# DISCRIMINATION IN THE AGE OF ALGORITHMS

*Jon Kleinberg*, Jens Ludwig**, Sendhil Mullainathan† and Cass R. Sunstein‡*

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

## ABSTRACT

The law forbids discrimination. But the ambiguity of human decision-making often makes it hard for the legal system to know whether anyone has discriminated. To understand how algorithms affect discrimination, we must understand how they affect the *detection* of discrimination. With the appropriate requirements in place, algorithms create the potential for new forms of transparency and hence opportunities to detect discrimination that are otherwise unavailable. The specificity of algorithms also makes transparent tradeoffs among competing values. This implies algorithms are not only a threat to be regulated; with the right safeguards, they can be a potential positive force for equity.

## I. INTRODUCTION

The law forbids discrimination, but it can be exceedingly difficult to find out whether human beings have discriminated.[1] Accused of violating the law, people might well dissemble. Some of the time, they themselves might not even be aware that they have discriminated. Human decisions are frequently opaque to outsiders, and they may not be much more transparent to insiders.

* Tisch University Professor, Cornell University.

** Edwin A. and Betty L. Bergman Distinguished Service Professor, University of Chicago.

† Roman Family University Professor of Computation and Behavioral Science, University of Chicago.

‡ Robert Walmsley University Professor, Harvard University, Office Harvard Law School, Areeda 225, 1563 Massachusetts Ave, Cambridge, MA 02138. Email: csunstei@law.harvard.edu.

Thanks to Michael Ridgway for his assistance with data analysis; to Justin McCrary for helpful discussions; to Solon Barocas, James Grenier, Saul Levmore, Karen Levy, Eric Posner, Manish Raghavan, and David Robinson for valuable comments; to the MacArthur, Simons, and Russell Sage Foundations for financial support for this work on algorithmic fairness; to the Program on Behavioral Economics and Public Policy at Harvard Law School; and to Tom and Susan Dunn, Ira Handler and the Pritzker Foundation for support of the University of Chicago Urban Labs more generally. Thanks to Andrew Heinrich and Cody Westphal for superb research assistance. All opinions and any errors are obviously ours alone.

1 For an instructive account in the constitutional context, see Strauss (1989). On the general problem, see Lee (2005).

© The Author(s) 2019. Published by Oxford University Press on behalf of The John M. Olin Center for Law, Economics and Business at Harvard Law School.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
doi:10.1093/jla/laz001

A defining preoccupation of discrimination law, to which we shall devote considerable attention, is how to handle the resulting problems of proof.[2] Those problems create serious epistemic challenges, and they produce predictable disagreements along ideological lines.

Our central claim here is that when algorithms are involved, proving discrimination will be easier—or at least it should be, and can be made to be. The law forbids discrimination by algorithm, and that prohibition can be implemented by regulating the process through which algorithms are designed. This implementation could codify the most common approach to building machine-learning classification algorithms in practice, and add detailed record-keeping requirements. Such an approach would provide valuable transparency about the decisions and choices made in building algorithms—and also about the tradeoffs among relevant values.

We are keenly aware that these propositions are jarring, and that they will require considerable elaboration. They ought to jar because in a crucial sense algorithms are not decipherable—one cannot determine what an algorithm will do by reading the underlying code. This is more than a cognitive limitation; it is a mathematical impossibility. To know what an algorithm will do, one must run it (Sipser 2012).[3] The task at hand, though, is to take an observed gap, such as differences in hiring rates by gender, and to decide whether the gap should be attributed to discrimination as the law defines it. Such attributions need not require that we read the code. Instead, they can be accomplished by examining the data given to the algorithm and probing its outputs, a process that (we will argue) is eminently feasible. The opacity of the algorithm does not prevent us from scrutinizing its construction or experimenting with its behavior—two activities that are impossible with humans.[4]

Crucially, these benefits will only be realized if policy changes are adopted, such as the requirement that all the components of an algorithm (including the training data) must be stored and made available for examination and experimentation. It is important to see that without the appropriate safeguards, the prospects for detecting discrimination in a world of unregulated algorithm design could become even more serious than they currently are.

2 See, e.g., *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973); *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977 (1988).

3 For a discussion of some of these issues in a legal setting, see Desai & Kroll (2017).

4 We employ this dichotomy—processes where decisions are made by algorithms and ones where decisions are made by humans—to make the issues clear. In practice, there can be a hybrid, with humans overriding algorithmic judgments with their own. These hybrid processes will clearly have the two elements of each process we describe here, with the additional observation that, with proper regulation, we can see the exact instances where humans overrode the algorithm.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

Our starting point is the current American legal system that has developed a complex framework for detecting and regulating discriminatory decisions.[5] It is increasingly clear that this framework must be adapted for regulating the growing number of questions—involving hiring, credit, admissions, criminal justice—where algorithms are now involved in how public and private institutions decide.[6] Algorithms provide new avenues for people to incorporate past discrimination, or to express their biases and thereby to exacerbate discrimination. Getting the proper regulatory system in place does not simply limit the possibility of discrimination from algorithms; it has the potential to turn algorithms into a powerful counterweight to human discrimination and a positive force for social good of multiple kinds.

We aim here to explore the application of discrimination law to a particularly important category of decisions: *screening decisions*, where a person (or set of people) is chosen from a pool of candidates to accomplish a particular goal, as when college students are admitted on the basis of academic potential or defendants are jailed on the basis of flight risk.[7] Algorithms can be used to produce predictions of the candidate's outcomes, such as future performance after acceptance of a job offer or admission to an academic program. We focus on one kind of machine-learning algorithm often applied to such problems, which uses training data to produce a function that takes inputs (such as the characteristics of an applicant) and produces relevant predictions (*id*). The terminology here can be confusing since there are actually two algorithms: one algorithm (the 'screener') that for every potential applicant produces an evaluative score (such as an estimate of future performance); and another algorithm (the 'trainer') that uses data to produce the screener that best optimizes some objective function.[8] The distinction is important and often overlooked; we shall emphasize it here.

The existing legal framework for these types of screening decisions is necessarily shaped by practical considerations involving the typical difficulty of uncovering human motivations. Simply knowing that there is a disparity in hiring outcomes is not enough to determine whether anyone has discriminated (*McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973)). Perhaps there are genuine differences in average performance across groups (Green 1999; Lee

5 For a clear, recent treatment, see Barocas & Selbst (2016).

6 This point has been made by a large and growing literature in computer science. While the literature is vast, some canonical papers include Dwork et al. (2012); Barocas & Selbst (2016), and the curated set of studies assembled at Fairness, Accountability, and Transparency in Machine Learning (2019).

7 For one example, see Kleinberg et al. (2018).

8 There are clearly other kinds of algorithms and decisions, and they will require an independent analysis.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

2005). A central aspect of the legal challenge is to determine how and why the hiring decisions were made and whether protected personal characteristics, such as race and gender, played a role.[9] Deciding whether there has been discrimination is difficult for one obvious and one less obvious reason. The obvious reason is generic to any legal system: people dissemble, obfuscate, and lie. The less obvious reason is that people may not even know themselves.

A large body of research from behavioral science, described below, tells us that people themselves may not know why and how they are choosing—even (or perhaps especially) when they think that they do (Nisbett & Wilson 1977). Many choices happen automatically; the influences of choice can be subconscious; and the rationales we produce are constructed after the fact and on the fly (Wilson 2004). This means that witnesses and defendants may have difficulty accurately answering the core questions at the heart of most discrimination cases: What screening rule was used? And why? Even the most well-intentioned people may possess unconscious or implicit biases against certain groups (Banjali & Greenwald 2013).

In contrast, a well-regulated process involving algorithms stands out for its *transparency* and *specificity:* it is not obscured by the same haze of ambiguity that obfuscates human decision-making. Access to the algorithm allows us to ask questions that we cannot meaningfully ask of human beings. For any candidate, we can ask: "How would the screening rule's decision have been different if a particular feature (or features) of the applicant were changed?" We can ask exactly which data were made available for training the algorithm (and which were not), as well as the precise objective function that was maximized during the training. We will show that, as a result, we can attribute any observed disparity in outcomes to the different components of the algorithm design or conclude that the disparity is due to structural disadvantages outside of the decision process. In a nutshell: For the legal system, discovering "on what basis are they choosing?" and "why did they pick those factors?" becomes much more feasible.

It would be naïve—even dangerous—to conflate "algorithmic" with "objective," or to think that the use of algorithms will necessarily eliminate discrimination against protected groups (Barocas & Selbsst 2016). The reliance on data does not provide algorithms a presumption of truth; the data they are fed can be biased, perhaps because they are rooted in past discrimination (as, for example, where past arrest records are used to predict the likelihood of future crime) (*id.*; Mayson 2019; Goel et al. 2019). It would also be naive to imagine that the specificity of algorithmic code does not leave room for ambiguity elsewhere.

9 See, e.g., *Corning Glass Works v. Brennan*, 417 US 188 (1974).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

Algorithms do not build themselves. The Achilles' heel of all algorithms is the humans who build them and the choices they make about outcomes, candidate predictors for the algorithm to consider, and the training sample. A critical element of regulating algorithms is regulating humans.[10] Algorithms change the landscape—they do not eliminate the problem.

To make our ideas concrete, consider a case in which a real estate company is accused of gender discrimination in hiring. Both sides agree that the firm has hired fewer women than men in the past but disagree on the reason: the firm cites differences in qualifications while the plaintiff cites discrimination. A highly stylized example of what might happen at trial: The plaintiff's lawyers point to the fact that the plaintiff (like most female applicants to the firm) has more years of schooling than male applicants who were hired. The firm responds that while it weighs education, it also weighs years of work experience (and men have more); according to the firm, years of work experience are more important. The plaintiff's lawyers argue that work experience is being cited after the fact, as a pretext for why men were hired.

Whether work experience is really the deciding factor is not readily discernible. Suppose that like many firms, this one does not use a formal quantitative rule; instead, it looks at each applicant holistically and forms a qualitative sense of who is likely the "best worker." In these circumstances, the plaintiff's lawyers might call the managers involved in hiring to the stand, hoping one of them will respond to questions in a way that reveals discriminatory intentions. None does. The plaintiff's lawyers might also call employees who overheard misogynistic conservations by managers, hoping to show discriminatory intent, but let us stipulate that the conversations are not decisive. It is difficult to know whether the firm has, in fact, discriminated on the basis of gender.

How might things have been different had the firm involved an algorithm in their hiring process? In a well-regulated world with algorithms, the plaintiff's lawyers would ask for the screening and training algorithms, as well as the underlying dataset used. Expert witnesses might be asked to analyze the screening rule; they would likely use statistical techniques that simulate counterfactuals to evaluate how otherwise similar applicants of different genders are treated. Suppose that they discover no disparate treatment; men and women are not being treated differently. But suppose too that the experts observe that the algorithm was given a fairly specific objective in its training procedure—predict sales over the employee's first year. The use of this objective has a

10 As another example, in some cases, humans may choose to override the algorithm and this re-introduces human ambiguity. It is worth noting here that we can see when these overrides arise.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

disparate impact on women. Can that impact be justified under the relevant legal standard, such as business necessity? An analysis of the algorithm might be useful in answering that question as well. It might be able to show that other objectives, such as sales over the first two years, would not have a disparate impact. Moreover, hiring workers with this as the objective is shown to have little net effect on the firm's total sales overall. If so, it would seem difficult for the employer to defend its choice of its objective.

### 1.1 Implications of Our Framework

Five points are central to our analysis of discrimination law in an age of algorithms. First, the challenge of regulating discrimination is fundamentally one of *attribution*. When a screening process produces a disparity for a particular group, to what do we attribute that gap? It could come from the screening rule used. Perhaps the screening rule explicitly takes account of gender. Perhaps the chosen objective—the outcome the screening rule aims to optimize—disadvantages women (or some other protected group). Perhaps the disparity comes from the set of inputs made available to the screener. Perhaps the screening rule fails to optimize for a given outcome using the inputs. The answers to these questions may or may not be relevant, depending on what the pertinent law considers to be "discrimination."

Importantly, the disparity may not result from any of these problems with the screening rule itself. It could also be the consequence of average differences in the outcome distributions across groups. For example, if some groups truly have worse access to K-12 schooling opportunities—perhaps their schools have lower levels of resources—then their college application packets may be less strong on average. Decomposing the source of disparities in screening decisions can be enormously difficult, but it is critical for determining when legal remedies should be applied, and which ones.

Second, this decomposition becomes easier once an algorithm is in the decision loop. Now the decisions we examine are far more specific than "why was this particular candidate chosen?" For example, a key input into the training algorithm is a choice of objective—given the data, the trainer must produce a screening rule that identifies people predicted to do well on some outcome (for example, salespeople expected to have the highest revenues generated). Algorithms are *exceedingly sensitive* to these choices (Kleinberg et al. 2018). In searching for discrimination, the legal system may or may not make it important to ask, "Was the training algorithm given an appropriate outcome to predict?" It should be emphasized that the ability even to ask this question is a

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

luxury: instead of trying to infer why a salesperson was hired, the algorithm's objective function provides us with such information.

The luxury of this knowledge unlocks the power of scrutiny: was this a reasonable choice of outcome?[11] The same point holds for the other key choices for the trainer. Of course, the ability to obtain the relevant knowledge requires (and we argue for) a high degree of transparency. At a minimum, these records and data should be stored for purposes of discovery. Algorithms do not only provide the means to scrutinize the choices we make in building them, but they also *demand* that scrutiny: it is with respect to these choices that human bias can creep into the algorithm.

Third, such scrutiny pays a dividend: if we regulate the human choices well, we might be willing to be more permissive towards how the algorithm uses information about personal attributes in certain ways. When we ensure human choices are made appropriately, some of the concerns that animate the existing legal framework for human discrimination are rendered moot for the algorithm. Suppose, for example, that college applications require recommendations from high school teachers. Any racial bias by teachers could lead to differences in average letter quality across race groups. Interestingly, in some cases, the best way to mitigate the discriminatory effects of biased data is *to authorize the algorithm to have access to information about race* (Kleinberg, Ludwig et al. 2018; Gillis & Spiess 2019). To see why, note that *only* an algorithm that sees race can detect that someone from a given group has better college grades than their letters would suggest, and then adjust predicted performance to address this disparity. Yet much of the time, considerations of factors like race is what antidiscrimination law seeks to prevent (though in this setting, the legal result is not entirely clear) (*Loving v. Virginia*, 388 U.S. 1 (1967); *Miller v. Johnson*, 515 U.S. 800, 811 (1995)).[12]

Fourth, algorithms will force us to make more explicit judgments about underlying principles. If our goals are in tension—as, for example, if admitting more minority students into an elite college would reduce first-year GPAs

11 We are bracketing, for the moment, the question whether that is legally relevant.

12 One possible response to this example is to argue that if the data contain bias, we simply should not use them at all. But in many applications, it is difficult to imagine any alternative to data-driven decision-making. In the case of mortgage lending, for instances, absent information about the current income or assets of a borrower, or prior repayment history, on what basis should a lender decide risk? A middle-ground approach might be only to use those data that are not so biased, but as argued above, the algorithm has a much greater chance of being able to tell which data elements contain a differential signal for one group relative to another than would any human being.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

because of disparities in K-12 school quality or other structural disadvantages—the algorithm precisely quantifies this tradeoff. And we must now articulate a choice. What tradeoff do we find acceptable? We will now be in a position to grapple with such questions *quantitatively*.[13]

Our fifth and final point is that if appropriate regulation can protect against malfeasance in their deployment, then algorithms can become a potentially powerful force for good: they can dramatically reduce discrimination of multiple kinds. A variety of research shows that unstructured decision-making is exactly the sort of environment in which implicit biases can have their biggest impact (Yang 2015; Cohen & Yang 2018). Of course it is true that in building an algorithm, human beings can introduce biases in their choice of objectives and data; importantly, they might use data that are themselves a product of discrimination. But conditional on getting objectives and data right, the algorithm at least removes the human bias of an unstructured decision process. The algorithm, unlike the human being, has no intrinsic preference for discrimination, and no ulterior motives.

And this might not even be the source of the most important gains for disadvantaged groups. In many contexts, efficiency improvements alone have large disparate benefits for members of such groups. For example, Kleinberg et al. (2018) examine pre-trial release decisions in New York, and find that algorithms better distinguish low-risk from high-risk defendants. By prioritizing the highest-risk people to detain, it becomes feasible in principle to jail 42% fewer people with no increase in crime (*id.*). The biggest benefits would accrue to the two groups that currently account for nine of every ten jail inmates: African-Americans and Hispanics.

We develop these points at length, beginning with an exploration of discrimination law, the relevance of principles from behavioral science, and the tensions introduced into this framework by the use of algorithms.[14] Our central claim, stated in simple form, is that safeguards against the biases of the people who build algorithms, rather than against algorithms per se, could play a key role in ensuring that algorithms are not being built in a way that discriminates (recognizing the complexity and contested character of that term). If we do that, then algorithms go beyond merely being a threat to be regulated; they can also be a positive force for social justice.

13 See infra for details.

14 We are building on an emerging line of research connected to developments in computer science, including valuable recent work by Barocas and Selbst (2016) that seeks to situate algorithms within the framework of discrimination law.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

## 2. THE LAW OF DISCRIMINATION: A PRIMER

Discrimination law has long been focused on two different problems. The first is *disparate treatment*; the second is *disparate impact*. The Equal Protection Clause of the Constitution (*Vasquez v. Hillery*, 474 US 254 (1986)), and all civil rights laws, forbid disparate treatment.[15] The Equal Protection Clause of the Constitution does not concern itself with disparate impact (*Washington v. Davis*, 426 US 229 (1976); *McCleskey v. Kemp*, 481 US 279 (1987)), but some civil rights statutes do.[16]

### 2.1 Disparate Treatment

The prohibition on disparate treatment reflects a commitment to a kind of neutrality (Brest 1976). For example, public officials are not permitted to favor men over women or white people over black people. Civil rights statutes forbid disparate treatment along a variety of specified grounds, such as race, sex, national origin, religion, and age.[17]

In extreme cases, the existence of disparate treatment is obvious, because a facially discriminatory practice or rule can be shown to be in place ("no women may apply") (*Reed v. Reed*, 404 U.S. 71 (1971)). In other cases, no such practice or rule can be identified, and for that reason, violations are more difficult to police (Bartholet 1982; Jolls & Sunstein 2006). A plaintiff might claim that a facially neutral practice or requirement (such as a written test for employment) was actually adopted in order to favor one group (whites) or to disfavor another (Hispanics) (*Washington v. Davis*, 426 U.S. 229 (1976)). To police discrimination, the legal system is required to use what tools it has to discern the motivation of decision-makers.[18] To paraphrase the Supreme Court, the key question under the Equal Protection Clause is simple: *Was the requirement or practice chosen because of, rather than in spite of, its adverse effects on relevant group members?* (*Personnel Adm'r of Massachusetts v. Feeney*, 442 U.S. 256

15 For helpful discussion, see Franklin (2012); Bartholet (1982); Mendez (1980).

16 See, e.g., *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971); *Meacham v. Knolls Atomic Power Lab.*, 554 U.S. 84 (2008). In the context of age discrimination, see *Smith v. City of Jackson*, 544 U.S. 229 (2005). For discussion, see Rutherglen (2006); Primus (2003); Bagenstos (2006, p. 4); Selmi (2006, pp. 732-745). For objections, see Gold (1985).

17 See, e.g., 42 U.S.C. § 2000e-2 ("It shall be an unlawful employment practice for an employer - (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin.").

18 A defining framework can be found in *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973). See Green (1999).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

(1979)). That question might be exceedingly challenging to answer, but the law makes it necessary to try.[19]

It is important to see that the disparate treatment idea applies whether discrimination is taste-based or statistical.[20] An employer might discriminate (1) because he himself prefers working with men to working with women; (2) because the firm's coworkers prefer working with men to working with women; or (3) because customers prefer men in the relevant positions. In all of these cases, disparate treatment is strictly forbidden (Strauss 1991). Or suppose that an employer is relying on a statistical demonstration that (for example) women leave the workforce more frequently than men do, or that women over 50 are more likely than men over 50 to leave within ten years. Even if the statistical demonstration is convincing, facial discrimination against women, or some kind of boost for men or penalty for women, is strictly prohibited.[21]

All this is relatively straightforward as a matter of law, but as a matter of principle, it raises some serious puzzles, to which we will return.

### 2.2 Disparate Impact

The prohibition on disparate impact means, in brief, that if some requirement or practice has a disproportionate adverse effect on members of protected groups (such as women and African-Americans), the defendant must show that the requirement or practice is adequately justified.[22] Suppose, for example, that an employer requires members of its sales force to take some kind of written examination, or that the head of a police department institutes a rule requiring new employees to be able to run at a specified speed. If these practices have disproportionate adverse effects on members of protected groups, they will be invalidated unless the employers can show a strong connection to the actual requirements of the job (*Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).[23]

19 We are bracketing some of the differences between the Equal Protection Clause and the civil rights statutes. On the latter, see Green (1999).

20 The classic discussion of taste-based discrimination is Becker (1971). On statistical discrimination, see Phelps (1972). On the difference, see Guryan & Charles (2013); Sunstein (1991).

21 This is a clear implication of *Craig v. Boren*, 429 US 190 (1976), and *JEB v. Alabama ex rel. TB*, 511 U.S. 127 (1994). For some complications, see *Nguyen v. INS*, 533 U.S. 53 (2001); Donohue III (2006).

22 The defining decision is *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971). For our purposes, the full intricacies of the doctrine do not require elaboration. See sources cited in footnote 16 for discussion.

23 The disparate impact standard is now under constitutional scrutiny, but we bracket those issues here. See Primus (2003).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

Employers must show that the practices are justified by "business necessity" (Lye 1998; Jolls 2001; Struebing 2016).

The appropriate justification of the disparate impact standard is widely disputed and raises fundamental questions about the nature of the antidiscrimination principle (Strauss 1989; Rutherglen 2006). The standard can be defended in two different ways (Perry 1991).[24] First, it might be seen as a way of ferreting out some kind of illegitimate motive—and might therefore be essentially equivalent, at the level of basic principle, to the disparate treatment standard. Lacking the tools to uncover bad motives, the legal system might ask: Does the manager have a sufficiently neutral justification for adopting a practice that has adverse effects on (say) women? If not, we might suspect that some kind of discriminatory motive is at work.

An alternative defense of the disparate impact standard would not speak of motivation at all (Fiss 1976; Sunstein 1993). It would insist that if a practice has disproportionate adverse effects on members of traditionally subordinated groups, it should be struck down, unless it has a strong independent justification (Fiss 1976; Colker 1986). On this view, the motivation of the decisionmaker is not relevant. What matters is the elimination of social subordination of certain groups or something like a caste system (Carle, 2011). The disparate impact standard does not, of course, go nearly that far (Ayres & Siegelman 1996). But by requiring a strong justification for practices with discriminatory effects, it tends in that direction.

### 2.3 Fair Representation

Some people, of course, go beyond both disparate treatment and disparate impact. They want members of certain groups to be chosen at socially desirable rates. Call this *the principle of fair representation*.[25] The desire for fair representation may derive from numerous sources. For example, fair representation might have an instrumental value (Hong 2017). Having greater numbers of African-Americans in the police force could be important if it improves the functioning of the force, whose relationship to the relevant community might be better if it is not entirely or mostly white (Sullivan 1986, p. 94). There may also be collateral benefits—including externalities—to such inclusion. Numerous economic examples fit this condition. African-Americans lacking past credit records may not be viable borrowers for lenders, which in turn means they will not ever be able to build a credit record. Even if each individual

24 On the relevant history, see Siegel (2019). On the theory, see Balkin & Siegel (2003).

25 This principle was at the heart of *Mobile v. Bolden*, 446 U.S. 55 (1980).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

lender acts fairly in a local sense, the externality ensures a global unfairness (Sunstein 1991).

The law does not concern itself with fair representation as such, and indeed an effort to pursue that goal, through race-conscious policies, would itself raise serious legal problems under the Constitution and civil rights statutes (*Gratz v. Bollinger*, 539 U.S. 244 (2003)). Race-conscious affirmative action programs remain constitutional, but only under narrow conditions (*id.*). Our only claim here is that some people think that fair representation, as such, is a legitimate and important goal.

This discussion thus far should be sufficient to reveal that as a normative matter, the line between antidiscrimination principles and affirmative action is less clear than is often thought (Strauss 1989). Suppose that an employer is forbidden to take account of customer tastes (even if he deplores them), or that an employer is prohibited from considering statistical reality (even if he deeply wishes it were otherwise). By hypothesis, he is not biased in any way; he is simply trying to succeed in business. Nonetheless, he is required to sacrifice his own economic interests for the sake of attaining broader social goals (Sunstein 1993). In such cases, there is a clear tradeoff. The sometimes-blurry theoretical boundary between anti-discrimination law and affirmative action policy has not posed much of an issue in a world of human decision-making, since the nature of the tradeoffs are hard or even almost impossible to see when so little can be quantified. But this changes when algorithms are introduced, as we discuss below.

## 3. ANTI-DISCRIMINATION LAW FOR HUMANS

We will focus throughout on a single category of decisions. Call them "screening" decisions. For example, a manager or team of managers must screen candidates and decide whom to hire. At its essence, decisions in this category are those where:

- We must make a choice about people. Hire or not? Promote or not? Jail or not? Give credit to or not? The decision materially affects people's lives.
- We must use features of people within the relevant class to make this decision. What is their education? What is their past arrest record? How much do they earn? Have they ever defaulted on loans in the past?
- There is some output or outputs the decision-makers say they care about—job performance, loan repayment, flight risk, and public safety.
- There is some uncertainty (at the least from the perspective of a third party or a judge) as to which combination of features best predicts the outcome of interest. This uncertainty is both qualitative and quantitative. Does

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

height matter at all for job performance in a physical job? And how much exactly does occupation matter for predicting default given that we already have income data?

In each of these decisions, managers will in effect, implicitly or explicitly, have a screening rule, simple or complex, they are using. We discuss in what follows how the law currently tries to identify and prevent people from engaging in discrimination. But in order to understand how and why we have developed our current legal rules regarding discrimination for humans, it is useful to understand first how human cognition works and exactly what we are worried about.

### 3.1 Human Cognition and Biases

Remarkably, large numbers of Americans admit to discriminatory attitudes when asked. Among white respondents to the General Social Survey (GSS), 28 percent believe "it's okay to discriminate when selling a home" (Bobo et al. 1972; New York Times Upshot 2014). Thirty-four percent say "blacks shouldn't push themselves where they're not wanted," and 45 percent say "most blacks don't have the motivation or willpower to pull themselves out of poverty."[26]

But the problems run far deeper than even these striking self-reports would suggest. This goes beyond the obvious problem of dissembling—we cannot trust that everyone who discriminates will tell us. The key lesson of a large body of psychology research tells us that people who discriminate often are not aware of it. Much of the time, human cognition is the ultimate "black box," even to ourselves (Wilson 2004).

To understand this point and its implications for discrimination law, it will be useful to say a few words about psychology in general. Dual-processing theories suggest that human cognition consists of two families of cognitive operations: (1) deliberate, conscious thought (what is often called by psychologists "system II thinking"), which is effortful, and (2) rapid, automatic responses, which is not effortful, and of which people may not even be aware ("system I thinking").[27] Because conscious thinking requires effort, we tend to

26 This is not specific to American race relations; prejudice against out-groups occurs almost anywhere there are people. For example, data from the World Values Survey (2019) show:
35 percent of those in Turkey would not want to be neighbors with someone of a different religion;
41 percent of South Koreans say they would not wish to be neighbors with an immigrant;
66 percent of Russians say they would not wish to be neighbors with someone who was homosexual;
77 percent of respondents in India believe that men make for better political leaders than do women.

27 The literature is voluminous. For an influential summary and exposition of dual systems models in psychology, see Kahneman (2011). Thinking Fast and Slow (2011). Within economics, models of dual systems thinking include Cunningham (2013), and for impulse control, see Fudenberg & Levine (2006).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

rely on our automatic systems to conserve mental energy.[28] If someone says "two plus two," a number automatically comes to mind in response to this frequently-encountered problem. If you read the word "bread," you might also think "butter."

Automatic responses are not limited to these sorts of small situations; they also extend to behaviors that we would presume are quite mindful, such as how we interact with our social environment. For example, Langer et al. (1978) asked subjects in their study to make some copies. Just as they were about to do so, a confederate asked to jump in front of them. In some cases, the confederate gave no reason at all ("Excuse me, I have 5 pages. May I use the Xerox machine?"); in others, a reason was given ("Excuse me, I have 5 pages. May I use the Xerox machine, because I'm in a rush?") The third condition had the *veneer* of a reason, but it actually had no informational content ("Excuse me, I have 5 pages. May I use the Xerox machine, because I have to make copies?") People complied at similar rates to both the actual reason and the pseudo-reason (94 percent and 93 percent, versus 60 percent with no reason). What explains this finding? People see a situation they have seen before and their automatic response kicks in, to avoid the effort of processing the rest of this setting. Consistent with this view, when the costs of complying go up—when the confederate wants to copy 20 pages, not just 5—compliance to the pseudo-reason is similar to what it is in the no-reason condition.

We are often unaware of these automatic responses (Cohen 2003). Latane and Darley (1969) carried out experiments showing that people are less likely to help a stranger in need when there are relatively more people around. They then asked why subjects did not help, "every way we knew how: subtly, directly, tactfully, bluntly. Always we got the same answer. Subjects persistently claimed that their behavior was not influenced by the other people present. This denial occurred in the face of results showing that the presence of others did inhibit helping" (*id.*, 124). Subsequent research has repeatedly confirmed that we often fail to understand why we do what we do.[29] The title of Richard Nisbett and Timothy Wilson's influential essay from decades ago is prescient; in describing our own cognition, we are often "telling more than we can know" (Nisbett & Wilson 1977).

These implicit cognitions can sometimes even be in direct conflict with our conscious thoughts, including with respect to discrimination.[30] The tendency to categorize people and favor "in-groups" and disfavor "out-groups" is

28 Kahneman (1973) is a classic treatment.

29 See, e.g., Cohen (2003).

30 On implicit bias and law, see Jolls & Sunstein (2006). On whether implicit attitudes map onto behavior, see Arkes & Tetlock (2004); Tetlock & Mithcell (2009).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

ubiquitous,[31] and indeed the automatic system pays particularly close attention to other people's sex, age, and race (Cosmides et al. 2003). But the personal characteristics that distinguish groups do not even need to be so obviously distinctive. Consider the famous experiment by Sherif et al. (1961) involving two groups of middle-school youth at Robber's Cave State Park in Oklahoma. Over the course of the study, the two groups (the Eagles and the Rattlers) exhibited increasingly negative views about the trustworthiness, integrity, and athletic skill of the other group, even culminating in aggression. This out-group hostility arose even though *the two groups were actually formed by random assignment of a quite homogenous pool of white Protestant boys* (*id.*). There were no *actual* differences across groups that generated the out-group hostility. All it took was to pit them against each other in a few small competitions.

A major contribution of psychological research has been to show that even implicit biases can be measured (Banjali & Greenwald 2013). In an early study, Word, Zanna, and Cooper asked study subjects—white Princeton undergraduates—to carry out interviews of white and African-Americans confederates (Word et al. 1974). Study subjects were found to sit closer to and spend more time talking with white than African-American interviewees, both measures of how positively inclined one person is toward another. In a follow-up experiment, the subjects now were the ones being interviewed, by a set of confederates randomly assigned to do the interviews either the way whites were interviewed in the first experiment (sitting closer, more time) or how African-Americans were interviewed (further, shorter). Those interviewed as African-Americans had been in the first experiment were rated by observers in this second experiment as having worse interview responses. That is, the interviewer's behavior changes the applicant's responses. Beyond showing that we can measure implicit bias, this study also showed that these biases can create a self-fulfilling prophecy: people are basically creating their own reality.[32]

31 Indeed, when meeting someone new these are typically the most likely things to be remembered about the person, and often the only things that are remembered. This automatic encoding may stem from the same basic cognitive processes related to cooperation and paying attention to social coalitions. Some evidence to this effect comes from studies that create new social coalitions among study subjects in laboratory conditions that are unrelated to race, which seems to substantially reduce mental encoding of race (Cosmides et al. 2003).

32 A great deal of attention has been devoted to other ways of measuring implicit biases, such as the implicit association test, or IAT (Greenwald et al., 1998; Bertrand et al., 2005). The societal consequences of implicit bias remain difficult to determine, partly because we only have IAT results for convenient (rather than truly representative) samples of respondent (Arkes & Tetlock, 2004). But implicit bias seems to be prevalent among those who volunteer to take the tests, and some people argue that it is correlated with actual behavior. For example, counties or metropolitan areas with relatively higher rates of measured implicit bias have been shown to have higher rates of police use of force against blacks, and larger black-white disparities in low-birth weight or preterm birth rates

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

In sum, with respect to human cognition we are, as suggested by the title of Timothy Wilson's 2004 book, often "strangers to ourselves" (Wilson 2004). In important domains, human behavior often looks quite different from what we think of as the conscious optimization of some clear objective.

Consider in this light the consequence of either conscious or subconscious bias for screening decisions.[33] When managers make a rank-ordered list of job applicants, those who discriminate by (say) gender will put men higher on the list than they deserve to be based on expected productivity. Managers will rank some lower-productivity men above some higher-productivity women because, as in the canonical formulation from Gary Becker, they hire a woman over a man only if her productivity advantage is large enough to compensate the manager for the disutility he gets from hiring a woman (Becker 1957). Put differently, the discriminating manager creates a prioritized list of job applicants for hiring that is in some sense *reshuffled*; it is no longer rank-ordered purely by expected job performance. We now turn to the challenge that the law faces in determining when this type of reshuffling has taken place.

### B. The Challenge of Detecting Discrimination

For decades, the challenge of ferreting out illicit motivations has preoccupied analysts of discrimination against protected groups (Lawrence 1987). Indeed, it may be their central preoccupation. We know that racial disparities exist for important outcomes such as income, wealth, credit, schooling, criminal justice,

(Hehman et al., 2018; Orchard & Price 2017). One study administered IATs to managers in a French grocery store chain. They found that the productivity of minority workers was lower when they interacted with managers who had IAT scores indicating more bias (Glover et al., 2017). It turns out that managers with higher levels of implicit bias interacted less with minority workers, leading to reduced worker effort. This helped support biased manager stereotypes about minority workers being less productive—another example of a self-fulfilling prophecy.

33 As briefly noted in the text, differential treatment on grounds of race, sex, and other factors can arise even when people are not trying to express personal animosity, but merely looking out for the bottom line. Suppose that an employer favors whites over African-Americans, not because he wants to discriminate (he does not), but because his customers like dealing with whites, and he does not want to lose money. Or suppose that a firm relies on a statistical demonstration that on average women leave the workforce more frequently than men do. If a hiring manager cannot tell which specific women are more likely to leave, then—if turnover is bad for the bottom line—a profit--maximizing company might use gender as a proxy for the propensity to leave. This is an example of what economists call "statistical discrimination." See, e.g., Arrow (1973); Phelps (1972). Statistical discrimination might be good for the firm, but it penalizes women (including those who are at low risk of leaving the firm themselves). Under the Equal Protection Clause and civil rights laws, it is generally unlawful to discriminate on the basis of customer preferences or to engage in statistical discrimination, even if these are profit-maximizing (Strauss 1989).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

health, health care, and many others.[34] One cause of these disparities involves people's choices and behaviors, including disparate treatment. We know that because people explicitly tell us they discriminate (as in the GSS survey results above), and also from carefully controlled "audit studies." For example, Bertrand and Mullainathan (2004) sent fictitious resumes out to a large number of employers in Boston and Chicago that were identical except that half had a white-sounding name and the others had an African-American-sounding name. Resumes with white names received 50% more call-backs from employers who wanted to carry out an interview (*id.*).

But that evidence for discrimination is *statistical*. Without more, it cannot establish disparate treatment with respect to any *individual* decision. This question about what the former can tell us about the latter is at the heart of the debate about the defining case of *McCleskey v. Kemp*.[35] An African-American defendant convicted of killing a white police officer and then sentenced to death challenged his sentence by citing statistical evidence that capital punishment was administered at a relatively higher rate in cases with a black offender and a white victim. The Court rejected the challenge and insisted on the need for proof of discrimination in the particular case. It objected that the plaintiff's claim "that these statistics are sufficient proof of discrimination, without regard to the facts of a particular case, would extend to all capital cases in Georgia, at least where the victim was white and the defendant was black" (*McCleskey v. Kemp*, 481 U.S. 279, 296 (1987)).

Here is another way to understand the problem. Consider the fact that African-Americans constitute 15 percent of all college-age young people but only 6 percent of all students enrolled at top private universities in the USA (Ashkenas et al. 2018). This disparity seems more than sufficient to conclude there are important barriers *somewhere* in American life. But for the purposes of the law, we must identify a concrete decision that we believe was affected by discrimination, so that we can name a defendant. Was it the college admissions committee? Or someone else more "upstream" in the process whose decisions contributed to disparities in K-12 learning opportunities? The frequent inscrutability of human decision-making, combined with the relative rarity of explicit archival evidence (memos, verbal statements) of discriminatory motives, makes violations difficult to police (Lawrence 1987; Carle 2011).

Without some kind of formal discrimination ("no women need apply" (*Reed v. Reed*, 404 U.S. 71 (1971))) or a "smoking gun" document, the only other direct way to tell whether someone discriminated in a specific case may be to

34 See for example, Stanford Center on Poverty and Inequality (2017).

35 481 U.S. 279 (1987).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

ask them.[36] Even setting aside the risk they lie, as noted above, they honestly might not even know themselves. Many screening decisions in practice do not involve any sort of formula or guidelines, but instead are largely (or entirely) subjective. A manager may tell us they selected applicants for hiring because they seemed like "good workers." That could mean almost anything. And given the black-box nature of human cognition, even cooperative managers may not be able to explain it. Nor may they be able to articulate what predictor variables for future productivity were used, or why those were chosen over other candidate predictors. And given the possibility of implicit bias, can every manager really say in good faith "I did not pay attention to gender (or race, etc.) in making my decision"? In the analogy to algorithms raised in the introduction, we cannot understand the choices that went into the "trainer" that creates the screening rule the person used. In contrast, statistical tests to infer these choices can be quite complex.

The black-box nature of the human mind also means that we cannot easily simulate counterfactuals. If hiring managers cannot fully understand why they did what they did, how can even a cooperative manager answer a hypothetical about how he would have proceeded if an applicant had been of a different race or gender? And if it is hard to imagine a counterfactual that involves changing a major salient personal characteristic of that kind, what hope do we have for a counter-factual entailing some incremental change to a different, less salient qualification like educational attainment or years of work experience? Put differently, not only can we not readily understand the "trainer" behind the human's decision, but we may also be unable to understand the "screener" that was actually used.

Of course, courts have developed tools for addressing these challenges, including statistical analysis (Finkelstein 1980; Tobia 2017). If we cannot get a meaningful answer from people about what they did, we might at least try to reconstruct their decision-making by interrogating data, as courts often do (*Segar v. Smith*, 738 F.2d 1249 (D.C. Cir. 1984), *cert. denied*, 105 S. Ct. 2357 (1985)). But a data-driven investigation faces its own challenges (Greiner 2008). For example, it can be difficult for judges to know whether a screening rule treats equally productive candidates the same when they cannot really measure output, but only qualifications. Consider how difficult defining the "output" of a job is even for a relatively simple occupation like, say, retail cashier. Is output the fraction of days worked where the cash register and recorded sales balance out? Or how often the worker shows up on time, not late? Or how often customers come back to the store? Or total sales more generally?

36 Statistical measures may of course be relevant (Norris, 1986, p. 65; Browne, 1993).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

What if we focused on qualifications? With the aid of existing tools, determining whether applicants with the same qualifications are treated the same may be helpful or even sufficient, but it runs into the problem of which qualifications matter. Attempts to focus on those qualifications that are most predictive of output run into the problem of measuring output. We might be tempted to solve this problem by comparing applicants who are the same on *every* qualification. But this can often lead to a *very* long list. And in many cases there are just not that many people actually hired into a given job by a given firm over a given time period. If we have, say, 10 people hired into a job over the past 5 years but 20 plausibly relevant qualifications, it becomes impossible to tell if a firm really hired someone from an advantaged group over a "similarly qualified" person from a disadvantaged group.

With these questions, we do not mean to suggest that the problem of ferreting out discrimination is insuperable. A great deal of illuminating work casts light on that problem and on potential solutions (Baldus & Cole 1980; Finkelstein 1980; Conway & Roberts 1986; Greiner 2008). The only point is that when human behavior is involved, it can be extremely challenging to ascertain relevant motivations and hence to see whether disparate treatment was at work.

We have seen that in the face of a demonstration of disparate impact, statistical evidence alone is not enough to conclude there was unlawful discrimination; the defendant is given a chance to provide some qualitative evidence of a justifiable, neutral reason for the disparity (such as "business necessity").[37] For example, imagine that the head of a private security firm institutes a rule requiring new employees to be able to run at a specified speed. If these practices have disproportionate adverse effects on some group, such as women, they will be invalidated unless they can show a strong connection to the actual requirements of the job.

Notice this takes us back not only to the difficulty of defining and measuring "output," noted above, but more generally puts judges and jurors in the position of having to make potentially difficult judgments about the best way to carry out some task that is far from their own expertise. The head of the security firm, for example, argues that the ability to run fast is an important part of the job. He notes that there is variability in the running speed of suspects, and variability in the distance of foot-chases, but he believes that a security guard who could run 1.5 miles in 12 minutes should be able to catch the suspect in "most" chases. It is challenging for judges to decide on the merits whether any of the security firm head's assertions are correct. It is no wonder, in these

37 On the dynamics, see Barocas & Selbst (2016); Lye (1998); Wax (2011)

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

circumstances, that much of the debate turns on an institutional question: how aggressively the legal system should scrutinize those assertions (Tobia 2017).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

## 6. USING ALGORITHMS TO PROMOTE EQUITY

Algorithms have the potential to help us to excise disparate treatment, to reduce discrimination relative to human decision-making, to limit disparate impacts, and also to predict much more accurately than humans can in ways that disproportionately benefit disadvantaged groups—what we call the "disparate benefit" of algorithms.

### 6.1 De-bias Relative to Humans

As we have seen, the tendency to classify others into "in-groups" and "out-groups" is a key feature of human psychology that contributes to explicit and implicit biases throughout society. For example, Bertrand and Mullainathan show that resumes with a typically African-American name are much less likely to receive a call-back than similar resumes that list a common white name (Bertrand & Mullainathan 2004). To overcome the effects of these human biases, we could try to exhort hiring managers to ignore irrelevant information on resumes. Or we could provide them with some sort of implicit bias training. But given the opacity of human decision-making, it would be very difficult to determine whether such efforts had actually been successful (Jackson et al. 2014).

The use of an algorithm is an alternative way to try to deal with the bias of human decision-making. To the algorithm, the name on a resume, race, age, sex, or any other applicant characteristic are candidate predictors like any other: variable X42. If this variable is not predictive of the outcome, the algorithm will not use it. And since predicting the outcome is *all* the algorithm is designed to do, we do not have to worry about any hidden agenda on the part of the algorithm itself. And recall that if use of a characteristic *is* predictive but would violate antidiscrimination law, use of that characteristic can and should be prohibited. But here there is an important qualification, to which we now turn.

### 6.2 Access to the Protected Variable Promotes Equity for the Algorithm

Consider a firm that is trying to decide which sales people to steer towards its most lucrative clients based on a prediction of their future sales level. Candidate predictors include (1) past sales levels and (2) manager ratings. Suppose that for men, managers provide meaningful assessments that include useful signals about employee performance that are not fully captured in the past sales data. But suppose that for women, the managers discriminate and give the lowest possible ratings. (This is of course disparate treatment and therefore unlawful.) An algorithm that is *prohibited* from knowing gender might well

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

use manager ratings as a predictor, because it has a useful signal for half the sample. And because the algorithm in this case does not know who is male and who is female, it has no choice but to assume that manager ratings mean the same thing for all workers. The resulting predictions would understate the future productivity of women and hence contribute to gender gaps in earnings.

But what happens if we instead allowed the algorithm to be *aware* of gender? With adequate training data, the algorithm could detect that manager ratings are predictive of future sales for men but not for women. Since the algorithm is tasked with one and only one job—predict the outcome as accurately as possible—and in this case has access to gender, it would on its own choose to use manager ratings to predict outcomes for men but not for women. The consequence would be to mitigate the gender bias in the data.[56] Clearly, such a mitigating effect will not result from the use of a protected attribute in every situation, but we can easily find additional scenarios where similar considerations arise.

For example, allowing the algorithm to have access to protected-class membership can also promote what many people would consider to be equity in cases where the relationship between the candidate predictors and outcomes differ between the advantaged and disadvantaged groups. To use an admittedly more controversial example, suppose that we have two college applicants who both score 1,100 on the SAT. One of them is from New Trier, an affluent northshore suburb of Chicago where the median family income is $145,000 (three times the national average) with public schools among the nation's best. The other is from Englewood, a south side Chicago neighborhood with median family income under $20,000 and among the city's highest homicide rates. Outside of extraordinarily unusual circumstances, it *cannot* be the case that the amount of effort, persistence, and extra learning on one's own required to

56 As another example, consider a city in which half of all residents are white, half are black. Suppose that white residents and black residents engage in criminal activity at exactly the same rates, and that the police force in this city manages to arrest everyone who commits a crime. But in addition, the police department includes some discriminatory officers and as a result half of all arrests made to black residents are actually false arrests (that is, an arrest when the person was not reasonably thought to have committed a crime). In this case, the arrest rate will be higher for blacks than for whites even though true rates of crime are the same. Now suppose we use these data to predict failure to appear in court, or FTA, which (for the moment) we will assume is accurately measured. An algorithm blinded to race has no choice but to assume that the relationship between prior arrests and FTA risk is the same for both groups; since the rate of prior arrests is higher for blacks as whites, the race-blind algorithm would generally predict a higher FTA risk for blacks. In contrast, an algorithm able to use race would detect in the data that (in our stylized example) the effect of each prior arrest on FTA risk was smaller for black residents than for whites; with more prior arrests for blacks than whites, the net result would be equalized predicted FTA risks for blacks and whites by the race-aware algorithm.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

score 1,100 on the SAT is the same for the student who starts with every possible advantage as for the one forced to overcome a long list of difficult obstacles.

Yet, when we prohibit an algorithm from having access to information about the college applicant's disadvantaged group membership (in this case, let us stipulate, race as well as economic circumstances and background),[57] that is *exactly* what we are forcing the algorithm to assume. From the standpoint of current law, it is not clear that the algorithm can permissibly consider race, even if it ought to be authorized to do so; the Supreme Court allows consideration of race only to promote diversity in education (*Grutter v. Bollinger*, 539 U.S. 306 (2003)). Whatever the law requires, many people would find it tempting to say that while the algorithm should focus on factors independent of race (poor neighborhood, violent crime, and so forth), it should not focus on race itself (*Gratz v. Bollinger*, 539 U.S. 244 (2003)). But sometimes race itself is relevant as a predictor, and on plausible assumptions, there is nothing invidious about insisting on that point.

These are not only hypothetical examples; we can see them play out in actual data. In Kleinberg et al. (2018),[58] we use data on a nationally representative sample of teens to predict college performance and simulate different admission decisions. Table 1 shows that, consistent with past studies, high school outcomes differ on average by race, and as a result we see differences in average college outcomes as well. Figure 1 shows that a race-aware algorithm allows us *substantially to increase the share of admitted students who are minority* (holding average GPA constant) relative to ignoring information about race. The share of the admitted class that has GPA $< 2.75$ is on the $y$-axis and the share that is African-American is on the $x$-axis. We show results from a race-aware algorithm, a race-blind algorithm, and an algorithm that pre-processes the data so that the average of all high school predictors is the same on average for white and black students.[59] The algorithms let us rank-order applicants by predicted GPA outcome.

57 We acknowledge that the stipulation is controversial. To skeptics, we emphasize that it is a stipulation.

58 The dataset we use is the U.S. Department of Education's National Education Longitudinal Study of 1988, which captures information on a nationally representative sample of students who entered 8th grade in the fall of 1988. Follow-up surveys were conducted in 1990, 1992, 1994, and in 2000 (when respondents were in their mid-20s) (U.S. Department of Education, 2000). We limit our analysis sample to those who participated in the 2000 survey wave, and had ever attended a four-year post-secondary institution. To simplify the analysis, we focus on comparing just non-Hispanic white students ($N = 4{,}274$) with black students ($N = 469$). In the public-use version of the NELS we use here, we do not have access to ACT or SAT scores. But we do have the results of how students did on standardized academic achievement tests that the NELS administered to students in four academic areas: math, reading, science, and social studies.

59 We might do this if, for example, we believed that systematic differences across groups in the average values of the predictor variables (Xs) was due to some sort of societal unfairness, or to biased data.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

**Table 1. Descriptive statistics from NELS dataset for predicting college performance**

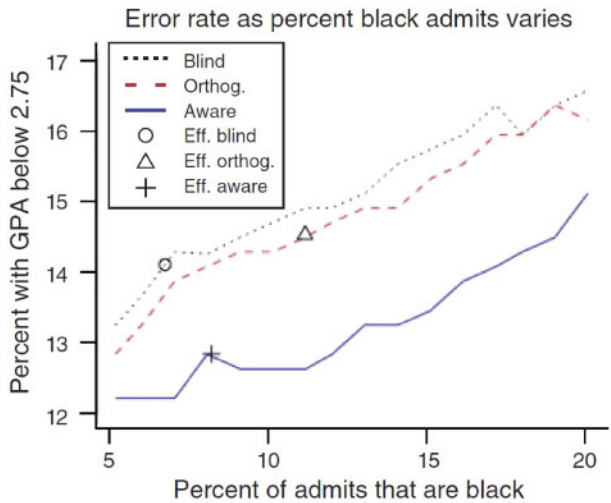
| | Whites | Blacks |
|---|---|---|
| *N* | 4,274 | 469 |
| College outcomes (2000 follow-up, %) | | |
| Graduate with BA by 2000 | 67.4 | 50.9 |
| College GPA $\geq$ 2.75 | 82.2 | 69.5 |
| College GPA $\geq$ 3.25 | 48.4 | 31.1 |
| Female (%) | 53.2 | 56.7 |
| 10th grade NELS test scores (standard deviation = 10) | | |
| Reading | 56.3 | 50.0 |
| Math | 57.1 | 49.5 |
| Science | 56.2 | 47.5 |
| History | 56.0 | 50.0 |
| 10th grade course grades (%) | | |
| English: Mostly A's | 29.4 | 19.6 |
| Math: Mostly A's | 26.9 | 14.9 |
| Science: Mostly A's | 26.9 | 14.4 |
| History: Mostly A's | 29.1 | 15.7 |
| 10th grade extracurricular hours per week (%) | | |
| None | 21.4 | 30.2 |
| 0-1 | 16.9 | 14.4 |
| 1-4 | 21.9 | 23.4 |
| 5-9 | 16.3 | 13.6 |
| 10-19 | 17.7 | 11.3 |
| 20+ | 2.1 | 1.2 |
| 12th grade NELS test scores (standard deviation = 10) | | |
| Reading | 56.0 | 50.0 |
| Math | 57.1 | 49.5 |
| Science | 56.2 | 47.5 |
| History | 56.0 | 50.0 |
| Number of credits taken by 12th grade by subject | | |
| English | 3.6 | 3.4 |
| Math | 3.2 | 2.9 |
| Science | 2.9 | 2.5 |
| Social Studies | 3.1 | 2.8 |
| Took SAT (%) | 60.4 | 63.1 |
| Took ACT (%) | 52.6 | 37.9 |

If we do this separately by race, *we can select whatever share of the incoming class we wish to be minority by deciding how far down the minority student list to go*. The tradeoff curve is upward-sloping for each algorithm (more diversity, higher share GPA $< 2.75$), but the race-aware algorithm dominates the others. For example, if we wanted to hold the share of admitted students with GPA $< 2.75$ at 14 percent, the share of admitted students who are black would be 7 percent with the race-blind model, 11 percent with the pre-processed model, and *19 percent with the race-aware algorithm.*

Figure 2 shows why the race-aware algorithm dominates. This "heat map" shows the share of black students in different predicted GPA "bins" according

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

**Figure 1. Fairness versus efficiency tradeoffs in college admissions using race-aware versus race-blind algorithms.**

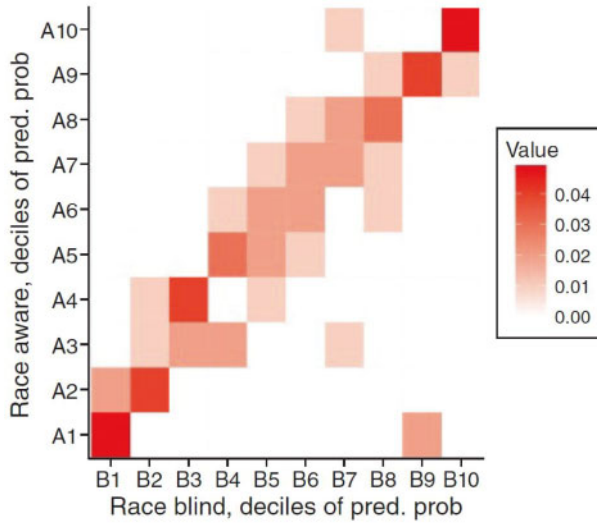



***Source:*** **Results from Kleinberg et al. (2018) using data from the NELS: 88 to predict college performance for those students who attended a four-year college (as measured by GPA < 2.75). The graph shows the results of rank-ordering all students by predicted performance and simulating an admission rule that admits the top half; the *y*-axis shows the percent of the "admitted class" that goes on to have a GPA below 2.75 ("mostly B's"), while the *x*-axis shows the share of the admitted students who are African-American. The points in the graph show the result of rank-ordering applicants using linear probability models that (a) are "blinded" to any information about each applicant's race [circle]; (b) use information about each applicant's race to pre-process the data and orthogonalize the predictors to race, so that the average value of high school grades, test scores, etc. are now forced to be equal for black and white applicants [triangle]; and (c) use information about applicant race to form the prediction model [cross]. For each candidate algorithm, we also show what happens if we use our decision-making framework to achieve a target fairness level (share of the admitted college class that is African-American). Specifically, we take the predictions from a given algorithm, create a list of white applicants rank-ordered by predicted college performance, and a different list of black applicants rank-ordered by their predicted college performance, then work down the list of black applicants until we hit the target for minority enrollment in the hypothetical "admitted class." We can see that admission decisions using the race-aware algorithm dominate those from the race-blind or the race-orthogonalized algorithms, in the sense that for any given level of academic performance of the incoming freshman class, the race-aware algorithm lets us substantially increase the share of the admitted class that is African-American.**

to the race-blind (*x*-axis) or race-aware (*y*-axis) algorithms. Observations that lie off of the 45-degree line show the models disagree. For example, the race-blind algorithm says the students in the right-hand lower corner have high likelihood of GPA < 2.75 (9th decile), but the race-aware predictor (accurately)

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

**Figure 2. Mis-ranking of African-American applicants by race-blind algorithm.**




***Source:*** **Kleinberg et al. (2018). Using data from the NELS: 88 dataset, we first predict for each observation their predicted college performance (measured as y = 1 if GPA < 2.75) using an algorithm that is blinded to applicant race, and then again using an algorithm that has access to each applicant's race. We then take the sample of black students in the NELS: 88 and use their predicted values to bin them into deciles based on the race-blind predictions (x-axis) and race-aware predictions (y-axis). If the two models rank-ordered everyone the same way, all the data would be along the 45-degree line. The "off diagonals" in the figure show mis-ranking.**

tells us they are actually in the *lowest-risk* decile. The race-blind model mis-ranks these students because the relationships between high school predictors of college success (such as test scores) turn out to be different for white versus black students.[60]

This discussion highlights how the mechanical application of current anti-discrimination law to algorithms might actually have harmful effects on exactly those populations we seek to protect. Our goal here is not to offer a final view on whether and when current law would forbid decision-makers—whether human beings or algorithms—from explicitly considering race. It is enough

60 When we use the race-blind algorithm, we force the statistical model to assume the relationship is the same for both whites and blacks, which distorts the predicted outcomes for black applicants. This leads us inadvertently to say some high-performing black applicants are actually low-performing, that is, to understate their academic potential and hence reject them, while simultaneously admitting lower-performing black applicants instead.

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

to say that courts would be very uncomfortable with that practice, and their discomfort, in the context of algorithms at least, might well be a mistake.

### 6.3 Disparate Benefit from Improved Prediction

There is a less controversial but comparably important point. The largest potential equity gains may come from simply predicting more accurately than humans can. This increase in accuracy can generate benefits that disproportionately accrue to disadvantaged groups, leading to what we call "disparate benefit."

For example, landlords in the private housing market have difficulty predicting which tenants are at high risk for skipping rent payments. This difficulty may lead landlords to implement blanket rules, such as requiring first and last month rent to protect against missed payments. For affluent renters, this is a minor inconvenience. But for low-income, low-wealth families, this can be the difference between leasing an apartment and doubling up with someone else (or even becoming homeless). Better prediction could allow landlords to relax collateral requirements, which would *change* the system in ways that helps disadvantaged families.

Better prediction could also let us *expand* some systems in ways that help disadvantaged groups. Compared to affluent white families, low-income minority families tend to live in neighborhoods that do not only have higher poverty rates, but are also located further from good schools, health care, and jobs (Pendall 2014). This distance is particularly a problem for families who rely on public transportation, which in the USA remains, as Pendall puts it, "slow, inconvenient, and [lacks] sufficient metropolitan-wide coverage to rival the automobile" (*id.*). While car loans have increased in recent years, interest rates often remain high in part because of non-payment: 6 million people are at least 90 days behind on their car loans (Guilford 2017). High interest rates surely contribute to differences in car ownership by income and employment status (Raphael & Rice 2002). Better prediction of payment risk could allow lenders to lower interest rates for low-risk families and expand access to car ownership in ways that disproportionately benefit the disadvantaged.

We are offering somewhat conjectural examples, but these potential gains are not merely conjectural. Consider an application where better prediction lets us shrink a system that disproportionately harms disadvantaged groups: pre-trial release decisions for criminal defendants (Kleinberg et al. 2018). State law in New York requires judges make these decisions based on a prediction of defendant risk of failure to appear (FTA) in court in the future. We use data from New York City on all cases that were continued at arraignment over a five-year period and build a model to predict FTA risk. The predictor variables in the

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

model consist of age (a legally allowable variable), current offense, and prior criminal record. The data show that the judges' risk predictions (implicit in their release decisions) are correlated with the algorithm, but that compared to the algorithm, the judges make numerous serious mistakes: they detain many low-risk people and release many high-risk ones.[61]

If we made these decisions using the algorithm's risk prediction rather than those of the judge, we could prioritize detention just for those people who are high risk. Indeed, some of the high-risk people judges release are *so* high risk that if we detained them, we could release multiple low-risk people without increasing FTAs. Indeed, if we followed the algorithm's release recommendations, we could reduce the jail population by 42 percent without increasing FTA rates at all (Kleinberg et al. 2018).

Figure 3 provides a more concrete sense for what this means in practice. In a city in which around 50,000 people spend time in jail each year, it would mean about 20,000 fewer people spending time behind bars on Riker's Island each year (New York City Department of Correction 2018). This would be the equivalent of closing Riker's Island at the end of July every year, with no increase in FTA or re-arrests. And note who benefits the most from this large reduction in detentions: the two groups who together account for nearly 90 percent of all current jail spells—African-American and Hispanic defendants.[62] This is a disparate benefit, and an especially large one.

### 6.4 Algorithms Can Reveal Our Own Biases

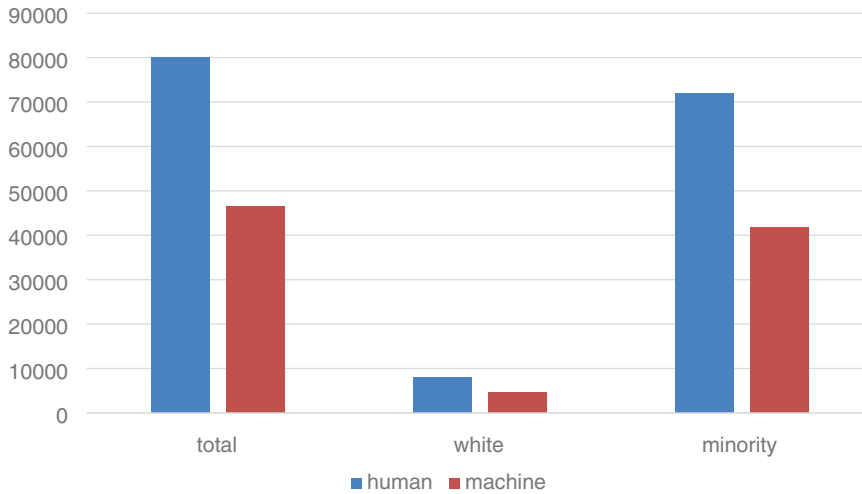
A final potential benefit arises from the fact that algorithms can help reveal our own biases. Imagine a large, growing firm that is inundated with job applications. To help prioritize which resumes to consider as it hires and expands, the

61 In particular, judges seem substantially to over-weight the severity of the current charge (i.e., whatever offense the person was arrested for that led them to wind up in court) (Kleinberg et al. 2018; Sunstein 2019). When the judges form their own implicit list of defendants to prioritize for detention based on their subjective assessment of defendant risk, the judges wind up putting too many felony arrestees towards the top of their list and too many misdemeanor arrestees towards the bottom of the list. But the machine-learning algorithm makes clear that prior record matters a lot in terms of future risk, and human judges are not fully taking this into account (Kleinberg et al. 2018).

62 These potential gains are not limited to pre-trial release decisions (Chalfin et al. 2016). Using algorithmic predictions rather than human judgment can help police departments hire officers at lower risk for adverse outcomes like police-involved shootings, complaints for verbal abuse, or complaints for physical abuse, and can help school systems hire more effective teachers. Because police misconduct so disproportionately involves disadvantaged population (Fryer, 2018), and similarly because the most effective teachers tend to prefer working with the most affluent students leaving the least effective to serve disproportionately low-income and often predominantly minority schools, see, e.g., Engel et al. (2014).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

**Figure 3. Pre-trial detention rates in New York City under current human (judge) decisions versus algorithmic release rule that holds failure to appear (FTA) rate constant at current level.**




***Source*: Kleinberg et al. (2018). "Human decisions" represent current outcomes from the existing criminal justice system, which technically combines the input of judges, district attorneys, public defenders and a six-item risk check-list created by a local non-profit, the Criminal Justice Agency. "Machine decisions" represent the hypothetical detention outcomes if one were to make pre-trial detention versus release decisions using a machine-learning algorithm (gradient-boosted trees) to predict defendant risk, rank-order defendants by risk, and detain only the number of defendants needed to keep the failure to appear rate constant at current levels.**

firm builds an algorithm to rank applications. The outcome predicted by this algorithm is based on the people the firm's managers hired in the past. The firm then notices that the new algorithm mostly hires men. A possible response would stem from the recognition that the algorithm, given how it was constructed, reveals implicit or explicit biases among the firm's HR team. In that light, the firm might try to overcome biased human judgments by building a new algorithm that predicts an outcome less infected by human bias, such as a more objective measure of worker productivity once hired. A perhaps less helpful response would lose sight of what is responsible for the algorithm's bias, and drop the algorithm altogether to revert back to a purely human hiring process.[63]

63 Interestingly, this is what some media accounts suggest may have happened in 2018 at Amazon (Reuters 2018).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

The transparency of algorithms will have other consequences that might be uncomfortable for many people. Recall that the disparate impact standard forbids disproportionate adverse effects on members of certain groups, unless there is a strong independent justification for the requirement or practice that creates those adverse effects. Here, we can see the lines blurring between antidiscrimination principles and affirmative action (Strauss 1989). Suppose that we have two candidate algorithms to predict worker productivity. One of them would lead to hiring a set of workers that is 1 percent more productive than those hired by the other, but reduces the number of minorities hired by 10 percent. Is this productivity gain large enough to provide "strong independent justification" for that algorithm? Answering that question unavoidably winds up requiring value judgments about how to trade off avoidance of disparate racial impacts against other social objectives (such as output).

Another way to put it is to say that those who favor affirmative action programs are sometimes willing to sacrifice some value for the sake of other goals (such as racial justice as they see it). Those who want to prevent disparate impacts, or who want to ensure that any disparate impact is strongly justified, are willing to do *exactly the same thing*. In practice, current antidiscrimination law says that it is worth suffering something, but not too much, to stop disparate impacts on (say) African-Americans or women. But how much? Algorithms permit unprecedented clarity about these questions by allowing us to specify the magnitude of tradeoffs. What if, instead of a 1 percent productivity gain for a 10 percent decline in minority hiring, it was a 10 percent productivity gain for a 1 percent decline in minority hiring? Many people might respond that the relevant judgments will depend on the precise numbers, which are typically impossible to quantify with "black box" human decision-making.

## 7. Conclusion

It is tempting to think that human decision-making is transparent and that algorithms are opaque. We have argued here that with respect to discrimination, the opposite is true. The use of algorithms offers far greater clarity and transparency about the ingredients and motivations of decisions, and hence far greater opportunity to ferret out discrimination.

For those who wish to reduce discriminatory behavior, this is a massive opportunity. Countless decisions have the flavor of the screening problems we discuss here; they hinge on a prediction. There is powerful evidence of discrimination in current human decision-making. To offer just a few examples:

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022

- Audit studies that randomly assign otherwise-equivalent white and black applicants to apply at different firms find that white applicants are called back at more than twice the rate of black applicants, 34 percent versus 14 percent.[64] Reducing this bias would do an enormous amount of social good given there are over 6 million job openings in the USA at any point in time (Bureau of Labor Statistics 2019).
- Audit studies of the U.S. housing market, which originates $2 trillion in new mortgages each year (Kapfidze 2018), find that minority borrowers are treated differently from and worse than white borrowers (Turner et al. 2002).
- Bias also arises in the health sector, which accounts for $3.5 trillion in spending each year in the USA alone (equal to 18 percent of GDP) (Centers for Medicare & Medicaid Services 2018; World Bank 2019). For example, when doctors were shown two equivalent patient histories, the chances of recommending a beneficial procedure (cardiac catheterization) were 40 percent lower for women and minorities than white males (Schulman et al. 1999).

For each of these critically important decisions, more and more data are becoming available over time on how people in the past made these decisions, the characteristics of these people and their decision-making environments, and the consequences of different decisions. These data make it increasingly possible to build data-driven statistical prediction models. Those models might be used to detect, reduce, or eliminate existing discrimination.

Algorithms have extraordinary promise. They have the *potential* to make important strides in combating discrimination, at least as the legal system has long understood it. But principles of transparency and auditability, fair and nondiscriminatory choice of data, and reasonable algorithmic objective are essential, not least to help understand and select the tradeoffs that people use algorithms to make.

64 This figure is for whites and blacks without criminal records. Among those with criminal records, we see similarly large differences in call-back rates, equal to 17 percent versus 5 percent, respectively (Pager 2003).

Downloaded from https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086 by guest on 13 September 2022