University Press. Professor Patty's research focuses on mathematical models of economic and political institutions. His substantive interests include organizational decision-making, political legitimacy, legislative procedures, bureaucratic institutions, and public policy. His work has been published in American Journal of Political Science, American Political Science Review, Annual Review of Political Science, Economics & Politics, Electoral Studies, Games & Economic Behavior, Journal of Law and Courts, Journal of Politics, Journal of Public Policy, Journal of Theoretical Politics, Mathematical & Computer Modelling, Politics, Philosophy, & Economics, Political Science Research and Methods, PS: Political Science & Politics, Public Choice, Quarterly Journal of Political Science, Social Choice & Welfare, and The Good Society. He also coauthored Learning While Governing (University of Chicago Press, 2012) with Sean Gailmard, which won the 2013 William H. Riker book award and the 2017 Herbert A. Simon book award, and Social Choice and Legitimacy: The Possibilities of Impossibility (Cambridge University Press, 2014) with Elizabeth Maggie Penn. He currently serves on the editorial boards of American Politics Research, Games, Journal of Politics, Political Analysis, and Political Science Research and Methods. Professor Patty received his Ph.D. in Social Sciences in 2001 and his M.S. in Economics in 1999 from the California Institute of Technology after receiving his B.A. in Mathematics and Economics from the University of North Carolina-Chapel Hill in 1996 and graduating from the North Carolina School for Science & Mathematics in 1992. Prior to coming to Emory, Professor Patty was Assistant Professor of Political Economy & Decision Sciences at Carnegie Mellon University, Assistant Professor of Government at Harvard University, Professor of Political Science & Director of the Center for New Institutional Social Sciences at Washington University in Saint Louis, and Professor of Political Science at the University of Chicago.

Elizabeth Maggie Penn is a Professor of Political Science and Quantitative Theory and Methods at Emory University. She is a formal political theorist whose work focuses on social choice theory, political institutions, and inter-group dynamics. She regularly teaches courses on electoral systems, positive political theory, and social choice. Her work has been published with Cambridge University Press and in the American Political Science Review, American Journal of Political Science, Journal of Politics, Journal of Theoretical Politics, Annual Review of Political Science, Mathematical & Computer Modelling, Political Analysis, Political Science Research & Methods, Journal of Political Institutions and Political Economy, PS: Political Science and Politics, Public Choice, Social Choice & Welfare, The Good Society, and Complexity. Since the fall of 2015 she has been a Managing Editor of Social Choice & Welfare. She also co-edits the Political Economy of Institutions and Decisions book series at Cambridge University Press. Penn received her Ph.D. in Social Science from the California Institute of Technology in 2003 and her B.A. in Economics and Applied Mathematics from the University of California at Berkeley in 1999. Prior to arriving at Emory she was Assistant Professor of Social and Decision Sciences at Carnegie Mellon University (2003–2005), Assistant Professor of Government at Harvard University (2005–2009), Associate Professor of Political Science at Washington University in St. Louis (2009–2015), and Professor of Political Science at the University of Chicago (2015–2018). She is a 1995 graduate of the Los Angeles County High School for the Arts.