DATA-DRIVEN DISCRIMINATION AT WORK

PAULINE T. KIM*

ABSTRACT

A data revolution is transforming the workplace. Employers are increasingly relying on algorithms to decide who gets interviewed, hired, or promoted. Although data algorithms can help to avoid biased human decision-making, they also risk introducing new sources of bias. Algorithms built on inaccurate, biased, or unrepresentative data can produce outcomes biased along lines of race, sex, or other protected characteristics. Data mining techniques may cause employment decisions to be based on correlations rather than causal relationships; they may obscure the basis on which employment decisions are made; and they may further exacerbate inequality because error detection is limited and feedback effects compound the bias. Given these risks, I argue for a legal response to classification bias—a term that describes the use of classification schemes, such as data algorithms, to sort or score workers in ways that worsen inequality or disadvantage along the lines of race, sex, or other protected characteristics.

Addressing classification bias requires fundamentally rethinking antidiscrimination doctrine. When decision-making algorithms produce biased outcomes, they may seem to resemble familiar disparate

* Daniel Noyes Kirby Professor of Law, Washington University School of Law, St. Louis, Missouri. This Article benefitted from the comments of participants at workshops at Washington University School of Law, the University of Denver Sturm College of Law, and the 2015 Colloquium on Scholarship in Employment and Labor Law. I would like to thank Scott Baker, Marion Crain, Lee Epstein, Peggie Smith, David Law, and Neil Richards for their helpful comments. I am also indebted to participants at the June 2016 Privacy Law Scholars' Conference—especially danah boyd, Solon Barocas, Andrew Selbst, and Mark MacCarthy—for their valuable feedback, and to Erika Hanson and Jae Ryu for outstanding research assistance.

impact cases; however, mechanical application of existing doctrine will fail to address the real sources of bias when discrimination is data-driven. A close reading of the statutory text suggests that Title VII directly prohibits classification bias. Framing the problem in terms of classification bias leads to some quite different conclusions about how to apply the antidiscrimination norm to algorithms, suggesting both the possibilities and limits of Title VII's liability-focused model.

TABLE OF CONTENTS

INTRODUCTION

The data revolution has come to the workplace. Just as the analysis of large datasets has transformed the businesses of baseball, advertising, medical care, and policing, it is radically altering how employers manage their workforces. Employers are increasingly relying on data analytic tools to make personnel decisions, thereby affecting who gets interviewed, hired, or promoted.[1] Using highly granular data about workers' behavior both on and off the job, entrepreneurs are building models that they claim can predict future job performance.[2] Sometimes called workforce or people analytics, these technologies aim to help employers recruit talented workers, screen for eligible candidates in an applicant pool, and predict an individual's likelihood of success at a particular job.[3]

Proponents of the new data science claim that it will not only help employers make better decisions faster, but that it is fairer as well because it can replace biased human decision makers with "neutral" data.[4] However, as many scholars have pointed out, data are not neutral, and algorithms can discriminate.[5] Large datasets often

1. *See, e.g.*, George Anders, *Who Should You Hire? LinkedIn Says: Try Our Algorithm*, FORBES (Apr. 10, 2013, 4:31 PM), http://www.forbes.com/sites/georgeanders/2013/04/10/who-should-you-hire-linkedin-says-try-our-algorithm [https://perma.cc/M7NF-SJJD]; Jeanne Meister, *2014: The Year Social HR Matters*, FORBES (Jan. 6, 2014, 10:21 AM), http://www.forbes.com/sites/jeannemeister/2014/01/06/2014-the-year-social-hr-matters/ [https://perma.cc/L6SJ-VMJE]; Claire Cain Miller, *Can an Algorithm Hire Better Than a Human?*, N.Y. TIMES: THEUPSHOT (June 25, 2015), http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html [https://perma.cc/PKM6-4JY4].

2. *See, e.g.*, Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES (Apr. 20, 2013), http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html [https://perma.cc/3X99-EM4X].

3. *See* Josh Bersin, *Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age*, FORBES (Feb. 17, 2013, 8:00 PM), http://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/ [https://perma.cc/W69F-3BAM].

4. *See, e.g.*, *id.* (discussing workforce analytics as the superior alternative to employment decisions "made on gut feel"); Lohr, *supra* note 2 (examining views of many proponents of workforce analytics).

5. *See, e.g.*, Solon Barocas & Andrew D. Selbst, Essay, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 674 (2016); danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural Technological, and Scholarly Phenomenon*, 15 INFO. COMM. & SOC'Y 662, 666-68 (2012); Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35, 35 (2013); Joshua A. Kroll, Joanna Huey, Solon Barocas,

contain errors in individual records, and these errors may not be randomly distributed. Algorithms that are built on inaccurate, biased, or unrepresentative data can in turn produce outcomes biased along lines of race, sex, or other protected characteristics. When these automated decisions are used to control access to employment opportunities, the results may look very similar to the systematic patterns of disadvantage that motivated antidiscrimination laws. What is novel is that the discriminatory effects are data-driven.

Of course, employers have always done things such as recruiting, hiring, evaluating, promoting, and terminating employees, but data models do not rely on traditional indicia like formal education or on-the-job experience. Instead, they exploit the information in large datasets containing thousands of bits of information about individual attributes and behaviors. Third-party aggregators harvest information from the internet about job applicants, including detailed information about their social networking habits—how many contacts they have, who those contacts are, how often they post messages, who follows them, and what they like.[6] Similarly, monitoring devices collect data on the workplace behaviors of current employees, recording information such as where they go during the day, how often they speak with others and for how long, and who initiates the conversation and who terminates it.[7] Employers can also obtain information about their employees' off-duty behavior. As employees spend more of their personal time online, third parties can collect information on those activities, aggregate it with other data, and share it with employers.[8] Growing participation in wellness programs means that employees increasingly share

---

Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. (forthcoming 2017) (manuscript at 29-35), https://ssrn.com/abstract=2765268 [https://perma.cc/CL85-DUKK]; Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 10, 2013), http://foreignpolicy.com/2013/05/10/think-again-big-data/ [https://perma.cc/V9XM-MNJ6].

6. *See* Michael Fertik, *Your Future Employer Is Watching You Online. You Should Be, Too.*, HARV. BUS. REV. (Apr. 3, 2012), https://hbr.org/2012/04/your-future-employer-is-watchi [https://perma.cc/XZ58-D5DC]; Meister, *supra* note 1.

7. *See* Don Peck, *They're Watching You at Work*, ATLANTIC (Dec. 2013), https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/ [https://perma.cc/92WJ-6VUD].

8. *See, e.g.*, Esther Kaplan, *The Spy Who Fired Me: The Human Costs of Workplace Monitoring*, HARPER'S MAG., Mar. 2015, at 31-32, 35.

information about their offline behaviors as well, reporting such things as how often they exercise or what they eat.[9] Data miners use this information to make health-related predictions, such as whether an employee is pregnant or trying to conceive.[10] Aggregating these various data sources can produce a rich and highly detailed profile of individual workers.[11]

This volume of information requires some form of automatic processing. No human brain can keep in view all of the thousands of data points about an individual. And so, algorithms are developed to make sense of it all—to screen, score, and evaluate individual workers for particular jobs. These algorithms are the tools of workforce analytics. For example, a company called Gild offers a "smart hiring platform" to help companies find "the right talent quicker."[12] Gild uses an algorithm that

> crunches thousands of bits of information in calculating around 300 larger variables about an individual: the sites where a person hangs out; the types of language, positive or negative, that he or she uses to describe technology of various kinds; self-reported skills on LinkedIn; [and] the projects a person has worked on, and for how long

as well as traditional criteria such as education and college major.[13]

Other services screen large pools of applicants, automating the

---

9. *See generally* Jay Hancock, *Workplace Wellness Programs Put Employee Privacy at Risk*, CNN (Oct. 2, 2015, 12:37 PM), http://www.cnn.com/2015/09/28/health/workplace-wellness-privacy-risk-exclusive/ [https://perma.cc/X9RY-X4VZ].

10. *See* Valentina Zarya, *Employers Are Quietly Using Big Data to Track Employee Pregnancies*, FORTUNE (Feb. 17, 2016, 5:36 PM), http://fortune.com/2016/02/17/castlight-pregnancy-data/ [https://perma.cc/MA3W-DDZQ].

11. *See, e.g.*, Peck, *supra* note 7; Sanjeev & Sandeep Sardana, *Big Data: It's Not a Buzzword, It's a Movement*, FORBES (Nov. 20, 2013, 12:05 PM), http://www.forbes.com/sites/sanjeevsardana/2013/11/20/bigdata/ [https://perma.cc/PU97-VFA9].

12. *Our Story*, GILD, https://www.gild.com/company [https://perma.cc/Q8QF-RPGB]; *see* Matt Richtel, *How Big Data Is Playing Recruiter for Specialized Workers*, N.Y. TIMES (Apr. 27, 2013), http://www.nytimes.com/2013/04/28/technology/how-big-data-is-playing-recruiter-for-specialized-workers.html [https://perma.cc/XAF6-SKXC].

13. Richtel, *supra* note 12; *see also* Vivian Giang, *Why New Hiring Algorithms Are More Efficient—Even If They Filter Out Qualified Candidates*, BUS. INSIDER (Oct. 25, 2013, 10:51 AM), http://www.businessinsider.com/why-its-ok-that-employers-filter-out-qualified-candidates-2013-10 [https://perma.cc/3XLE-GH6V] (describing how Bright.com uses "data and algorithms to match candidates up with potential jobs and hiring managers with star performers").

process of selecting the most promising candidates for employers.[14] One company examines hundreds of variables about job seekers, analyzes a firm's past hiring practices, and then recommends only those applicants it believes the employer will be interested in hiring. Other firms are developing computer games that record thousands of data points about how individuals play, such as what decisions they make and how long they hesitate before deciding, in order to uncover patterns that can identify successful employees.[15] Employers can then use these tools to make hiring or promotion decisions.

The actual impact on employment opportunities is difficult to document because information about how developers construct these algorithms is considered proprietary, and personnel data is confidential. Nevertheless, some publicly available examples suggest there is reason for concern. One company seeking to identify which employees would stay longer found that the distance between home and the workplace is a strong predictor of job tenure.[16] If a hiring algorithm relied on that factor, it would likely have a racially disproportionate impact, given that discrimination has shaped residential patterns in many cities. Other studies involving internet advertising illustrate how algorithms that learn from behavioral patterns can discriminate. For example, Latanya Sweeney has shown that Google searches for African American-associated names produce more advertisements for criminal background checks than searches for Caucasian-associated names, likely reflecting past patterns in users' search behavior.[17] Amit Datta, Michael Carl Tschantz, and Anupam Datta have demonstrated gender differences in the delivery of online ads to jobseekers, with identified male users "receiv[ing] more ads for a career coaching service that promoted high pay jobs," while female users received more generic ads.[18] Similarly, a field study by Anja Lambrecht and Catherine

14. *See* Miller, *supra* note 1.

15. *See, e.g.*, Peck, *supra* note 7.

16. *See* Dustin Volz, *Silicon Valley Thinks It Has the Answer to Its Diversity Problem*, ATLANTIC (Sept. 26, 2014), http://www.theatlantic.com/politics/archive/2014/09/silicon-valley-thinks-it-has-the-answer-to-its-diversity-problem/431334/ [https://perma.cc/VA6N-6W53].

17. *See* Latanya Sweeney, *Discrimination in Online Ad Delivery*, COMM. ACM, May 2013, at 44, 46-47.

18. *See* Amit Datta, Michael Carl Tschantz & Anupam Datta, *Automated Experiments on Ad Privacy Settings*, PROC. ON PRIVACY ENHANCING TECHS., Apr. 2015, at 92, 92-93; *see also* Amit Datta, Anupam Datta, Deirdre K. Mulligan & Michael Carl Tschantz, Discrimination

Tucker revealed that an internet ad for STEM (science, technology, engineering and math) jobs was far less likely to be shown to women than men.[19] These examples did not necessarily result from intentional bias, but the discriminatory effects were nevertheless real.

While workforce analytics are transforming employers' personnel practices, the legal world has only just begun to take notice. Privacy law scholars have raised concerns about the growth of big data, asking what limits the law should place on the collection of particularly sensitive personal information, or whether it should regulate "data flows" or downstream uses of this information.[20] Although much of the focus has been on problems caused by inaccurate data records or unexpected and invasive uses of sensitive personal information,[21] these scholars have also sounded alarms that big data may produce biased outcomes. Of the handful of commenters who have addressed the employment context, most have simply raised questions about the discriminatory potential of data analytics,[22] without deeply theorizing the nature of the harms that these technologies threaten for workers. And to the extent that legal scholars have considered how the law might respond, they have confined their analysis to narrowly applying existing doctrine.[23]

---

in Online Personalization: A Multidisciplinary Inquiry 3-5 (Mar. 13, 2016) (unpublished manuscript) (on file with author) (describing experiment and analyzing possible legal response).

19. *See* Anja Lambrecht & Catherine Tucker, Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads 2, 10-12 (Oct. 13, 2016) (unpublished manuscript), https://ssrn.com/abstract=2852260 [https://perma.cc/3PGF-CVTW].

20. *See, e.g.*, Danielle Keats Citron & Frank Pasquale, Essay, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4, 7-8, 18-22 (2014); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 94-96, 98-99, 101, 103-09, 123-27 (2014). *See generally* Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 409 (2014); Neil M. Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 20 STAN. TECH. L. REV. (forthcoming 2017), https://ssrn.com/abstract=2655719 [https://perma.cc/58A8-SCZB].

21. *See* Citron & Pasquale, *supra* note 20, at 4; Crawford & Schultz, *supra* note 20, at 96-99.

22. *See, e.g.*, EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 51-53 (2014), https://perma.cc/LE9N-PA9D; Citron & Pasquale, *supra* note 20, at 4; danah boyd & Kate Crawford, Six Provocations for Big Data (Sept. 21, 2011) (unpublished manuscript), https://ssrn.com/abstract=1926431 [https://perma. cc/3JUG-FAFJ]; Alex Rosenblat, Kate Wikelius, danah boyd, Seeta Peña Gangadhoran & Corrine Yu, Data & Civil Rights: Employment Primer (Oct. 30, 2014) (unpublished manuscript), https://ssrn.com/abstract=2541512 [https://perma.cc/398Q-F5MZ].

23. *See, e.g.*, Barocas & Selbst, *supra* note 5, at 694-712 (applying existing Title VII doc-

Workforce analytics pose an entirely new set of challenges to equality that calls for fundamentally rethinking antidiscrimination doctrine. Proponents of workforce analytics argue that data models can avoid reliance on biased human decision-making.[24] Skeptics warn that data is not neutral and that workforce analytics threaten to introduce new forms of bias or exacerbate existing ones.[25] But there is a third possibility as well—employers and researchers can use data to diagnose where and how cognitive or structural biases are currently operating in ways harmful to disadvantaged groups. Thus, the impact of workforce analytics will depend to a large extent on the choices that are made about how to deploy these technologies. And those choices will be shaped in turn by the legal environment in which firms operate.

The harms threatened by biased algorithms are not easily captured by traditional antidiscrimination law, which tends to focus on a specific "bad actor" and individual victims. Of course, a prejudiced employer might hide its discriminatory intent behind a biased data model. Such a scenario poses no particular conceptual challenge, although proof may be difficult as a practical matter. Even without any deliberate intent, a model may be biased in the statistical sense. Choices in the coding of information, errors in the data, reliance on unrepresentative samples, or the selection of variables for exclusion or inclusion might produce a model that is inaccurate in a systematic way.[26] When those systematic errors coincide with protected class status and operate to reduce opportunities for already disadvantaged groups, it should trigger the same concerns about workplace equality that motivated antidiscrimination laws.

The nature of algorithmic decision-making raises particular concern when employers rely on these models to make personnel decisions. Data mining techniques used to build the algorithms seek to uncover any statistical relationship between variables present in the data, regardless of whether the reasons for the relationship are understood. As a result, if employers rely on these models, they may deny employees opportunities based on unexplained correlations and make decisions that turn on factors with no clear causal

---

trine).

24. *See supra* note 4 and accompanying text.

25. *See supra* note 5.

26. *See infra* Part I.B.

connection to effective job performance. Because of limited opportunities for error correction, and the possibility of reinforcing feedback effects, these models may not only introduce but actually worsen bias and inequality. Given these risks, the law ought to be concerned with what I call "classification bias." Classification bias occurs when employers rely on classification schemes, such as data algorithms, to sort or score workers in ways that worsen inequality or disadvantage along the lines of race, sex, or other protected characteristics.

Classification bias may seem amenable to challenge under disparate impact doctrine, which targets facially neutral employment practices that have disparate effects on racial minorities or other protected classes.[27] However, a mechanical application of existing disparate impact doctrine will fail to meet the particular risks that workforce analytics pose. That doctrine evolved to address employer use of tests purporting to measure workers' abilities, and therefore focused on the validity of those measures and their relevance to a particular job.[28] In contrast, data mining models do not rest on psychological or any other theories of human behavior. Instead, these models simply mine the available data, looking for statistical correlations that connect seemingly unrelated variables, such as patterns of social media behavior, with workplace performance.[29] As a result, they pose a different set of risks—risks that existing doctrine does not address well.

As an example, disparate impact doctrine provides a defense if an employer can show that a test is "job related ... and consistent with business necessity."[30] In the case of workforce analytics, the data algorithm by definition relies on variables that are correlated in some sense with the job. So to ask whether the model is "job related" in the sense of "statistically correlated" is tautological. The more important question in the context of data mining is what does the correlation mean? Is the statistical relationship it uncovers causal, such that it provides a reliable basis for predicting future behavior?

---

27. *See* Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(2) (2012); Griggs v. Duke Power Co., 401 U.S. 424, 430-31 (1971).

28. *See* Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 755-60 (2006).

29. *See* Fertik, *supra* note 6.

30. *See* Civil Rights Act of 1991 § 105, 42 U.S.C. § 2000e-2(k)(1)(A)(i).

Or does it result from erroneously coded information, an unrepresentative sample, omitted variable bias, or some other data problems? Because the risks to workplace equality posed by data mining algorithms arise from different sources, existing disparate impact doctrine will not be adequate to address the risks they pose.

Addressing the possibilities and risks of data analytics for workplace equality requires taking a fresh look at antidiscrimination law, unencumbered by the specific doctrinal details that have developed under Title VII. Revisiting the statutory text suggests that Title VII directly prohibits classification bias. More specifically, section 703(a)(2) forbids employer practices that "classify" employees or applicants "in any way which would deprive or tend to deprive" them of employment opportunities because of protected class characteristics.[31] By focusing on the consequences of employers' classification schemes, this reading offers a more relevant frame for addressing the challenges that workforce analytics pose.

Thinking about the problem in terms of classification bias leads to some quite different conclusions about how the antidiscrimination norm should apply to data models.[32] For example, if the goal is to discourage classification bias, then the law should not forbid the inclusion of race, sex, or other sensitive information as variables, but seek to preserve these variables, and perhaps even include them in some complex models.[33] Similarly, this perspective suggests that those who use data mining models should bear the burden of demonstrating the accuracy and representativeness of the data used to construct the models, rather than requiring complainants to identify the flaws giving rise to biased outcomes.[34]

Addressing the challenges of workforce analytics using a theory of classification bias also reveals the limitations of the backward-looking, liability-focused model of legal regulation embodied by Title

---

31. The full text of subsection (a)(2) reads:

> It shall be an unlawful employment practice for an employer ... to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.

Civil Rights Act of 1964 § 703(a)(2), 42 U.S.C. § 2000e-2(a)(2).

32. *See infra* Part III.C.

33. *See infra* Part III.C.1.

34. *See infra* Part III.C.3.

VII.[35] Because of the diffuse nature of the harms and the significant resources that would be required to challenge biased algorithms, it may be difficult to incentivize individual plaintiffs to enforce a prohibition on classification bias. Even more problematic, a strong liability regime intended to address the use of biased algorithms may discourage employers from trying to understand whether these tools have disparate effects or may discourage them from using algorithms at all. If the law swings too far in this direction, it would avoid the costs of biased algorithms but also eliminate any potential positive effects that data analytics might have on diagnosing and counteracting cognitive and structural biases already affecting workplaces. Resolving this dilemma may require looking beyond liability-focused legal models to alternatives such as ex ante regulation, licensing models, or the development of technological solutions.

In considering the impact of data analytics on workplace equality and the appropriate legal response, this Article proceeds as follows. Part I surveys the psychological and structural factors that contribute to bias in the contemporary workplace and considers the potential for data models to eliminate that bias. Replacing human decision makers with a computer algorithm may prevent certain types of cognitive biases from operating but is unlikely to reach other types of structural disadvantage that may result from the way work is organized. At the same time, widespread reliance on decision-making algorithms risks introducing new forms of bias or exacerbating existing ones. Part I surveys those risks, catalogues the types of harm that may result from reliance on algorithms in the workplace, and then argues for recognizing classification bias as a distinct type of threat to workplace equality.

Part II considers whether other responses—aside from antidiscrimination law—can effectively address classification bias and concludes that neither market forces nor traditional forms of privacy protection are likely to be successful. The nature of labor markets are such that employers will not reliably receive signals if their employment practices produce bias against minority groups. And privacy protections typically focus on individual harms rather than addressing the group-based disadvantages that are the principal concern of antidiscrimination law.

35. *See infra* Part III.E.

In Part III, I consider the limits and possibilities of existing antidiscrimination law. Mechanical application of existing Title VII doctrine is unlikely to be successful in addressing the equality challenges that workforce analytics pose. Neither disparate treatment nor current disparate impact doctrine completely captures the types of risks threatened by data models. Instead, antidiscrimination law should be adapted to meet these unique risks. Part III argues that a close reading of the statutory text shows that Title VII *does* prohibit classification bias, and considers what a robust response to this form of discrimination should look like.

More specifically, it argues that an effective legal response will depart from traditional disparate impact doctrine in several ways. For example, employers should not be able to justify reliance on a biased model merely by showing a statistical relationship but should bear the burden of showing that the model is statistically valid and substantively meaningful. At the same time, an employer should be permitted to rely on a "bottom-line" defense if its use of a model as part of a larger selection process does not produce discriminatory results.

After considering how the law should respond, Part III briefly explains why the Supreme Court's decision in *Ricci v. DeStefano* poses no obstacle to enforcing a prohibition on classification bias. Finally, it considers the limitations of classification bias theory and suggests some alternatives to a liability-based regime.

I. THE IMPACT OF DATA ANALYTICS ON WORKPLACE EQUALITY

A. The Promise of Workforce Analytics

The use of data analytics offers the potential to *reduce* bias in employment. Proponents of the technology claim that algorithms do just that by eliminating the subjective biases and personal predilections of a human resources manager. For example, the goal of Gild is "to build machines that ... eliminate human bias."[36] Pointing to the many ways in which human decision-making is biased, these services offer to find overlooked talent that better matches a company's needs and, in turn, to produce a more diverse workforce.

36. *See* Richtel, *supra* note 12.

These claims are consistent with scholarly accounts of how human bias distorts personnel decisions, even in the absence of a conscious discriminatory motive.[37] Charles Lawrence argues that unconscious prejudices may lead to discrimination even when the decision maker is unaware of, and would disclaim, any prejudicial intent.[38] Similarly, Linda Krieger and other scholars explain how ordinary cognitive processes naturally lead people to create mental categories.[39] When these categories coincide with race or gender differences, they can distort the perceptions of supervisors and managers in ways that tend to confirm societal biases. More recently, a great deal of attention has focused on implicit bias.[40] Scholars point to the results of the Implicit Associations Test to argue that people typically associate negative characteristics more strongly with disfavored groups.[41] These negative associations can result in adverse decisions for members of those groups, even when the decision maker intends to act fairly and believes that she is doing so.[42]

Although these theories differ as to the precise mechanism at work, they are alike in pointing to processes that occur outside of conscious awareness. They suggest that automatic processes—the ways in which our brains naturally function—can produce biased judgments. As a result, these effects are not readily visible to the decision maker, even upon self-reflection.[43] Individuals who strongly embrace nondiscrimination and equality norms may be particularly

37. *See, e.g.*, Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317, 322 (1987).

38. *See id.*

39. *See, e.g.*, Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1186-87 (1995).

40. *See, e.g.*, R. Richard Banks, Jennifer L. Eberhardt & Lee Ross, *Discrimination and Implicit Bias in a Racially Unequal Society*, 94 CALIF. L. REV. 1169 (2006); Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CALIF. L. REV. 945 (2006); Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CALIF. L. REV. 969 (2006); Jerry Kang, *Rethinking Intent and Impact: Some Behavioral Realism About Equal Protection*, 66 ALA. L. REV. 627 (2015); Jerry Kang & Kristin Lane, *Seeing Through Colorblindness: Implicit Bias and the Law*, 58 UCLA L. REV. 465 (2010).

41. For reviews of the social science literature on implicit bias, see Greenwald & Krieger, *supra* note 40, at 951-58; Kang & Lane, *supra* note 40, at 473-81.

42. *See, e.g.*, Kang & Lane, *supra* note 40, at 468-89.

43. *See* Lawrence, *supra* note 37, at 336-39; *see also* Krieger, *supra* note 39, at 1217.

resistant to recognizing the operation of bias in their mental processing because of the cognitive dissonance that would result.[44]

The claim of workforce analytics is that algorithms can replace fallible human judgments with neutral, unbiased data to improve decision-making.[45] The chief scientist at Gild put it this way: "Let's put everything in and let the data speak for itself."[46] The proponents of data science are right to point out that traditional employment practices—relying as they often do on subjective assessments, intuition, and limited human cognition—may entail considerable amounts of bias. However, as discussed in Part I.B below, algorithms are not always neutral either. Depending on the choices made in collecting and coding information and building models, data analytics risk replicating existing biases or introducing new ones.[47] So although algorithms offer the potential for avoiding or minimizing bias, the real question is how the biases they may introduce compare with the human biases they avoid.

Whatever their promise for eliminating cognitive biases, algorithms will not counteract structural forms of workplace bias. This type of bias results not from cognitive processes but from structural forces that shape opportunities differently for different types of people. Numerous scholars have argued that workplaces are often organized in ways that systematically disadvantage women or minorities.[48] For example, when training and advancement opportunities are informally distributed in a firm through social networks, women or racial minorities who have less extensive networks may be disadvantaged. Similarly, work that requires long hours or unpredictable schedules may place particular burdens on women, who are often the primary caretakers of their children. These types of choices about workplace organization may not reflect intent to exclude, and, therefore, like the cognitive processes described above,

---

44. *See* Lawrence, *supra* note 37, at 337.

45. *See* Richtel, *supra* note 12.

46. *Id.* (quoting Vivienne Ming, chief scientist at Gild).

47. *See supra* note 5.

48. *See, e.g.*, Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 11 (2006); Tristin K. Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. C.R.-C.L. L. REV. 91, 104 (2003); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 468-74 (2001).

their impact on disadvantaged groups is not readily visible to managers.[49]

Relying on data models instead of human decision-making is unlikely to counter structural forms of bias because these models take existing workplace structures as givens. For example, if reduced access to social networks in a firm hampers minority employees' chances of promotion, relying on data to make those promotion decisions will not remedy the fact that minority employees are receiving less mentoring and training. Similarly, data-driven hiring decisions will not alter the reality that unpredictable work schedules will take a greater toll on workers with caregiving responsibilities, who are more often women. Thus, merely relying on data analytics instead of human judgments will not address forms of disadvantage that result from biased workplace structures.

On the other hand, data can be a useful tool for *diagnosing* both cognitive and structural forms of bias. Rather than using workforce analytics to *make* decisions, firms could deploy close analysis of employment-related data to assess the decision-making process itself, thereby uncovering hidden biases and prompting efforts to counteract them. One service, Textio, used language analysis to determine that certain phrases in job postings—for example, military analogies like "mission critical"—appear to reduce the proportion of women who apply.[50] Employers committed to recruiting a diverse workforce might learn how to craft language likely to attract a more diverse applicant pool from such a program. Cognitive science teaches that individuals tend to remember facts that confirm their preexisting beliefs about the world.[51] Krieger and others explain how this phenomenon might lead supervisors to remember negative information about members of disfavored groups but to disregard similar information about in-group members.[52] Data could be a useful corrective to such biased perceptions, highlighting for managers when their recall about particular workers may be faulty.

Supervisors who are not themselves biased might nevertheless fail to recognize how earlier discriminatory decisions continue to shape current outcomes. An initial discriminatory decision that

49. *See id.* at 470-71.
50. *See* Miller, *supra* note 1.
51. *See* Krieger, *supra* note 39, at 1203.
52. *See id.* at 1209.

created a pay differential between men and women can have effects years later, even if every subsequent decision regarding individual raises is entirely fair and neutral.[53] A current supervisor, having directly observed only unbiased decisions in recent years, might view the differential in wages as justified. By decomposing the factors contributing to current salary or by comparing salary to discrete measures of productivity, data analysis might make visible the current effects of past discrimination, rather than allowing those outcomes to appear natural and inevitable.

Employers can also use data to identify sources of structural bias that disadvantage certain groups. In the example cited in the introduction, Evolv, the company that identified the distance between home and the workplace as a predictor of employee job tenure, decided not to use this factor in its hiring algorithm because it understood that housing patterns are correlated with race and that relying on that correlation might result in discrimination.[54] In addition to eliminating the factor as a basis for decision-making, an employer might use the information to examine whether its workplace practices make it more difficult for employees who travel long distances to succeed. A firm committed to a diverse workforce but located in a city with a segregated housing market might consider policies like flex-time or benefits like public transit passes in order to relieve a commuting burden that falls more heavily on already disadvantaged groups.

Data analytics thus hold the potential to reduce biases and increase opportunities in the workplace for traditionally disadvantaged groups. But much depends on how data are used. When employers use analytics to evaluate personnel policies and procedures, data can help to diagnose where workplace structures or

53. Consider, for example, the facts in *Ledbetter v. Goodyear Tire & Rubber Co.*, 550 U.S. 618, 621-22 (2007). The plaintiff in that case, Lilly Ledbetter, worked for Goodyear Tire for nearly twenty years. *Id.* at 621. She alleged that several supervisors had given her poor evaluations because of her sex and that those discriminatory evaluations continued to result in her receiving lower pay than her male counterparts throughout her employment with the defendant. *Id.* at 622. The Supreme Court dismissed her claims on the grounds that no discriminatory pay decisions had been made during the statutory "charging period"—the last 180 days before she filed with the Equal Employment Opportunity Commission. *Id.* at 624-32. Congress eventually overturned the decision in the Lily Ledbetter Fair Pay Act of 2009, Pub. L. No. 111-2, § 3, 123 Stat. 5, 5-6 (2009) (amending 42 U.S.C. § 2000e-5(e)).

54. *See* Volz, *supra* note 16.

organizations inadvertently disadvantage or exclude members of certain groups. Relying on data analytics to sort applicants and employees may also reduce bias if these models are less biased than the subjective human decision makers they replace. Whether that is the case, however, depends a great deal on how the algorithms are constructed and deployed. As the next Section explores, there are numerous reasons to be concerned that workplace analytics may introduce bias or worsen existing patterns of disadvantage.

B. The Risks of Workforce Analytics

Although data analytic tools offer the potential for countering biased decision-making processes and workplace structures, these same tools also risk reinforcing existing discrimination or introducing new forms of bias. Employers have long used data to sort and rank workers—for example, through preemployment tests, psychological screens, or productivity requirements. These traditional uses of data metrics to measure and evaluate can raise concerns about bias, and they have faced legal challenges.[55] However, the new workforce science poses distinct risks. With traditional forms of testing, employers generally started by identifying skills or attributes thought relevant to job performance and then relied on test professionals to develop measures of those skills or attributes. These forms of testing collected limited amounts of targeted information about applicants or employees. In contrast, data models today take advantage of the vastly greater quantity of data available and mine it to discover novel correlations. That data may contain information about attributes or behaviors, such as social media usage, that have no clear connection with job performance.

In order to build a model, its creators must select the data that they will use to build it—the "training data."[56] The actual data mining occurs when the data are analyzed using statistical techniques

55. *See, e.g.*, Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975); Griggs v. Duke Power Co., 401 U.S. 424 (1971).

56. *See* Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY: DATA MINING AND PROFILING IN LARGE DATABASES 3, 3-4 (Bart Custers, Toon Calders, Bart Schermer & Tal Zarsky eds., 2013).

to uncover patterns.[57] The data miner is not testing any particular hypotheses or explanations; instead, the process reveals statistical relationships among variables present in the data.[58] What the data miner finds thus depends on the data examined. The correlations may be causal or the relationship may be entirely coincidental.[59] Data mining is generally unconcerned with the reasons for the correlation.[60] So long as the relationships discovered are thought to be robust, the data model may use them to classify or predict future cases.[61] So, for example, a data model might find that individuals who "like" certain items on Facebook have higher intelligence.[62] Data mining cannot explain this relationship, but a model may nevertheless predict that applicants who share that characteristic are better workers and recommend their selection over those who do not.

In their article *Big Data's Disparate Impact*, Solon Barocas and Andrew Selbst provide a taxonomy of ways that the data mining process can result in adverse impact on protected groups.[63] One of the first steps in building a model is identifying the target variable—in other words, defining the outcome of interest[64]—and defining which outcomes are desired by categorizing them.[65] Doing so in the employment context is not simple. Unlike credit card charges, which can be categorized with complete certainty as fraudulent or not, the category of "good employee" is not self-

57. *See id.* at 9 ("[T]he ... data-mining stage ... [occurs when] the data are analyzed in order to find patterns or relations. This is done using mathematical algorithms.").

58. *See id.* at 7 (explaining that data mining differs from traditional statistical analysis, which begins with a hypothesis, because data mining generates hypotheses from the data itself).

59. *See id.* at 16-17.

60. *See id.*

61. *See id.* at 16.

62. *See, e.g.*, Michal Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. NAT'L ACAD. SCI. U.S. 5802, 5805 (2013) (showing that records of an individual's Facebook "likes" can be used to accurately predict personal characteristics such as race, gender, sexual orientation, religious and political views, and intelligence); *see also* Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 56, at 43, 45-47.

63. *See* Barocas & Selbst, *supra* note 5, at 677-93.

64. *See id.* at 678.

65. This process is referred to as defining "class labels" for the target variable. *See id.* at 678-79.

evident.[66] In order to build a model, the meaning of "good employee" must be specified in a way that the machine can understand, namely "in ways that correspond to measurable outcomes: relatively higher sales, shorter production time, or longer tenure, for example."[67] Using a more holistic definition of "good" would require someone to create a measure that captures that quality and to apply it to particular individuals in order for the machine to know what it is looking for in future cases.[68]

As Barocas and Selbst explain, this process of classifying individuals risks reintroducing the human biases the data analysts are seeking to avoid.[69] If the data miner chooses to rely on only "objective" measures for the target variable, this will introduce bias of a different kind, by valuing quantifiable measures of performance over softer skills like leadership or collaboration. In order to build a predictive model, the data miner must label and classify the training data—a "necessarily subjective process of translation"[70]—and these choices may introduce biases against protected groups.[71]

The selection of the training data will affect the outcome of the model as well. As Barocas and Selbst explain, "what a model learns depends on the examples to which it has been exposed."[72] The training data may incorporate biased judgments, as, for example, when they include supervisors' evaluations or previous hiring decisions that were colored by prejudice or distorted by cognitive bias.[73] Because the model will accept those characterizations "as ground truth,"[74] it will inevitably reflect those biases in the outcomes it produces. Factual errors may exist in the data as well, and those errors may be more frequent for members of certain groups, rendering the model less accurate when applied to members of those groups.[75] Another concern is that the data may be unrepresentative in that different groups are not represented in proportion to their

66. *See id.* at 679.
67. *Id.*
68. *See id.*
69. *Id.* at 680.
70. *Id.* at 678.
71. *See id.*
72. *Id.* at 680.
73. *See id.* at 682.
74. *Id.*
75. *See id.* at 684; EXEC. OFFICE OF THE PRESIDENT, *supra* note 22, at 52.

presence in the population.[76] Big datasets, which often supply the training data for workforce analytics, are more likely to exclude members of minority groups and disadvantaged populations, those "who live on big data's margins ... and whose lives are less 'datafied' than the general population's."[77] If the data collection process systematically captures less information about certain groups, then the resulting decision-making algorithm may produce biased results.[78] Barocas and Selbst offer the example of an employer that relies on data about online expressions of interest to target its recruitment efforts.[79] Because of differences in access to broadband in different communities, relying on such data may cause an employer to underestimate the level of interest and qualifications in underrepresented communities. A recruiting strategy based on such data is likely to produce biased outcomes.

Barocas and Selbst identify several other mechanisms by which data models may produce biased outcomes. The process of "feature selection"—choosing which attributes to include in the analysis—can have "serious implications for the treatment of protected classes."[80] If the attributes that explain variation within a protected class are not incorporated, the model may be unable to distinguish among members of the group, leading it to rely on broad generalizations that disadvantage individual members of the group.[81] Data models may also discriminate when neutral factors act as "proxies" for sensitive characteristics like race or sex.[82] Those neutral factors may be highly correlated with membership in a protected class, and also correlate with outcomes of interest.[83] In such a situation, those neutral factors may produce results that systematically disadvantage protected groups, even though the model's creators have no discriminatory intent, and the sensitive characteristics have been removed from the data.[84] Finally, Barocas and Selbst point out that

76. *See* Barocas & Selbst, *supra* note 5, at 684.

77. Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55, 57 (2013); *see also* Crawford, *supra* note 5.

78. *See* Barocas & Selbst, *supra* note 5, at 684-86.

79. *Id.* at 685.

80. *Id.* at 688.

81. *See id.* at 689-90.

82. *See id.* at 691-92.

83. *See id.* at 691.

84. *See id.*

employers may use data models to intentionally discriminate against certain groups. Because data mining can often infer protected class status from other neutral variables, employers could use data analytics as cover for intentional discrimination.[85]

In addition to the mechanisms that Barocas and Selbst identify, other characteristics of data models raise particular concerns when employers rely on them to make personnel decisions. Contrasting data mining techniques with traditional social science methodologies illuminates the problems. Social scientists articulate theories about the world, develop hypotheses based on those theories, and then subject those hypotheses to rigorous empirical testing, often by using data.[86] Their goal is to understand and explain patterns observed in the world.[87] An important part of designing an empirical test is determining what population the data should be drawn from and what variables should be included in the statistical model.[88] The theory motivating the study informs each of these decisions and each decision is consequential for the accuracy of the results.[89]

Suppose a researcher has a theory that past military service makes employees more successful in managerial positions. Testing this hypothesis will require examining how military service and on-the-job success are related using data about a representative group of workers. Looking only at those two variables might suggest that military service is negatively associated with future job performance. But a social scientist would also want to include other variables that could independently influence job performance. Unless the researcher controls for these factors, the study might reach an erroneous conclusion—a problem referred to as "omitted variable bias."[90] If military recruits are significantly less educated than the rest of the population, looking only at the relationship between service and later job performance could be misleading. Including a variable for education in the model might show that

85. *See id.* at 692-93.

86. *See* Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 19-20 (2002).

87. *See id.* at 20-21, 60-61.

88. *See id.* at 54-55, 99-102.

89. *See id.*

90. *See id.* at 78.

military service is in fact associated with better job performance, after controlling for an individual's level of education.

The concern about omitted variable bias can apply to sensitive characteristics like race and sex in some circumstances. To extend the example above, suppose that for African Americans military service is highly positively correlated with subsequent work performance, while for white workers it has a somewhat negative effect. If the dataset includes far more observations about white workers, then a statistical model that omits race as a variable might predict that workers with past military service are less successful employees, even though the opposite is true for African Americans. If an employer relied on the model to disfavor workers with military experience, then the failure to include race as a control variable would ultimately disadvantage African Americans.

The solution is not to throw every possible variable into the statistical model.[91] Including too many variables might also bias results, especially if some variables are highly correlated. In such a situation, real effects are obscured, suggesting that no relationship exists among variables that are in fact related. Thus, for a social scientist trying to accurately describe relationships and effects in the real world, choices about which variables to include are crucial. Because the results of a statistical model are very sensitive to those choices, the norms of social science dictate that researchers be transparent about their choices and justify them by reference to the theory motivating the study. Those norms also encourage data sharing, to allow other researchers to replicate the study, to further test the results, and to criticize and revise the findings when necessary.

In contrast, data mining is inductive and atheoretical.[92] Data miners have no particular theory they are trying to test, nor are they necessarily interested in explaining observed relationships between different variables. Instead, data mining exploits enormous datasets with thousands of variables to uncover whatever statistical correlations might exist in the data. With no motivating theory to

91. *See id.* at 79-80.

92. *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 12-14 (2013) (explaining that big data shifts the focus from discovering causal relationships to uncovering patterns in the data); Custers, *supra* note 56, at 16.

justify the choices made, it is difficult to assess whether the data relied on is sufficiently representative, or whether the appropriate variables have been included to ensure the accuracy of the model. And in the absence of data sharing or transparency about the choices made in constructing the model, others cannot test the robustness and validity of the results.

Concerns that a data model may systematically disadvantage traditionally protected groups cannot be resolved simply by eliminating protected characteristics like race and sex from the data. As Barocas and Selbst explain, other types of information that closely correlate with those protected characteristics may serve as proxies, producing the same results without expressly relying on those categories.[93] At the same time, the possibility of omitted variable bias means that excluding race and gender variables will sometimes increase the risk of bias by failing to capture relevant differences between groups. The remedy is therefore not to exclude or include variables for sensitive characteristics in every case.

Because data mining is concerned only with identifying relationships, the model's creators often do not know whether correlations that are uncovered represent genuine relationships between factors in the real world or are artifacts of the data mining process. Social scientists expend a great deal of effort trying to determine whether an observed relationship between variables is causal. Because of the difficulty of establishing causality through statistics alone, a claim that two variables are related is subject to retesting and constantly open to challenge. By contrast, data mining models make predictions based on the strength of the statistical correlation alone.

In some contexts, we may not care much about the limitations of data mining. For example, if a computer algorithm can correctly flag which purchases made on my credit card are fraudulent and notify me, it does not matter whether I, or my bank, understand which variables triggered the alert or why. The difference between correlation and causation becomes important, however, if employers are basing their decisions on these statistical relationships. Suppose, for example, that data mining shows a strong statistical relationship between intelligence and "liking" curly fries on Facebook.[94] An

93. *See, e.g.*, *supra* notes 82-83 and accompanying text.

94. *See* Kosinski et al., *supra* note 62, at 5804.

employer seeking highly intelligent employees might justify reliance on that correlation in selecting employees, even if it has a racially disproportionate effect, on the grounds that intelligence is a relevant job criterion.[95] If, however, the variables are merely correlated and not causally related, there is no necessary connection between them, and the correlation may not hold in the future. An employer relying on the statistical correlation may continue to make decisions disadvantaging minority applicants, even after the statistical relationship no longer holds true. Although it may seem clear that "liking" curly fries is not causally related to intelligence, in other cases it will not be intuitively obvious whether a given correlation is meaningful or spurious. But the same risk is present—that the algorithm is relying on a factor that has a discriminatory effect but is not actually connected to job performance.

Another novel challenge posed by data mining models is their lack of transparency. Many algorithms are built using machine learning techniques, which do not require the human programmer to specify in advance which factors the model should consider or what weight each should be given. Instead, the computer constructs a model by exploiting the relationships it uncovers between variables in the data. These relationships may be quite complex, such that in some cases the resulting model is completely opaque, even to its creators. When such a model is relied on to screen or rank applicants, it obscures the basis on which employers are making ultimate employment decisions. This lack of transparency makes it difficult to know if any observed bias is simply a byproduct of justifiable business considerations or the result of flaws in the model's construction.

A related concern is that mechanisms to improve the accuracy of predictive models may not work in the context of employment. Big data enthusiasts often defend the use of algorithms on the ground that if the predictions are inaccurate, the machine will "learn" over time, such that any errors will be eliminated.[96] To return to the

95. The study by Kosinski and others also found that "liking" "I Love Being a Mom" is predictive of low intelligence. *Id.* The discriminatory impact on women that would result from relying on that apparent correlation is obvious.

96. *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 92, at 12 (arguing that big data systems can "improve themselves over time" by continuing to look for signals and patterns as they receive new data).

example of credit card fraud detection, if the algorithm makes an error in classifying a charge on my credit card, I will discover the error sooner or later and report it. My feedback will be incorporated and the model will update and refine its decision process, becoming more accurate over time.

When applied to employment decisions, however, the process of error detection and learning is far less likely to occur. In the case of credit card fraud, consumers can observe and report the error. In the case of employment decisions, not all types of errors will be observable. Suppose an employer relies on an algorithm to sort applicants into "qualified" and "unqualified" pools. After hiring an applicant, the employer can observe the new employee's work performance and will learn if the model made a mistake in classifying the applicant as qualified. However, if the algorithm mistakenly labeled an applicant as "unqualified," the employer will not hire her and therefore will never observe her work performance. As a result, there will be no opportunity to learn of the error and update the model.

Once bias enters the system, feedback loops may reinforce that bias. Recall the example of Google's algorithm which advertised criminal background checks more often when searches were conducted for African American-associated names than for Caucasian-associated names.[97] Those results likely reflect patterns in past search behavior, rather than any discriminatory bias on the part of the programmers who created the algorithm.[98] Nevertheless, the ads might nudge even the nonprejudiced employer, who otherwise would not treat applicants differently because of race, to scrutinize the criminal history of African American applicants more closely than white applicants. If, as a result of the nudge, the employer conducts criminal background checks more often for African American applicants than for white applicants, it will find more instances of criminal history in that population, further reinforcing a cycle of bias.

Feedback effects could also reinforce biased outcomes if disfavored groups are aware of the bias. If members of a particular group perceive that selection processes are systematically biased against

97. *See* Sweeney, *supra* note 17, at 46-47.

98. *See id.* at 52.

them and their chances of success are much less than for others, they may reduce their investment in developing their human capital.[99] This risk may be particularly significant if the patterns they observe suggest that the types of signals they have some control over—education, training, and the like—are not decisive and that other unknown or uncontrollable factors are shaping their employment opportunities.

Data mining models are thus far from neutral. Choices are made at every step of the process—selecting the target variable, choosing the training data, labeling cases, determining which variables to include or exclude—and each of these choices may introduce bias along the lines of race, sex, or other protected characteristics. Because of the atheoretical nature of data mining, once these biases are introduced, they may be difficult to detect and eliminate. Mere correlation may be mistaken for causation, and the true basis for employer decision-making is obscured. Moreover, these biases may persist or even worsen over time because of limited opportunities for error detection and the operation of feedback effects. For all of these reasons, identifying and addressing the potential harms that biased algorithms cause should be matters of policy concern.

C. Types of Harm

Although many scholars have raised alarms that data analytics can produce biased outcomes, they have not articulated the precise nature of the harms that biased algorithms impose, or explained why they should be matters of policy concern. A common assumption among critics is that any type of bias in an algorithm is normatively troubling and requires policy or legal interventions. However, this assumption is unwarranted and overly broad. Virtually any decision-making process will produce disproportionate effects, and sometimes those effects will fall along protected class lines. What matters are the reasons unequal outcomes are occurring and whether those reasons are normatively acceptable.

---

99. *See* Samuel R. Bagenstos, *Subordination, Stigma, and "Disability,"* 86 VA. L. REV. 397, 464 & n.254 (2000) (citing economics literature that discrimination can be self-perpetuating if it discourages members of groups facing discrimination from investing in their human capital).

In this Section, I explain why certain types of bias in data models produce cognizable harms. Barocas and Selbst's taxonomy, discussed in the last Section, sought to explain the specific technical issues that can cause data models to discriminate.[100] My focus in this Section is different—namely, to identify the different types of harm that might result when employers rely on biased data models. Because the nature of the harm depends in part on the source of the bias, my typology of harms partially overlaps, but does not coincide, with their taxonomy. In what follows, I identify four distinct types of equality harms that may occur when employers rely on data analytics to distribute employment opportunities.

1. Intentional Discrimination

One type of harm results when an employer uses data analytics to intentionally discriminate against a protected group.[101] In such a scenario, the employer relies on an algorithm to make hiring or promotion decisions because it *knows* the model produces a discriminatory result and *intends* that result to occur. The discriminatory decision simply masquerades behind the neutral façade of data analysis.[102] This type of discrimination is familiar as a form of intentional disparate treatment, only with the twist that the pretext—the "legitimate business reason" given for the decision—is the output of a computer model.

Although an employer might use data analytics as a screen for race or sex discrimination, an algorithm may be particularly effective in masking discrimination where the protected characteristic is not readily observable—for example, genetic traits and some kinds of disabilities. The law currently attempts to prevent these types of discrimination by restricting access to information about the protected characteristics. Thus, the Americans with Disabilities Act restricts an employer's ability to conduct medical exams or to

100. *See* Barocas & Selbst, *supra* note 5, at 677.

101. In Barocas and Selbst's taxonomy, this is referred to as "masking." *Id.* at 692. Other scholars also catalogue the different ways that an algorithm can enable intentional discrimination. *See, e.g.*, Dwork & Mulligan, *supra* note 5, at 36-38; Kroll et al., *supra* note 5 (manuscript at 32-34).

102. *See* Custers, *supra* note 56, at 9-10.

inquire about a disability prior to making a job offer,[103] and the Genetic Information Nondiscrimination Act forbids employers from seeking any kind of genetic information about applicants or employees.[104] An employer who believes that certain individuals are more costly to employ might use data profiles to identify and screen them out without ever explicitly asking for medical or genetic information. Several years ago, Target Stores used purchasing information to identify consumers who were in the early stages of pregnancy in order to send them coupons for baby products.[105] An employer with access to large amounts of behavioral data might similarly use that information to predict which applicants or employees pose future medical risks.[106]

When employers use data simply to mask intentional discrimination, the individual who loses out on an employment opportunity suffers the same type of harm as any other victim of intentional discrimination. The harm is direct and specific to the individual with the targeted characteristic.

2. Record Errors

A second type of harm arises when errors in an individual's record lead to the denial of an employment opportunity. For example, data collected from public sites might suggest that an individual has a criminal record or has defaulted on a loan, when in fact that is not true. The privacy literature, discussed in Part II.B, has focused on this type of harm. Inaccurate information does not inherently raise equality concerns, as errors may be randomly distributed, infecting the records of members of privileged groups as well as protected groups. However, evidence suggests that errors are more likely to

103. *See* Americans with Disabilities Act of 1990 (ADA) § 102, 42 U.S.C. § 12112(d)(2)(A) (2012) ("[A] covered entity shall not conduct a medical examination or make inquiries of a job applicant as to whether such applicant is an individual with a disability or as to the nature or severity of such disability.").

104. *See* Genetic Information Nondiscrimination Act of 2008 (GINA) § 202, 42 U.S.C. § 2000ff-1(b) ("It shall be an unlawful employment practice for an employer to request, require, or purchase genetic information with respect to an employee or a family member of the employee.").

105. *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [https://perma.cc/88XA-HHT7].

106. *See* Zarya, *supra* note 10.

occur for members of subgroups that are farther from the mainstream. For example, individuals whose names have less common spellings—most likely ethnic names—have greater rates of error in records relating to their employability.[107] Similarly, people with two surnames—disproportionately Hispanics—or who have changed their names—disproportionately women—are more likely to have inaccuracies in their records.[108]

When an algorithm makes a prediction based on error-ridden data about an applicant, it may unfairly deprive that individual of an employment opportunity. The overall operation of the model may be unbiased in the sense that it accurately predicts outcomes for individuals about whom it has reliable data. If, however, errors are not randomly distributed, then the model's predictions may be more likely to produce erroneous predictions for some, and could result in outcomes systematically biased against members of certain groups.[109] In such a situation, it is theoretically possible to identify individual victims who can be made whole by granting access to the opportunities they would have had absent the errors in their records.[110] Of course, significant practical challenges may make it difficult to detect when errors are present in an individual's records and to prove that they caused the adverse outcome. Although proof may be difficult, the harm is easily conceptualized—identifiable individuals have lost out on specific employment opportunities.

3. Statistical Bias

A third type of harm may result from data models that are statistically biased, in the sense that they systematically disfavor a protected class because of the way the underlying model was created. Social scientists refer to statistical bias when problems

107. *See* AM. CIVIL LIBERTIES UNION, PROVE YOURSELF TO WORK: THE 10 BIG PROBLEMS WITH E-VERIFY (2013), https://www.aclu.org/files/assets/everify_white_paper.pdf [https://perma.cc/9K8A-N8L5].

108. *See* EXEC. OFFICE OF THE PRESIDENT, *supra* note 22, at 52.

109. If record errors are pervasive for certain protected classes in the training data, they may also bias the model as a whole, such that even when applied to a population for whom accurate records are available, the outcomes will be biased. *See* Barocas & Selbst, *supra* note 5, at 684-85. This is a type of statistical bias, discussed in Part I.C.3.

110. *Cf.* Citron & Pasquale, *supra* note 20, at 4-5; Crawford & Schultz, *supra* note 20, at 101.

such as selection effects or omitted variables cause a model to be biased in the sense that it is systematically inaccurate in some way.[111] Similarly, data mining models built using biased, error-ridden, or unrepresentative data may be statistically biased.[112] Because of problems with the data or the model's construction, an algorithm may inaccurately capture relationships in the data, leading to imprecise or even erroneous predictions.

When statistical bias coincides with systematic disadvantage to protected classes, it causes discriminatory harm. The algorithm's creators may not have made the choices that produced the discriminatory effects with conscious intent to discriminate or even awareness of their biasing effects. Nevertheless, the resulting outcomes are not only biased in a statistical sense, but also in the colloquial sense of unfairly disadvantaging members of protected groups. The employer's practice has a discriminatory effect, and the statistical unreliability of the model undermines any justification for its use.

This type of bias, which results from the operation of a model, is structural in nature rather than individual. Correcting errors in the data about particular individuals will not solve the problem. Even if all the data used to predict future cases are entirely accurate, the algorithm produces results that are systematically biased against a protected group. The harm is also structural in the sense that it cannot be corrected for just one individual applicant or employee. The harmful effects on a protected group result from the operation of the model as a whole. This means that it may be difficult, if not impossible, to identify specific individual victims of discrimination.

Imagine a situation in which an employer relies on a biased algorithm to hire 100 employees from a pool of 1000 applicants. Suppose that 200 of the applicants (20 percent) are African American, but the employer only hires five. Of the 195 African American applicants who were not hired, it will be difficult to determine who *would* have been hired if the employer had not used the biased algorithm. Doing so requires making assumptions about what the model or the decision process would have looked like if constructed without the biasing choices. The difficulty is that there is not likely

111. *See* GARY KING, ROBERT O. KEOHANE & SIDNEY VERBA, DESIGNING SOCIAL INQUIRY: SCIENTIFIC INFERENCE IN QUALITATIVE RESEARCH 28 (1996).

112. *See* Barocas & Selbst, *supra* note 5, at 684-87.

to be a single unbiased alternative. Because model creation entails so many choices, multiple unbiased or less biased alternatives are possible, each of which might have selected a different set of individuals from the applicant pool for hire.

In the absence of a clear baseline against which to compare the outcomes, it is difficult to say that a particular individual in a protected class has been harmed while another has not. The harm to any given individual might be more accurately characterized as a reduction in their probability of selection rather than the loss of a job.[113] This uncertainty in identifying individual harms does not mitigate the fact that the operation of the model overall threatens a social harm if its effects are to entrench the disadvantage that subordinated groups experience.

4. Structural Disadvantage

Even in the absence of statistical bias, an algorithm may produce disproportionate effects on a protected class. It may accurately capture the relationships between various attributes in the data in a way that produces outcomes that systematically disadvantage certain groups.[114] Note that with data mining models using large datasets, it may be practically difficult, if not impossible, to rule out the possibility that statistical bias has caused the discriminatory effects. At least as a theoretical matter, however, it is possible that a model is not biased in the statistical sense, but its operation systematically disadvantages members of a protected class. It might do so because the members of the protected class in fact differ in some systematic way relevant to characteristics that the model is trying to predict.[115]

In such a case, whether a rejected applicant has been harmed depends upon societal judgments about the fairness of the model. And whether a model should be considered fair depends on what attributes it leverages to make its predictions and on the normative

113. For a similar argument that affirmative action programs should be understood as altering the odds of success rather than actually depriving any particular individual of an opportunity, see Pauline T. Kim, Essay, *The Colorblind Lottery*, 72 FORDHAM L. REV. 9, 12, 30-35 (2003).

114. *See* Barocas & Selbst, *supra* note 5, at 691.

115. *See id.*

acceptability of relying on those factors. Certain attributes may be sufficiently related to job performance that the law should allow employers to rely on them regardless of their impact. For example, a company might reasonably screen applicants for legal positions to ensure that they are licensed to practice law, even if that selection criterion disadvantages certain groups. Whether employers should rely on other criteria, such as credit scores or criminal record history, is far more debatable, and resolving those questions turns on contested normative judgments.

The nature of data mining complicates our ability to make these types of judgments. Algorithms based on machine learning may be agnostic about what qualities make a good employee, and the resulting model may be opaque as to how it is sorting applicants or employees. Alternatively, the quality or characteristic the model seeks to maximize (the target variable) may be clearly specified, but the algorithm is so complex that it is not possible to explain which factors drive the model's predictions. Even when the factors are identifiable, a pure data mining model will not reveal whether the relationships uncovered are causal or merely coincidental.[116] Thus, in addition to familiar debates about whether certain selection criteria are closely enough related to the job, data analytics raise new questions about whether the law should permit employers to rely on unknown or unexplained correlations when they have the effect of disadvantaging certain groups.

Consider a simple example. Suppose a model analyzing tens of thousands of observations finds that residents of certain zip codes tend to perform more poorly at a particular job. Because residence is often associated with race, the model may effectively screen out minority applicants at higher rates. The data and methods used to build the model may be unimpeachable, such that there are no concerns about statistical bias. Or, put differently, the available evidence might suggest that the correlation is a genuine one. Nevertheless, as a normative matter, relying on this association may be unacceptable, not only because residence does not measure ability, but also because our country has a long history of housing segregation along racial lines.

116. *See* Custers, *supra* note 56, at 16.

A more difficult question is raised if the algorithmic bias results from a factor less clearly identified with past racial harms. Suppose, for example, that an algorithm uncovers a strong statistical correlation between job performance and a seemingly arbitrary factor like what kind of automobile someone drives, but the effect of relying on that factor is to reduce opportunities for members of a minority group. Some models may be so complex that it is impossible to specify which factors influence the results, or what precise weights different factors have in determining the model's predictions. Without knowing the precise mechanism producing the outcome, it is impossible to judge whether it is normatively acceptable to rely on the factors it leverages.

Thus, when an algorithm produces structural disadvantage that is not caused by statistical bias, the nature of the harm is more difficult to characterize. In such a case, the model's disparate outcomes may reflect genuine differences between groups that are relevant to job performance, or it may simply be capturing arbitrary and meaningless correlations. Whether it causes social harm depends on which differences the model leverages to make its predictions and on contested normative judgments about the acceptability of relying on those factors. To the extent that a harm occurs, however, it is a group-based harm. As with discriminatory statistical bias, the disadvantage is structural, and therefore identifying particular individual victims will be difficult.

D. Classification Bias

As discussed in Part I.C, algorithmic decision-making can produce various types of harms for individuals or protected groups deprived of employment opportunities. Apart from the first type—intentional discrimination—these harms do not easily fit traditional notions of discrimination as motivated by prejudice or animus. And yet, the growing use of big data and data analytics in the workplace risks creating or reinforcing patterns of disadvantage and subordination that will be very similar in effect to more familiar forms of discrimination from the past.

These risks raise a concern about what I call "classification bias"—namely, the use of classification schemes that have the effect of exacerbating inequality or disadvantage along lines of race, sex,

or other protected characteristics. I use the term classification bias to emphasize concerns about inequality and disadvantage, and at the same time to underscore that this type of bias results from mechanisms that are quite distinct from familiar forms of discrimination. More specifically, classification bias is *data-driven*, which means that the traditional legal tools for responding to discrimination are in many ways inadequate, as discussed in Parts II and III below.

The term "classification bias" resonates with the data science literature, which identifies "classification" as one of several basic data mining techniques;[117] however, I do not use the phrase in any technical sense. Other data mining techniques that are used to sort and score workers may also systematically disadvantage certain groups. Thus, classification bias applies whenever an algorithm—regardless of its logical structure—systemically biases applicants' or employees' access to opportunities.

In speaking of classification bias, I do not mean to invoke what is sometimes referred to as "anticlassification" theory.[118] Scholars have long debated what principles underlie antidiscrimination law. Some scholars have argued that the guiding principle should be one of formal equality—namely, that the law's protections extend only as far as forbidding employers from making decisions based on an individual's race, sex, or other protected characteristics.[119] This perspective, sometimes referred to as the "anticlassification principle," identifies discriminatory harm primarily in the use of classifications—like race—to make decisions.[120] Anticlassification theory stands in contrast to antisubordination theory, which aims to promote equality by redressing structures and practices that disadvantage historically subordinated groups, regardless of whether the employer expressly or intentionally relied on race or other

117. *See, e.g.*, Toon Calders & Bart Custers, *What Is Data Mining and How Does It Work?*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 56, at 27, 31-34.

118. *See, e.g.*, Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 10 (2003) (describing the anticlassification principle as holding that the government may not classify people on the basis of a forbidden category such as race and explaining that it exists in tension with an antisubordination principle).

119. *See, e.g.*, William Van Alstyne, *Rites of Passage: Race, the Supreme Court, and the Constitution*, 46 U. CHI. L. REV. 775, 797-98, 809-10 (1979).

120. *See* Balkin & Siegel, *supra* note 118, at 10.

categories in making its decisions.[121] Like antisubordination theory, the concept of classification bias proposed here looks at the consequences of employers' decisions. By asking whether neutral classification schemes work to systematically deprive already disadvantaged groups of opportunities, it shares the concerns of antisubordination theorists.

In Part III, I examine to what extent antidiscrimination law can respond to concerns about classification bias. But first, in Part II, I consider two other possible responses and explain why they are likely inadequate to meet the challenges posed by data-driven discrimination.

II. ALTERNATIVE SYSTEMS OF REGULATION

This Part explores whether market forces or privacy law protections can be relied on to eliminate classification bias, and concludes that neither approach is likely to successfully meet concerns about inequality raised by workforce analytics.

A. The Market Response

Proponents of market-based solutions might argue that the growing use of data mining models in employment raises no particular concerns because employers will rely on them only if they are effective. Collecting and analyzing data is expensive and employers will not do so, or pay a third party to do so, unless the benefits exceed the costs. The promised benefit of workforce analytics is that they will save employers time and money when making personnel decisions and will produce a better workforce.[122] Rational employers will not rely on these tools if they do not actually help them hire and retain good employees, and, therefore, market forces should eliminate models that are biased.

Michael Lewis's book, *Moneyball: The Art of Winning an Unfair Game*, has contributed to the idea that data analytics can accurately predict performance. *Moneyball* tells the story of Billy Beane, the

121. *See, e.g.*, *id.* at 9; Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 157-58 (1976); Lawrence, *supra* note 37, at 319-20.

122. *See* Bersin, *supra* note 3.

Oakland Athletics manager who built a competitive baseball team with a limited payroll.[123] By substituting statistical analysis for hunches, intuition, and conventional wisdom, Beane was able to identify undervalued ballplayers and recruit them at a fraction of the cost of their true worth.[124] Since then, statistical analysis has become a standard tool that major league baseball teams use to identify talent.[125] The lesson seemed to be that statistics can not only help identify talent, but that they succeed in doing so because they are more "objective" and can overcome traditional prejudices.[126]

The success of statistics in baseball scouting does not translate easily to more ordinary jobs, however. As Nate Silver points out, baseball is unique in that it "offers perhaps the world's richest data set."[127] Not only are there *lots* of data about almost everything that happens in baseball games, but also the nature of the sport permits the collection of objective measures of individual performance under well-specified conditions—for example, batting statistics in a given ballpark against a particular pitcher.[128] Statistics revolutionized baseball to the extent that it did "because of the sport's unique combination of rapidly developing technology, well-aligned incentives, tough competition, and rich data."[129]

---

123. *See generally* MICHAEL LEWIS, MONEYBALL: THE ART OF WINNING AN UNFAIR GAME (2003).

124. *See id.* at 18, 37-42, 127-29.

125. *See* NATE SILVER, THE SIGNAL AND THE NOISE: WHY SO MANY PREDICTIONS FAIL—BUT SOME DON'T 86-88 (2012).

126. *See id.* at 91-92.

127. *See id.* at 80.

128. *Id.* ("[A]lthough baseball is a team sport, it proceeds in a highly orderly way: pitchers take their turn in the rotation, hitters take their turn in the batting order, and they are largely responsible for their own statistics."). This type of data is harder to come by in other professional sports in which statistics have had less of an impact to date. *See* Leigh Steinberg, *Changing the Game: The Rise of Sports Analytics*, FORBES (Aug. 18, 2015, 3:08 PM), http://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/ [https://perma.cc/WXV6-EDL4] (noting that although use of data analytics in all professional sports has increased, it is harder to adapt analytics to basketball than baseball); Reeves Wiedeman, *The Sabermetrics of Football*, NEW YORKER (Sept. 23, 2011), http://www.newyorker.com/news/sporting-scene/the-sabermetrics-of-football[https://perma.cc/7TRY-F3FE] (discussing why baseball is "more receptive to stats" than football). Even in baseball, statistics have not eliminated the role of scouts, and successful teams today use a combination of quantitative and qualitative information. *See* SILVER, *supra* note 125, at 91-92, 99-101.

129. SILVER, *supra* note 125, at 106.

In the more ordinary workplace, data models are more likely to exhibit bias,[130] and market competition will not reliably eliminate them. First, biased data models may be accurate *enough* to persist in a competitive market, even though they are biased against certain groups. Second, feedback effects may appear to confirm the accuracy of biased data models, entrenching their use. And finally, biased data models may be efficient precisely *because* they are discriminatory, and therefore pressures toward efficiency will not eliminate them.

The first reason that market pressures are unlikely to drive out classification bias is that a model may be sufficiently accurate to benefit employers who use them, even if, at the same time, they have a discriminatory effect. Consider an algorithm that selects candidates who are predicted to be more successful at a particular job. It may be highly effective in identifying strong candidates, even though it disproportionately excludes members of disadvantaged groups. So long as the algorithm is accurate enough to make the employer's process less costly, neither the employer nor the vendor will have sufficient incentive to identify and remove the bias.

This difficulty is compounded when considering singular, high-level positions for which there are few objective measures of performance. In baseball, the availability of highly detailed, objective, and publicly available data about performance means that a team will have numerous observations for comparing the performances of players in nearly identical circumstances.[131] In the case of other highly skilled workers, comparing performance is far more difficult. Finding an objective measure may not be possible, and even if one exists, comparisons will be difficult because a firm cannot observe the performance of the accepted and rejected candidates under identical circumstances. Without this information, it is difficult to assess the benefit or the cost of the choice actually made.

Imagine a company that relies on a data algorithm to choose among several applicants for a management position. The model might be biased in a way that discounts the leadership styles more typical of female candidates, such that it systematically assigns

130. *See supra* Part I.B.

131. Offensive ability in baseball is reliably captured by statistics, but defensive ability has proven somewhat more challenging to measure objectively. *See* LEWIS, *supra* note 123, at 136.

them lower scores, but nevertheless accurately identifies some candidates who are capable of performing the job. The employer may not recognize that the model is biased—particularly if its predictions match the decision maker's prior implicit assumptions or expectations. In other words, the same cognitive biases that data purportedly help to avoid may cause the human decision makers not to notice when the model is biased.

If, relying on such a model, the employer selects a man for the job, and that man is ultimately successful in the position, the employer will have no reason to question the algorithm, even though an unbiased model might have prioritized others, including more female candidates. A female candidate might also have been successful in the job—maybe even *more* successful—but the employer will have no way of knowing that. So long as the algorithm is accurate *enough*, the employer would have no reason to distrust it.

Employers may persist in using biased algorithms to select for low skill positions as well. For these jobs, the basic skills may be widely available in the labor pool, and the relevant performance metrics may be easier to measure and compare across time. For example, an employer concerned with high turnover in low-skilled positions can easily measure the length of job tenure of different employees. The employer may utilize data mining tools in an effort to select employees who will stay longer at the job, and then compare the job tenure of employees hired before and after adopting the model. If the employer observes that employees hired using the model stay on the job longer, it may take that as confirmation of its accuracy. In fact, the model may not have identified the factors that actually increase job tenure. Some other factor, such as a decrease in alternative employment options, may have caused the observed increase in job tenure and would have similarly influenced those applicants not selected to stay on the job longer as well. Alternatively, an unbiased model might have similarly increased tenure without the discriminatory impact. Nevertheless, the employer's observations would not lead it to question the model, and it would likely continue to use it, even though the effect is to disproportionately screen out minority applicants.

A second reason market forces may not reliably squeeze out classification bias is that feedback effects may cause biased models to become more accurate over time—the model in effect becoming a

self-fulfilling prophecy. Suppose, for example, that an employer uses a data model to select employees for an entry-level position for which many applicants meet the minimum qualifications. If the model is biased, such that it overselects individuals in a dominant group, then fewer minority group members will be hired, and the employer will have little opportunity to observe their performance in the position. At the same time, members of the minority group—particularly if similar processes restrict their access to other employment opportunities—may perceive a lower return to effort and therefore lose the incentive to invest in learning relevant skills.[132] A model which erroneously underpredicted minority performance may become more accurate over time. If similar biases operate across multiple domains, affecting access to other critical resources like housing and credit, then these feedback effects will multiply. Thus, when biased selection processes create feedback effects, market forces will tend to affirm rather than disconfirm their usefulness.

Finally, in those cases in which a data model is accurate *because* it is discriminatory, market forces will not eliminate classification bias. As discussed earlier, a model may incorporate biased judgments—for example, ratings by supervisors that are themselves biased—as a measure of job performance. If employers use such a model to predict future cases, and the performances of the selected employees are then evaluated using the same biased measure, the outcomes will simply confirm the "correctness" of the model. In other situations, a model might capture real market differences between employees, but those differences are themselves the product of discriminatory forces. One can imagine, for example, that women are less productive in nontraditional employment settings if they face resistance to their presence that is manifested in harassment and noncooperation from their coworkers. A model that predicts future performance based on the past would both reflect prior discrimination *and* be highly accurate. Once again, an employer focused on efficiency gains is unlikely to abandon the model.

Thus, market forces will not reliably eliminate classification bias. The market may squeeze out highly inaccurate models that fail to provide enough benefit to justify the cost to employers. In many

132. *See* Bagenstos, *supra* note 99, at 464.

cases, however, algorithms are likely to have some predictive value even if they are biased against certain protected groups. If they are accurate enough, employers will not have strong market incentives to abandon them or to incur the costs of searching for less biased alternatives.

B. Privacy Rights

If market forces will not reliably eliminate biased algorithms, then what about regulation aimed at protecting informational privacy? Can restrictions on the collection, disclosure, and use of personal information address the risks of classification bias that data analytics pose? Privacy law scholars argue for more robust rules regulating information flows, suggesting that such rules would not only protect dignitary and autonomy interests, but also address the risk of discrimination as well.[133] Although information rules can certainly mitigate some of the threats to workplace equality, they cannot entirely meet the challenges posed by workplace analytics. A full exploration of the complex relationship between privacy and discrimination is beyond the scope of this Article. Instead, this Section briefly explains why even robust privacy protections are unlikely to fully resolve concerns about data-driven discrimination in the workplace.[134]

In some circumstances, privacy rights can prevent intentional discrimination from occurring. Thus, antidiscrimination statutes sometimes incorporate restrictions on employers' information gathering. For example, the Genetic Information Nondiscrimination Act prohibits employers from inquiring about or otherwise deliberately acquiring genetic information about applicants and employees.[135]

---

133. *See, e.g.*, Richards & King, *supra* note 20, at 409-13.

134. In earlier work, I argued that protecting the privacy of sensitive information could prevent genetic discrimination from occurring. *See* Pauline T. Kim, *Genetic Discrimination, Genetic Privacy: Rethinking Employee Protections for a Brave New Workplace*, 96 NW. U. L. REV. 1497, 1501-02 (2002). That argument turned on the goal of preventing intentional discrimination and the fact that unexpressed genetic characteristics are not identifiable through casual observation. *See id.* at 1517, 1521. My observations about the connections between privacy and discrimination in that context do not necessarily apply in a data-rich environment where the discriminatory outcomes may not be intentional.

135. Genetic Information Nondiscrimination Act of 2008 (GINA) § 202, 42 U.S.C. § 2000ff-1(b) (2012).

The Americans with Disabilities Act similarly limits employers' access to medical information that might reveal the existence of a disability at certain stages of the employment process.[136] This strategy works when the protected characteristic is not readily observable.[137] If the employer does not know about a protected characteristic, such as a disability or a genetic predisposition to disease, it cannot discriminate on that basis. This strategy will obviously be less successful in preventing discrimination on the basis of highly salient characteristics like race and sex. Title VII of the Civil Rights Act does not contain a similar prohibition on acquiring information, although employer inquiries—for example, about an employee's plans to have children—may raise an inference that a later adverse action was taken on a prohibited basis. Restricting access to information can be effective in preventing intentional discrimination when the employer would not otherwise know about the protected characteristic and therefore would be unable to act on that basis.[138]

However, restricting access to sensitive information is not likely to be effective in preventing classification bias that results from data analytic models. If the data being mined is rich enough, other seemingly neutral factors may closely correlate with a protected characteristic, permitting a model to effectively sort along the lines of race or another protected characteristic.[139] Factors such as where someone went to school or where they currently live may be highly correlated with race. Behavioral data, such as an individual's Facebook "likes," can also predict sensitive characteristics like race and sex with a high degree of accuracy.[140] Because other information contained in large datasets can serve as a proxy for race, disability, or other protected statuses, simply eliminating data on those characteristics cannot prevent models that are biased along these dimensions. On the other hand, the problem of omitted variable bias means that prohibiting the collection or use of sensitive data may

136. Americans with Disabilities Act of 1990 (ADA) § 102, 42 U.S.C. § 12112(d)(2)(A).

137. *See* Kim, *supra* note 134, at 1517; *see also* CAL. GOV'T CODE § 12940(d) (West 2016).

138. *Cf.* Kim, *supra* note 134, at 1521.

139. *See* Custers, *supra* note 56, at 9-10.

140. *See, e.g.*, Woodrow Hartzog & Evan Selinger, *Big Data in Small Hands*, 66 STAN. L. REV. ONLINE 81, 83 (2013); Kosinski et al., *supra* note 62, at 5804 fig.4.

sometimes increase the biased effects of a data model.[141] Thus, a simple prohibition on access to sensitive information will not prevent classification bias, and in some cases could make it worse.

Another approach to protecting privacy focuses on procedural protections. Fair information practices emphasize the right of individuals to know when and how personal data is collected, to ensure its accuracy, and to consent to its use.[142] However, these procedural rights have not significantly limited the types of data collected or how employers use that information. Applicants and employees often have little choice but to acquiesce to employer requests for information, and the law grants employers wide discretion in making employment decisions.[143] As a result, the emphasis on consent and data accuracy has had limited practical effect in restricting the information available to employers to make employment decisions.

Experience with the Fair Credit Reporting Act (FCRA), which embodies fair information practice principles, is illustrative.[144] The FCRA requires an employer to obtain an applicant's consent before it accesses a consumer report,[145] to provide notice of an adverse action based on a consumer report along with a copy of the report, and to provide information about the individual's rights to dispute the report's accuracy.[146] These requirements put few obstacles in the path of employers who wish to use consumer data to make personnel decisions. Job applicants have little choice but to consent to the use of credit reports if they wish to be considered for a job. If an

---

141. *See supra* Part I.B.

142. *See, e.g.*, EXEC. OFFICE OF THE PRESIDENT, *supra* note 22, at 17.

143. *See, e.g.*, Pauline T. Kim, *Privacy Rights, Public Policy, and the Employment Relationship*, 57 OHIO ST. L.J. 671, 717 (1996).

144. *See* Fair Credit Reporting Act (FCRA) § 602, 15 U.S.C. § 1681 (2012). For a more detailed discussion of the FCRA's limited ability to address concerns about algorithmic bias, see generally Pauline T. Kim & Erika Hanson, *People Analytics and the Regulation of Information Under the Fair Credit Reporting Act*, 61 ST. LOUIS U. L.J. (forthcoming 2017), https://ssrn.com/abstract=2809910 [https://perma.cc/N35G-P9FR].

145. The FCRA defines a "consumer report" as

> [A]ny written, oral, or other communication of any information by a consumer reporting agency bearing on a consumer's credit worthiness, credit standing, credit capacity, character, general reputation, personal characteristics, or mode of living which is used or expected to be used or collected in whole or in part for the purpose of serving as a factor in establishing the consumer's eligibility for ... employment purposes.

15 U.S.C. § 1681a(d)(1).

146. *Id.* § 1681b(b).

employer denies employment based on the report, the applicant's recourse is to try to correct any errors in that record.[147] The FCRA provides no remedy against an employer for failure to hire even when the employer relied on an inaccurate credit report. Relying on an accurate record to make decisions violates no legal prohibitions either, as long as all of the procedural steps have been followed. Thus, fair information practice principles are unlikely to significantly limit employer use of data models.

Scholars have widely criticized the reliance on notice and consent to protect privacy interests, especially in the era of big data.[148] Lengthy, jargon-filled disclosures encountered in nearly every internet transaction do not provide real notice,[149] and because the alternative to accepting those terms is to refuse the service or transaction, consumers have little real choice about how their personal information will be handled. The processing of big data exacerbates the problem of obtaining meaningful consent. Separate data streams can be combined, and, once aggregated, data may reveal far more about an individual's habits, tastes, and opinions than the individual data points alone would suggest.[150] As the example of Target Stores predicting which consumers were pregnant demonstrates,[151] the disclosure of relatively trivial bits of information may reveal far more sensitive information when data is aggregated and analyzed. Thus, consent obtained at the moment data is collected is not meaningful, given that it is impossible to know all subsequent uses of that information and its impact in advance.[152]

In response to the challenges posed by big data, privacy scholars have proposed forms of regulation that go beyond traditional fair information practice principles. As Neil Richards and Jonathan King point out, privacy rules are not just about secrecy or restricting

147. *See id.* § 1681i(f)(2)(B)(i).

148. *See, e.g.*, Citron & Pasquale, *supra* note 20, at 27-28; Crawford & Schultz, *supra* note 20, at 108; Richards & Hartzog, *supra* note 20 (manuscript at 17-21); Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1880-81 (2013).

149. *See* Richards & Hartzog, *supra* note 20 (manuscript at 18) (citing studies).

150. *See* Solove, *supra* note 148, at 1889-90.

151. *See supra* note 105 and accompanying text.

152. Solove, *supra* note 148, at 1889-90.

access to personal information.[153] Rather, privacy should be understood as "the rules that govern how information flows."[154] For example, Kate Crawford and Jason Schultz advocate for a form of procedural data due process entitling individuals to know when predictive analytics are used and to challenge the fairness of the process.[155] Danielle Citron and Frank Pasquale similarly assert that data subjects should have the right to correct inaccurate data and that regulatory oversight should ensure the fairness of scoring systems.[156]

Requiring data transparency, auditing for accuracy, and substantively regulating downstream uses of data are important steps in ensuring the fair use of data; however, these types of interventions cannot fully address the risk of classification bias in employment. Inaccuracies in an individual's record may unfairly deprive her of a particular opportunity, but accurate records do not guarantee unbiased outcomes. If an individual is excluded because of errors in her individual record, procedural rights can help correct the errors. However, fixing errors in an individual's record will not prevent statistical bias or structural disadvantage—harms which result from the overall operation, rather than any individual application, of an algorithm. Because these harms operate by reducing opportunities for members of a group as a whole, merely correcting individual errors will not eliminate them. Thus, even robust privacy law regimes that focus on data accuracy are likely insufficient to address concerns about classification bias in employment.

III. THE ANTIDISCRIMINATION RESPONSE

If neither the market nor privacy protections can reliably prevent classification bias, what about antidiscrimination law? In the employment context, Title VII of the Civil Rights Act of 1964 was the landmark piece of legislation establishing the antidiscrimination norm by forbidding discrimination on the basis of race, color, religion, sex, and national origin.[157] Later federal enactments extended

153. *See* Richards & King, *supra* note 20, at 411-12.
154. *Id.* at 411.
155. *See* Crawford & Schultz, *supra* note 20, at 126-27.
156. *See* Citron & Pasquale, *supra* note 20, at 20-22.
157. *See* Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a) (2012).

protections to older workers[158] and individuals with disabilities,[159] and prohibited discrimination based on genetic traits.[160] How do these laws apply to bias that is data-driven? Barocas and Selbst examined this question and concluded that "Title VII would appear to bless" the use of algorithms, even when they work to disadvantage protected groups.[161] In this Part, I reject that conclusion, arguing instead that employment discrimination law can provide a vehicle for addressing classification bias, so long as the doctrine accounts for its data-driven sources. The discussion below focuses on Title VII, because both statutory text and judicial interpretation of other employment discrimination laws often follows that of Title VII.[162]

In Section A, I review the conventional understanding of Title VII which divides prohibited discrimination into two categories—disparate treatment and disparate impact—and explain its limitations in addressing classification bias. Section B argues that a close reading of the statutory text supports a finding that Title VII directly prohibits classification bias. In Section C, I consider what an effective legal response to classification bias might look like, and how it should differ from conventional disparate impact theory in order to more closely meet the unique challenges that biased algorithms pose. The last two Sections of this Part, D and E, consider whether there are any legal or practical limits to relying on antidiscrimination law to address classification bias.

A. The Conventional Account of Title VII

Judges, litigants, and scholars commonly recite that Title VII prohibits two types of discrimination: disparate treatment and

158. *See* Age Discrimination in Employment Act of 1967 §§ 2-12, 14-15, 17, 29 U.S.C. §§ 621-634 (2012).

159. *See* Americans with Disabilities Act of 1990 §§ 2-4, 101-102, 42 U.S.C. §§ 12101-12112.

160. *See* Genetic Information Nondiscrimination Act of 2008 §§ 201-212, 42 U.S.C. §§ 2000ff to 2000ff-11.

161. *See* Barocas & Selbst, *supra* note 5, at 672.

162. Of course, there are differences between Title VII and the other antidiscrimination statutes, which might affect the analysis, but a close examination of Title VII is a reasonable starting point. Further work should explore the extent to which the arguments advanced here do or do not apply to prohibitions on discrimination based on age, disability, or genetic traits.

disparate impact.[163] The standard account holds that disparate treatment cases involve intentional discrimination based on a protected characteristic, whereas disparate impact cases target employer practices that are facially neutral but have discriminatory effects.[164] As many scholars have argued, this neat division of actionable discrimination into two discrete types oversimplifies the reality of how bias can operate in the workplace.[165] It also arguably oversimplifies the relationship between these types of discrimination as a doctrinal matter.[166] And, as I argue in Section B of this Part, it may not be the best reading of the statutory text, or even an entirely accurate explanation of current doctrine.

Nevertheless, the conventional understanding is the place to begin. Read narrowly, existing Title VII doctrine does not appear to match the particular risks to workplace equality that classification bias poses.[167] Only one of the types of harm identified in Part I.C.—intentional discrimination—easily fits within the conventional framework. When an employer intends to discriminate but relies on an apparently neutral data model to justify its decisions, the traditional disparate treatment doctrine clearly applies.[168] The

163. *See, e.g.*, Furnco Constr. Corp. v. Waters, 438 U.S. 567, 569 (1978); Charles A. Sullivan, *Disparate Impact: Looking Past the* Desert Palace *Mirage*, 47 WM. & MARY L. REV. 911, 914 (2005) ("Early in its history, the Supreme Court adopted two definitions of the term ['discriminate':] ... disparate impact ... [and] disparate treatment.").

164. *See* Ricci v. DeStefano, 557 U.S. 557, 577-78 (2009); Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 986-87 (1988).

165. *See, e.g.*, Green, *supra* note 48, at 92; Krieger, *supra* note 39, at 1164-65; David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899, 899 (1993); Sturm, *supra* note 48, at 461.

166. *See, e.g.*, Jed Rubenfeld, Essay, *Affirmative Action*, 107 YALE L.J. 427, 436-37 (1997); George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313, 2313 (2006); Stacy E. Seicshnaydre, *Is the Road to Disparate Impact Paved with Good Intentions?: Stuck on State of Mind in Antidiscrimination Law*, 42 WAKE FOREST L. REV. 1141, 1142 (2007).

167. Barocas and Selbst similarly concluded that Title VII is "not well equipped" to address the various discriminatory features of data mining. *See* Barocas & Selbst, *supra* note 5, at 694.

168. Under the familiar *McDonnell Douglas* burden-shifting framework, the plaintiff has the initial burden of establishing a prima facie case of discrimination. *See* McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801-03 (1973). The employer must then "articulate some legitimate, nondiscriminatory reason" for the adverse employment action. *Id.* Finally, the plaintiff has the opportunity to show that the employer's proffered justification is pretext for discrimination. *Id.* at 804. If an employer were to point to the predictions of a data model to justify an adverse decision, the plaintiff could try to prove that the model is merely a pretext for intentional discrimination.

plaintiff may find it quite difficult as a practical matter to prove the employer's discriminatory intent in using a biased data model;[169] however, this scenario poses no conceptual difficulties under the disparate treatment theory.

As discussed in Part I.B, simply prohibiting use of protected characteristics will not prevent classification bias. Other nonsensitive variables can act as proxies, such that a model that does not explicitly consider race or sex may nevertheless have discriminatory effects along those lines. Moreover, because of the problem of omitted variable bias, forbidding the use of protected class variables could exacerbate discriminatory effects under certain circumstances. Thus, a blanket prohibition on the explicit use of race or other prohibited characteristics does not avoid, and may even worsen, the discriminatory impact of relying on a data model.[170]

The other types of harm resulting from classification bias—due to individual record errors, statistical bias, and structural disadvantage—can occur without any conscious intent or awareness on the part of the employer. Disparate impact doctrine would thus seem the natural place to look for a response. First articulated by the Supreme Court in *Griggs v. Duke Power Co.*, the disparate impact theory holds that Title VII forbids not only overt discrimination, but also "practices that are fair in form, but discriminatory in operation."[171] The *Griggs* Court held that Duke Power could not require applicants to have a high school diploma or a passing score on a written test unless those requirements had "a demonstrable relationship to successful performance."[172]

The disparate impact theory recognized in *Griggs* was rooted in Title VII's purpose—"to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees."[173] Given that purpose, the Court held that Title VII required "the

169. *See* Barocas & Selbst, *supra* note 5, at 712-14.

170. In fact, mitigating the risk of biased outcomes arguably requires *preserving* data on race, sex, and other protected characteristics. *See infra* Part III.C.

171. 401 U.S. 424, 431 (1971). The *Griggs* Court explained that "artificial, arbitrary, and unnecessary barriers to employment" that "operate invidiously to discriminate on the basis of racial or other impermissible classification" are forbidden unless they "bear a demonstrable relationship to successful performance" of the job. *Id.*

172. *Id.*

173. *Id.* at 429-30.

removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate on the basis of racial or other impermissible classification."[174] The lack of discriminatory intent did not absolve the employer, for it "does not redeem employment procedures or testing mechanisms that operate as 'built-in headwinds' for minority groups and are unrelated" to the worker's ability to do the job.[175]

As described in *Griggs*, the disparate impact theory would appear well-suited to address classification bias. Reliance on algorithms will typically be a facially neutral employment practice. Data models that do not explicitly categorize on the basis of race or other protected categories may nevertheless operate as "built-in headwinds" for disadvantaged groups. However, since the Court first articulated the concept of disparate impact in *Griggs*, a doctrinal superstructure has developed around the theory, which does not fit well when bias is data driven.[176]

As refined in subsequent cases and eventually codified by the Civil Rights Act of 1991, disparate impact liability attaches when a plaintiff has shown that an employment practice produces a disparate impact on the basis of a protected characteristic and the employer "fails to demonstrate that the challenged practice is job

174. *Id.* at 431.

175. *Id.* at 432.

176. Numerous scholars have noted the limitations of the doctrine and its failure to meet initial expectations of its transformative potential. Civil rights advocates initially heralded the *Griggs* decision as monumentally important in advancing the cause of workplace equality. *See, e.g.*, Robert Belton, *Title VII at Forty: A Brief Look at the Birth, Death, and Resurrection of the Disparate Impact Theory of Discrimination*, 22 HOFSTRA LAB. & EMP. L.J. 431, 433 (2005) ("Aside from *Brown v. Board of Education*, the single most influential civil rights case during the past forty years that has profoundly shaped, and continues to shape, civil rights jurisprudence and the discourse on equality is *Griggs v. Duke Power Co.*"); Alfred W. Blumrosen, *The Legacy of* Griggs*: Social Progress and Subjective Judgments*, 63 CHI.-KENT L. REV. 1, 1-2 (1987) ("Few decisions in our time—perhaps only *Brown v. Board of Education*—have had such momentous social consequences [as *Griggs*]." (footnote omitted)). However, many others have viewed the doctrine more skeptically, arguing that it has been narrowly applied, is inherently limited, and lacks a clear theoretical basis. *See, e.g.*, Rutherglen, *supra* note 166, at 2314; Selmi, *supra* note 28, at 706 ("[D]isparate impact claims are more difficult—not easier—to prove than claims of intentional discrimination."); Sullivan, *supra* note 163, at 970, 975-76; Amy L. Wax, *Disparate Impact Realism*, 53 WM. & MARY L. REV. 621, 626 (2011); Steven L. Willborn, *The Disparate Impact Model of Discrimination: Theory and Limits*, 34 AM. U. L. REV. 799, 804 (1985); Nicole J. DeSario, Note, *Reconceptualizing Meritocracy: The Decline of Disparate Impact Discrimination Law*, 38 HARV. C.R.-C.L. L. REV. 479, 484, 507 (2003).

related for the position in question and consistent with business necessity."[177] Even if the employer satisfies this burden, complainants might still prevail by demonstrating the existence of a less discriminatory alternative.[178] More specifically, a complaining party could "show that other tests or selection devices, without a similarly undesirable ... effect [on the protected class], would also serve the employer's legitimate interest."[179]

Michael Selmi argues that the disparate impact doctrine is not well suited to application outside the contexts in which the doctrine developed.[180] He points out that the early cases focused on seniority systems and written tests that employers used to perpetuate discrimination that had been lawful prior to the passage of Title VII.[181] Contemporaneous commentators understood the significance of the *Griggs* case as defining what was required to validate written employment tests.[182] The next disparate impact case decided by the Supreme Court, *Albemarle Paper Co. v. Moody*, also involved a challenge to preemployment tests, as well as an employer's seniority system.[183] According to Selmi, application of disparate impact doctrine to these practices was relatively straightforward because they involved "specific practices that were easy to identify and for which there was no difficult causal question" and "[t]he employers'

177. *See* Civil Rights Act of 1991 § 105, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

178. *See* 42 U.S.C. § 2000e-2(K)(1)(A)(ii).

179. *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975). The exact standard for establishing liability based on the existence of an alternative employment practice is uncertain because rather than defining the standard, Congress in the Civil Rights Act of 1991 simply referred to "the law as it existed on June 4, 1989, with respect to the concept of 'alternative employment practice.'" *See* 42 U.S.C. § 2000e-2(k)(1)(c). In effect, Congress restored the law as it existed before the Supreme Court's decision in *Wards Cove Packing Co. v. Atonio*, decided on June 5, 1989. 490 U.S. 642 (1989), *superseded by statute*, Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1074, *as recognized in* Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507 (2015). In doing so, Congress repudiated the Court's suggestion in *Wards Cove* that "any alternative practices ... must be equally effective as [the employer's] chosen hiring procedures in achieving [its] legitimate employment goals," including factors such as cost and other burdens on the employer. *See id.* at 661. However, because there was disagreement prior to *Wards Cove* about what exactly was required to show the existence of an alternative employment practice, the Civil Rights Act of 1991 did not resolve the issue.

180. *See* Selmi, *supra* note 28, at 705.

181. *See id.* at 708-16.

182. *See id.* at 723.

183. *See* 422 U.S. at 408-09.

rationales were likewise relatively easy to define."[184] When applied in other contexts lacking these characteristics, however, the doctrine does not fit well, and liability is far more difficult to prove. As a result, very few disparate impact cases have been successful outside of the specific contexts in which the doctrine developed.[185]

Similarly, traditional disparate impact doctrine is a poor fit for addressing classification bias. Most data models have none of the characteristics that Selmi identifies as making disparate impact doctrine workable. Rather than providing specific selection criteria that are justified by clearly stated rationales, data models typically involve opaque decision processes, rest on unexplained correlations, and lack clearly articulated employer justifications.

The written employment tests targeted in early disparate impact litigation were grounded in psychological theories regarding aptitude and ability.[186] These tests focused on identifying and measuring skills or personal characteristics relevant to successful performance of a job, and their validity could be evaluated in light of standards set by an established scientific discipline.[187] In contrast, data mining is entirely atheoretical.[188] The models exploit whatever data are available, rather than selecting which factors should be included or controlled for based on theoretical expectations.[189] As a result, if existing disparate impact doctrine is applied

184. Selmi, *supra* note 28, at 716. Once adopted, disparate impact doctrine came to be seen as a generalized method of proving discrimination in situations far removed from seniority systems and written tests. *See, e.g.*, N.Y.C. Transit Auth. v. Beazer, 440 U.S. 568, 584-85 (1979) (applying disparate impact doctrine to claim that a transit authority's regulation prohibiting the use of narcotics by employees violated Title VII); Dothard v. Rawlinson, 433 U.S. 321, 328-29 (1977) (applying disparate impact doctrine to claim that height and weight requirements for employment discriminated against women).

185. *See* Selmi, *supra* note 28, at 739-43 (describing how intentional discrimination cases may be easier to prove, with many cases asserting claims under both disparate impact and disparate treatment doctrine, and succeeding on the disparate treatment claim but not on the disparate impact claim); *id.* at 753 ("[O]utside of the testing cases, there has been no area where the disparate impact theory has proved transformative or even particularly successful.").

186. *See, e.g.*, Ablemarle Paper Co. v. Moody, 422 U.S. 405, 410-13 (1975) (challenging employer use of Revised Beta Examination and Wonderlic Personnel Testing); Griggs v. Duke Power Co., 401 U.S. 424, 428-29 (1971) (challenging employer use of Wonderlic Personnel Test and Bennett Mechanical Comprehension Test).

187. *See* Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.15 (2016).

188. *See supra* note 92 and accompanying text.

189. *See supra* Part I.B.

mechanically, it will fail to address the mechanics underlying classification bias.

A couple of examples are illustrative. Under disparate impact doctrine, if a plaintiff shows that an employer practice has a disproportionate impact on a protected group, the employer may defend by showing that the practice is "job related ... and consistent with business necessity."[190] If an employer could meet this burden simply by showing that an algorithm rests on a statistical correlation with some aspect of job performance, then the test is entirely tautological, because, by definition, data mining is about uncovering statistical correlations. Any reasonably constructed model will satisfy the test, and the law would provide no effective check on data-driven forms of bias. Similarly, in disparate impact cases courts tend to defer to employer judgments about what abilities or skills are necessary for a job when evaluating employer justifications for a practice.[191] However, data mining models often rely on "discovered" relationships between variables rather than measuring previously identified job-related skills or attributes. When the employer has not considered and clearly articulated the reasons for relying on particular criteria, it is unclear why any deference is warranted.

The differences between employment testing and data mining also mean that defenses based on section 703(h) of Title VII do not apply. That section excuses employers from liability for relying on "any professionally developed ability test" so long as the test is "not designed, intended or used to discriminate" on a protected basis.[192] Algorithms used to sort or score workers are not "ability tests" because they do not actually test ability—rather, they identify behavioral markers that appear to correlate with on-the-job success. The legislative history of section 703(h) indicates that Congress added it to the statute to immunize the practice—common at the time—of relying on standardized tests to select applicants for hire or promotion.[193] Reflecting this understanding, the Equal Employment

190. Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

191. *See* Selmi, *supra* note 28, at 753; Wax, *supra* note 176, at 633-34.

192. 42 U.S.C. § 2000e-2(h).

193. The version ultimately adopted made clear that reliance on these types of tests was not permitted if "designed, intended or used to discriminate." *Id.* The opinion in *Griggs* focused primarily on this language, adopting the Equal Employment Opportunity Commission's

Opportunity Commission (EEOC) Uniform Guidelines on Employee Selection Procedures, which interpret section 703(h), rely on and incorporate standards regarding test validation established by the American Psychological Association.[194] Because the EEOC wrote them to address an entirely different practice, those Guidelines are simply irrelevant when evaluating the use of atheoretical data mining models that result in classification bias.

To be clear, the *concept* of disparate impact—the idea that facially neutral employer practices can have discriminatory effects—applies to classification bias. The problem is that the ways the doctrine has been applied in the past are not well suited to address the data-driven nature of classification bias. Disparate impact theory *can* meet these specific challenges; however, doing so will require some adjustments in how it applies to workforce analytics. Section C below explains what types of adjustments are required, but first I consider whether Title VII can be read to address classification bias directly.

B. A Closer Reading

The conventional reading of Title VII assumes that disparate treatment and disparate impact exhaust the possibilities for proving a violation under the statute. Scholars concerned about implicit biases or workplace structures that disadvantage women or racial minorities have either argued that the disparate treatment or disparate impact theory ought to apply,[195] or expressed concern that neither theory fits.[196] Similarly, Barocas and Selbst's conclusion that Title VII "would appear to bless" the use of data models even when they produce discriminatory results[197] rests on the assumption that the only available alternatives are existing disparate treatment and disparate impact doctrines.

---

interpretation that section 703(h) requires that any test be "job related" and not merely professionally prepared. *See* Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971).

194. *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 430-31 (1975).

195. *See, e.g.*, Green, *supra* note 48, at 145; Krieger, *supra* note 39, at 1231.

196. *See, e.g.*, Samuel R. Bagenstos, *Bottlenecks and Antidiscrimination Theory*, 93 TEX. L. REV. 415, 434-35 (2014) (reviewing JOSEPH FISHKIN, BOTTLENECKS: A NEW THEORY OF EQUAL OPPORTUNITY (2014)); Sullivan, *supra* note 163, at 1000.

197. *See* Barocas & Selbst, *supra* note 5, at 672.

Perhaps, however, these two doctrines do not exhaust the options for demonstrating the discrimination forbidden by Title VII. The operative language of section 703 is divided into two parts:

> (a) 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; or
>
> (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.[198]

The conventional reading of section 703 is that (a)(1) is about disparate treatment—which turns on motive[199]—whereas (a)(2) is about disparate impact—which focuses on discriminatory effects. This reading reflects the doctrinal superstructure that has devel-

198. Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(1)-(2) (2012). The Age Discrimination in Employment Act and the Genetic Information Nondiscrimination Act contain nearly identical prohibitions. *See* Age Discrimination in Employment Act of 1967 § 4, 29 U.S.C. § 623(a)(1)-(2) (2012); Genetic Information Nondiscrimination Act of 2008 § 202, 42 U.S.C. § 2000ff-1(a)(1)-(2). The Americans with Disabilities Act (ADA) similarly forbids "limiting, segregating, or classifying a job applicant or employee in a way that adversely affects the opportunities or status of such applicant or employee because of ... disability." *See* American with Disabilities Act of 1990 § 102, 42 U.S.C. § 12112(b)(1). However, the operative provisions of the ADA differ from Title VII in other significant ways—for example, by making unlawful an employer's failure to reasonably accommodate otherwise qualified individuals with a disability, and its use of "qualification standards, employment tests or other selection criteria that screen out or tend to screen out" individuals with disabilities. *See* 42 U.S.C. § 12112(b)(5)-(6). These differences may mean that the ADA applies to biased data models in different ways than Title VII—a discussion that is beyond the scope of this Article.

199. Although the Supreme Court has at times suggested that disparate treatment cases require proof of discriminatory motive, *see, e.g.*, Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 335 n.15 (1977) (stating that "[p]roof of discriminatory motive is critical" in disparate treatment cases), subsection 703(a)(1) does not refer to "intent" or "motive" at all. Rather, interpretation of that provision hinges entirely on the words "because of." As Noah Zatz argues, however, "because of" could be interpreted to mean many things other than "motivated by." *See* Noah D. Zatz, *The Many Meanings of "Because Of": A Comment on* Inclusive Communities Project, 68 STAN. L. REV. ONLINE 68, 68-69 (2015).

oped around Title VII rather than a coherent underlying theory of discrimination. As numerous scholars have pointed out, the distinction between disparate treatment and disparate impact is far from clear, and the two theories overlap quite a bit both conceptually and as a matter of proof.[200] Nevertheless, the notion that disparate treatment and disparate impact capture the entire meaning of subsections 703(a)(1) and (a)(2), respectively, is often an unquestioned assumption.

However, the conventional reading does not inevitably flow from the statutory language. Focusing on the text suggests that Title VII also forbids what I have called classification bias—namely, the use of classification schemes that have the effect of exacerbating inequality or disadvantage along lines of race, sex, or other protected characteristics. The language of section 703(a)(2) specifically refers to employer practices that "classify" employees in ways that "deprive or tend to deprive" individuals of employment opportunities because of protected characteristics.[201] Obviously, Congress did not have in mind the problem of biased data mining models when it enacted the language of section 703(a)(2) in 1964. Nevertheless, the language sweeps broadly enough to reach unanticipated employer practices that exacerbate or entrench inequality on prohibited bases.

Differences in the texts of subsections 703(a)(1) and (a)(2) support the conclusion that section 703(a)(2) has broader reach than section 703(a)(1). Section 703(a)(2) restricts an employer's ability to "limit, segregate, or classify" its employees or applicants.[202] In contrast to section 703(a)(1), which focuses on actions, such as hiring, firing, setting compensation, or terms and conditions that are taken with respect to a particular employee, section 703(a)(2) focuses on group-based actions—limiting, segregating, or classifying—all actions that necessarily are taken along some generalizable dimension. Importantly, the prohibited actions are not defined as limiting, segregating, or classifying *on the basis* of race or other protected characteristics. Instead, the emphasis of the language is on actions (such as classifying) that "deprive or tend to deprive" employees of opportunities on a protected basis.

---

200. *See, e.g.*, Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1343-44 (2010); Rutherglen, *supra* note 166, at 2322-23, 2325, 2327, 2329-30.

201. 42 U.S.C. § 2000e-2(a)(2).

202. *See id.*

Many courts and commentators have simply assumed that section 703(a)(2) is synonymous with disparate impact doctrine. However, the text of (a)(2) makes no mention of "disparate impact," "discriminatory effects," "business necessity," or "job relatedness."[203] These concepts are codified in section 703(k), leaving open the possibility that section 703(a)(2) has meaning beyond or apart from established disparate impact doctrine.

When the Supreme Court first articulated the disparate impact theory in *Griggs*, it was only loosely connected to the language of section 703(a)(2).[204] In framing the question presented—whether the Duke Power Company's high school diploma and testing requirements were lawful under Title VII—the Court dropped a footnote citing to the language of section 703(a)(2).[205] The Court made no further mention of that particular statutory provision in the opinion. Instead, the Court rested its analysis on Congress's objective in enacting Title VII—namely, "to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees."[206] The only part of the text of Title VII that the Court engaged with at length was section 703(h), which permits employers to rely on professionally developed ability tests so long as they are not "designed, intended or used to discriminate."[207]

Subsequent cases cited primarily to *Griggs* as authority for the disparate impact doctrine,[208] although the Court eventually explained that *Griggs* was grounded in the text of section 703(a)(2).[209]

---

203. *See id.*

204. *See* Griggs v. Duke Power Co., 401 U.S. 424, 426 (1971).

205. *Id.* at 426 n.1.

206. *Id.* at 429-30.

207. *See id.* at 433 (emphasis removed) (quoting Civil Rights Act of 1964, Pub. L. No. 88-352, § 703, 78 Stat. 241, 257 (1964) (codified as amended in scattered sections of 42 U.S.C.)). Duke Power Company argued that section 703(h) authorized its use of general intelligence tests as a screening device. *See id.* Relying on guidance the EEOC had issued and the legislative history of section 703(h), the Court concluded that the employer could not rely on the provision to defend its testing requirement when the test was not job related. *Id.* at 433-36.

208. The next three disparate impact cases in the Supreme Court did not cite to section 703(a)(2) at all in the majority opinions. *See generally* N.Y.C. Transit Auth. v. Beazer, 440 U.S. 568 (1979); Dothard v. Rawlinson, 433 U.S. 321 (1977); Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975).

209. *See* Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 985-86 (1988); *see also* Smith v. City of Jackson, 544 U.S. 228, 235 (2005) (explaining that although *Griggs* "relied primarily on the purposes of the Act," the Court subsequently found that the disparate impact theory

Some commentators questioned whether Title VII authorized disparate impact claims at all,[210] but those concerns became moot when Congress enacted the Civil Rights Act of 1991.[211] Congress passed that legislation in response to several Supreme Court decisions in the late 1980s that were widely criticized as interpreting the protections of Title VII too narrowly.[212]

One of those cases was *Wards Cove Packing Co. v. Atonio*, a disparate impact case involving two companies that operated salmon canneries in remote areas of Alaska.[213] The plaintiffs alleged that the employers' hiring and promotion practices had produced a racially stratified workforce, in which skilled jobs (noncannery jobs) were held predominantly by white workers, while unskilled jobs (cannery jobs) were held predominantly by nonwhites.[214] The Court of Appeals found a prima facie case of disparate impact, but the Supreme Court reversed, holding that the appeals court had relied on the wrong statistics to conclude that a disparate impact existed.[215]

"represented the better reading of the statutory text as well"); Connecticut v. Teal, 457 U.S. 440, 445-47 (1982).

210. *See, e.g.*, Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2526 (2015) (Thomas, J., dissenting) ("[T]he foundation on which the Court builds its latest disparate-impact regime—*Griggs v. Duke Power Co.*—is made of sand." (citation omitted)); Nelson Lund, *The Law of Affirmative Action in and After the Civil Rights Act of 1991: Congress Invites Judicial Reform*, 6 GEO. MASON L. REV. 87, 94 (1997) (arguing that there was no basis for the Supreme Court's recognition of the disparate impact theory in *Griggs*); *see also* Selmi, *supra* note 28, at 708-24 (detailing the origins of the disparate impact cause of action).

211. *See* Pub. L. No. 102-166, 105 Stat. 1071 (1991) (codified as amended in scattered sections of 42 U.S.C.).

212. *See* Sullivan, *supra* note 163, at 961 ("In reaction to *Wards Cove* and other decisions issued during the 1988 Term of the Supreme Court, Congress passed, and President Bush signed, the Civil Rights Act of 1991.").

213. *See* 490 U.S. 642 (1989), *superseded by statute*, Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071, *as recognized in* Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507 (2015).

214. *See id.* at 647-48.

215. *See id.* at 655. The Court held that the Ninth Circuit Court of Appeals had erred by comparing the percentage of nonwhite workers in the cannery and noncannery positions, and concluding that the stark racial disparity between the two groups established a prima facie case of disparate impact discrimination. *Id.* The relevant statistical comparison, the Court explained, is "between the racial composition of [the at-issue jobs] and the racial composition of the qualified ... population in the relevant labor market." *Id.* at 650 (quoting Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 308 (1977) (alterations in original)). Because the cannery work force did not reflect the population of qualified workers for the noncannery jobs, the statistical disparity in racial composition between the two groups did not establish a

In remanding, the Court addressed several additional issues—arguably all dicta—regarding disparate impact litigation. First, it stated that plaintiffs must identify the specific employment practice that created the alleged disparate impact as part of the prima facie case.[216] Second, it lowered the burden placed on the employer to justify an employment practice—asking whether it "serves, in a significant way, the legitimate employment goals of the employer"[217] rather than whether it is job related or required by business necessity, as it had in earlier cases.[218] Finally, the Court reallocated the burden of proving the lack of a business necessity to the plaintiffs,[219] which made it more difficult for plaintiffs to establish liability by showing that a less discriminatory alternative existed that the employer failed to adopt.[220]

When Congress passed the Civil Rights Act of 1991, it responded to the Court's decision in *Wards Cove* by codifying the disparate impact doctrine and overturning or rejecting some of the Court's guidance on disparate impact cases. It did so by placing the burden on the employer to demonstrate that a challenged practice is "job related for the position in question and consistent with business necessity,"[221] and by making clear that if "the elements of a respondent's decision-making process are not capable of separation for analysis, the decision-making process may be analyzed as one

---

disparate impact. *See id.* at 651.

216. *Id.* at 656-58.

217. *Id.* at 659.

218. *See, e.g.*, Dothard v. Rawlinson, 433 U.S. 321, 331 n.14 (1977) (finding that "a discriminatory employment practice must be shown to be necessary to safe and efficient job performance"); Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971) (stating that in disparate impact cases, "[t]he touchstone is business necessity").

219. *See Wards Cove*, 490 U.S. at 659. After a prima facie case of disparate impact is established, "the employer carries the burden of producing evidence of a business justification for his employment practice," but the ultimate burden of persuasion remains with the plaintiff. *Id.*

220. *See id.* at 661. The Court wrote that

> [A]ny alternative practices which respondents offer up ... must be equally effective as [the employer's] chosen hiring procedures in achieving [its] legitimate employment goals. Moreover, "[f]actors such as the cost or other burdens of proposed alternative selection devices are relevant in determining whether they would be equally as effective as the challenged practice in serving the employer's legitimate business goals."

Id. (quoting Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 998 (1988) (fifth alteration in original)).

221. Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

employment practice."[222] With regard to establishing liability by showing the existence of an "alternative employment practice," the Act simply stated that such a showing "shall be in accordance with the law as it existed on June 4, 1989"—the day before the Supreme Court issued the *Wards Cove* decision—without trying to articulate the correct standard.[223]

Congress made these changes by adding a new subsection (k), which defined disparate impact liability, to section 703 of Title VII, and retaining the language of section 703(a)(2) intact. After the amendments, the statute continued to prohibit in section 703(a)(2) limiting, segregating, or classifying employees in ways that "deprive or tend to deprive any individual of employment opportunities" because of race, color, religion, sex, or national origin, separately from the prohibition in section 703(k) of employment practices that have a disparate impact. Thus, the Civil Rights Act of 1991 left open the possibility that the judicially elaborated theory of disparate impact, as codified in section 703(k), does not exhaust the meaning of section 703(a)(2).

Interestingly, dictum in the Court's *Wards Cove* decision is consistent with a reading that gives section 703(a)(2) meaning apart from traditional disparate impact doctrine. Because the canneries operated on a seasonal basis in a remote location, the employers provided housing and meals. Cannery and noncannery workers were assigned to separate dormitories and mess halls, which resulted in racially stratified living and eating quarters. In passing, the Supreme Court commented that the racially segregated facilities could give rise to a separate claim under section 703(a)(2), apart from the plaintiffs' claim of disparate impact in hiring and promotion.[224] The Court's language is admittedly ambiguous, but one way

222. *Id.* § 2000e-2(k)(1)(B)(i).

223. *Id.* § 2000e-2(k)(1)(C).

224. More specifically, the Court clarified the reach of its opinion in a footnote:

> The Court of Appeals did not purport to hold that any specified employment practice produced its own disparate impact that was actionable under Title VII. This is not to say that a specific practice, such as nepotism, if it were proved to exist, could not itself be subject to challenge if it had a disparate impact on minorities. *Nor is it to say that segregated dormitories and eating facilities in the workplace may not be challenged under 42 U.S.C. § 2000e-2(a)(2) without showing a disparate impact on hiring or promotion.*

Wards Cove, 490 U.S. at 655 n.9 (emphasis added).

In other words, even if no actionable disparate impact had produced the employer's racially

of reading it is that section 703(a)(2)'s meaning is not cabined by the disparate impact doctrine.

In any case, the fact that Congress left section 703(a)(2) intact when it responded to *Wards Cove* in the Civil Rights Act of 1991 supports the idea that (a)(2) continues to have independent force apart from the traditional disparate impact theory codified in subsection (k). Without the doctrinal elaboration of disparate impact theory, the text of (a)(2) supports a finding that Title VII prohibits classification bias.

C. Addressing Classification Bias

As discussed in Section III.B, Title VII could be read to directly prohibit classification bias when algorithms operate to systematically disadvantage protected groups. Alternatively, disparate impact doctrine might be adjusted in ways that address those concerns. In either case, an effective legal response will require developing the doctrine to meet the particular challenges posed by data-driven discrimination. This Section sketches what a legal prohibition of classification bias looks like and how it should differ from traditional disparate impact doctrine.

As a preliminary note, this exploration focuses on employer liability, leaving aside the question whether vendors who create these models and sell or license them to employers should bear any legal responsibility. Although Title VII does apply to employment agencies,[225] it is highly uncertain whether that provision reaches vendors. I do not attempt to answer that question here, focusing instead on how Title VII might be applied to employers to address classification bias caused by workplace analytics. Regardless of whether vendors are directly liable, employers who face potential legal responsibility will have an incentive to pressure vendors to avoid biased outcomes.

---

stratified workforce, the plaintiffs might still be able to use section 703(a)(2) to challenge the employer's use of a classification (cannery versus noncannery workers) that adversely affected the employees' status. In that case, the harm suffered by the workers was the segregated living and dining quarters, and the violation occurred because the employer relied on a neutral classification that had the effect of depriving individual workers of opportunities or status because of their race. *See* 42 U.S.C. § 2000e-2(a)(2).

225. *See* 42 U.S.C. § 2000e-2(b).

Prohibiting classification bias requires examining the actual impact of the algorithms used to sort applicants and employees, and asking whether they deprive individuals of employment opportunities along lines of race, sex, or other protected characteristics.[226] Like traditional disparate impact doctrine, classification bias focuses on facially neutral employment practices that have disproportionately adverse effects on disadvantaged groups.[227] And like disparate impact doctrine, classification bias is not concerned with employer intent or motive.[228] If an employer relies on a data-driven classification scheme to sort applicants or employees, then it should be responsible for the impact that selection device has on the opportunities of workers in protected classes.

Given the differing reasons that data analytics may produce biased outcomes, an effective legal response must differ from traditional disparate impact doctrine in a number of ways. First, the law should not require employers to purge sensitive information, such as race and sex, from datasets; instead, preserving such data is important to avoid bias. Second, the method of identifying the relevant labor market for statistical comparison should look quite different. Third, an employer's defense of an algorithm with biased effects should depend, not on a claim of job-relatedness, but on the employer proving that the underlying model is statistically valid and substantively meaningful. Fourth, unlike under traditional disparate impact doctrine, employers should be able to rely on a "bottom-line" defense.

1. Data on Protected Class Characteristics

Understanding the sources of classification bias suggests quite different rules regarding information about protected characteristics such as race and sex. A formalist reading of Title VII might appear

---

226. *See id.* § 2000e-2(a)(2).

227. *See* Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 988 (1988); Connecticut v. Teal, 457 U.S. 440, 446 (1982); Griggs v. Duke Power Co., 401 U.S. 424, 429-30 (1971).

228. See *Watson*, 487 U.S. at 988 ("This Court has repeatedly reaffirmed the principle that some facially neutral employment practices may violate Title VII even in the absence of a demonstrated discriminatory *intent*." (emphasis added)); *Griggs*, 401 U.S. at 430 ("Under [Title VII], practices, procedures, or tests neutral on their face, and *even neutral in terms of intent*, cannot be maintained if they operate to 'freeze' the status quo of prior discriminatory employment practices." (emphasis added)).

to prohibit any use of variables capturing sensitive characteristics in a data model.[229] Certainly, a simple model that relied on race or other protected characteristics as the basis for adverse decisions would run afoul of Title VII's prohibitions. However, when dealing with a complex statistical model involving multiple variables, the appropriate treatment of these sensitive variables is more complicated. If the goal is to reduce biased outcomes, then a simple prohibition on using data about race or sex could be either wholly ineffective or actually counterproductive due to the existence of class proxies and the risk of omitted variable bias.[230] Instead, avoiding classification bias may sometimes call for excluding sensitive demographic variables and at other times call for *including* them. Any response to biased data models must be sensitive to these nuances.

Regardless of whether a particular model should include variables for protected characteristics, preventing classification bias requires that, at the very least, model creators preserve these data when they are already present in the training data.[231] If developers purge demographic variables such as race and sex from the dataset, it becomes more difficult, if not impossible, to determine whether a model is systematically biased. Preserving these variables allows a model to be tested to determine its effect on the distribution of opportunities among different groups. Thus, unlike standard readings of Title VII which might suggest that data on sensitive characteristics should be disregarded or deleted,[232] a focus on classification bias argues for preserving this data and using it to assess the risks that a model produces biased outcomes.

2. Relevant Labor Market Statistics

The Supreme Court has stressed the importance of identifying the correct labor pool for comparison purposes when using statistical

229. *See* Barocas & Selbst, *supra* note 5, at 694-95.

230. *See supra* Part I.B.

231. *See* Dwork & Mulligan, *supra* note 5, at 37 (arguing that having data about legally protected characteristics is necessary to avoid unintended biased outcomes); *cf.* Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.15 (2016) (requiring employers to maintain records and disclose the impact of tests and other selection procedures on employment opportunities).

232. *See* Barocas & Selbst, *supra* note 5, at 694-95.

evidence to establish disparate impact.[233] According to the Court, the "proper comparison [is] between the racial composition of [the at-issue jobs] and the racial composition of the qualified ... population in the relevant labor market."[234] This requirement has led to conflicts in particular cases over how to define the comparison pool—for example, what indicia should be used to identify "qualified" applicants and what geographic area constitutes the "relevant labor market."[235] How a court resolves these questions can determine whether complainants are successful in establishing a prima facie case of discrimination.[236]

The search for the proper comparator group makes sense when trying to diagnose whether an independently developed selection device, such as a written ability test, will have a disproportionate impact when a particular employer administers it. With data mining, however, the employment practice at issue—the predictive model—is derived from preexisting data about large numbers of individuals who are taken to be representative of the target population. By constructing the model from the data, the data miners implicitly assume that the dataset used to train the model is complete enough and accurate enough to identify meaningful patterns among applicants or employees. If the operation of the model on the training data demonstrates an adverse effect on a protected class, that showing should be sufficient to establish a prima facie case. A court should not require a complainant to collect additional

233. *See, e.g.*, Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 650-51 (1989), *superseded by statute*, Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071, *as recognized in* Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507 (2015); *Watson*, 487 U.S. at 997; N.Y.C. Transit Auth. v. Beazer, 440 U.S. 568, 585-86 (1979).

234. *Wards Cove*, 490 U.S. at 650 (quoting Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 308 (1977) (alterations in original)).

235. *See, e.g.*, Dothard v. Rawlinson, 433 U.S. 321, 330 (1977) ("The appellants argue that a showing of disproportionate impact on women based on generalized national statistics should not suffice to establish a prima facie case.... There is no requirement, however, that a statistical showing of disproportionate impact must always be based on analysis of the characteristics of actual applicants."); *In re* Emp't Discrimination Litig. Against Ala., 198 F.3d 1305, 1312 (11th Cir. 1999) ("The focus during this first stage of the inquiry, and indeed during the whole of the disparate impact analysis, is on defining the qualified applicant pool.").

236. *See, e.g.*, Peightal v. Metropolitan Dade County, 26 F.3d 1545, 1554-55, 1557 (11th Cir. 1994); Maddox v. Claytor, 764 F.2d 1539, 1555 (11th Cir. 1985).

data about some relevant comparator pool to establish the adverse effects of the model.

On the other hand, even if a model does not exhibit discriminatory effects when run on the training data, that fact cannot be taken as conclusive evidence that outcomes will be unbiased when a particular employer applies the model in the real world. If the data relied on to build the model were not sufficiently representative or accurate, the model may be statistically biased in ways that systematically disadvantage certain groups when applied to actual applicants or employees. Thus, courts should also permit complainants to demonstrate that the operation of the model on real cases produces biased outcomes.

3. Employer Justifications

Under disparate impact doctrine, an employer may defend against a prima facie showing of disparate impact by demonstrating that the challenged practice is "job related ... and consistent with business necessity."[237] The exact meaning of this phrase is ambiguous, and the standard has proven difficult to apply consistently in practice.[238] When applied to data analytics, however, it is difficult to make sense of the standard at all. When an algorithm relies on seemingly arbitrary characteristics or behaviors interacting in some complex way to predict job performance, the claim that it is "job related" often reduces to the fact that there is an observed statistical correlation. If a statistical correlation were sufficient to satisfy the defense of job-relatedness, the standard would be a tautology rather than a meaningful legal test. In order to protect against discriminatory harms, something more must be required to justify the use of an algorithm that produces biased outcomes.

As discussed in Part I.C, error-ridden, biased, or unrepresentative data, or improper specification of variables can introduce statistical bias, undermining the accuracy of a data model. When these statistical biases coincide with class membership, reliance on the model can harm members of protected groups. In order for claimants to diagnose whether statistical bias has infected an algorithm,

237. *See* Civil Rights Act of 1991 § 105, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

238. *See* Selmi, *supra* note 28, at 721-24; Wax, *supra* note 176, at 628, 631-36.

they would need access to the training data and the underlying model. The claimants would have to trace how the data miners collected the data, determine what populations were sampled, and audit the records for errors. Conducting these types of checks for a dataset created by aggregating multiple, unrelated data sources containing hundreds of thousands of bits of information would be a daunting task for even the best-resourced plaintiffs. In addition, the algorithm's creators are likely to claim that both the training data and the algorithm itself are proprietary information. Thus, if the law required complainants to prove the source of bias, they would face insurmountable obstacles.

Given these hurdles and the employer's superior access to information about the model's construction, employers should bear the burden of establishing the model's validity. The existence of a statistical correlation should not be sufficient. Instead, because the employer's justification for using an algorithm amounts to a claim that it actually predicts something relevant to the job, the employer should carry the burden of demonstrating that statistical bias does not plague the underlying model. In other words, the employer should have to defend the accuracy of the correlations it relies on by showing that no problems exist with the data or model construction that are biasing the results, and not simply by showing a statistical correlation in the existing data.

If an employer were able to satisfy this burden—if we could be certain that no statistical biases affected the model—should that be sufficient to justify reliance on an algorithm, even if it produces biased outcomes? In other words, should an employer be permitted to use a model that creates structural disadvantage if it is clear that it is not caused by statistical bias? Answering that question turns on the legitimacy of the employer's justification for using the model. And making that judgment requires knowing something about what the model is measuring and how it relates to the particular job. When applied to data analytics, however, two distinct problems arise. The first is the issue of interpretability. The second is the difficulty of distinguishing meaningful from spurious correlations.

The problem of interpretability arises because the atheoretical nature of data mining and the availability of unguided machine-learning techniques often make it difficult to know what factors are driving outcomes. An algorithm may be a "black box" that sorts

applicants or employees and predicts who is most promising, without specifying what characteristics or qualities it is looking for. It may, for example, be trained simply to look for applicants who resemble individuals hired in the past. Alternatively, the target variable might be clearly defined—as, for example, when an employer seeks employees who will maximize sales or have the longest job tenure—but it may not be possible to identify which particular attributes or variables are driving the algorithm or to determine how they are weighted.

Even when a model is interpretable, its *meaning* may not be clear. Two variables may be strongly correlated in the data, but the existence of a statistical relationship does not tell us if the variables are causally related, or are influenced by some common unobservable factor, or are completely unrelated. For example, one study found that employees who installed new web browsers on their computers rather than using preinstalled software stayed longer on the job.[239] But it is unclear why this correlation exists. It is possible, although unlikely, that not using the default browser makes an employee more dedicated. More likely, some unobserved attribute leads some individuals to choose a nonstandard browser, and also affects their longevity on the job. Or, it could be that the observed relationship between browser choice and productivity is entirely coincidental. Other correlations seem much more likely to be spurious—an artifact of the data mining process rather than a meaningful relationship—such as the apparent correlation between "liking" curly fries on Facebook and higher intelligence.[240]

Given the significant risks that biased algorithms will reproduce or entrench existing disadvantage, employers should bear the burden of justifying their use when they have disproportionate effects on protected groups. When a model is interpretable, debate may ensue over whether its use is justified, but it is at least possible to have a conversation about whether relying on the behaviors or attributes that drive the outcomes is normatively acceptable. When a model is not interpretable, however, it is not even possible to have

239. *See* Joe Pinsker, *People Who Use Firefox or Chrome Are Better Employees*, ATLANTIC (Mar. 16, 2015), http://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/ [https://perma.cc/4ZAA-LFLS].

240. *See* Kosinski et al., *supra* note 62, at 5804.

the conversation. In such a case, the employer should not be able to justify its use merely because it captures a statistical relationship.

4. The Bottom-Line Defense

Another way that Title VII doctrine should be adjusted is to allow employers a bottom-line defense when an algorithm is part of a larger selection process that is not biased overall. In 1982 the Supreme Court rejected the "bottom-line defense" in a disparate impact case, *Connecticut v. Teal*.[241] The plaintiffs in *Teal* alleged that their employer had violated Title VII by using a written exam that had a disparate impact on black employees as the first step in a promotion process.[242] Because black and white employees had significantly different passing rates, the proportion of black employees who continued to be eligible for promotion was much lower than that of white employees. When the employer later promoted some of these employees, it over selected black employees from among the eligible candidates. The end result was that 22.9 percent of the black employees who initially took the test were ultimately promoted, as compared with 13.5 percent of white employees.[243]

The employer argued that this "bottom-line" result, in which black employees were promoted at higher rates than white employees, should be a defense to the plaintiffs' Title VII suit.[244] The Supreme Court, in a five-to-four decision, rejected the employer's argument on the grounds that the goal of Title VII, as interpreted in *Griggs*, is "to achieve equality of employment *opportunities* and remove barriers" to equality.[245] In the Court's view, the ultimate outcome of the promotion process was irrelevant because the plaintiffs' claim was that they were denied "the *opportunity* to compete equally with white workers on the basis of job-related criteria."[246] The Court also argued that the focus of the statute's protection was the individual, not groups, and therefore, Title VII

241. 457 U.S. 440, 442, 452-56 (1982).
242. *Id.* at 443-44.
243. *Id.*
244. *See id.* at 452-53.
245. *See id.* at 448 (quoting Griggs v. Duke Power Co., 401 U.S. 424, 429-30 (1971)).
246. *Id.* at 451.

required the employer to afford each applicant an equal opportunity to compete.[247]

Regardless of whether the Court's rejection of the bottom-line defense made sense given the facts in *Teal*, addressing classification bias calls for a different approach. It is possible that relying on certain elements or factors in a data model may tend to disadvantage a protected group, but those effects might disappear when they are part of a more complex model that allows for interactions among multiple factors. Thus, including race or sex as a variable might not cause an overall discriminatory effect at all. In some circumstances, including these variables might even make the model less likely to have a discriminatory effect—thereby contributing to a more equal bottom line. Similarly, the inclusion of some neutral variables may bias outcomes based on protected characteristics but will not always do so, depending on the overall structure of the model. Because isolating the effect of particular variables is difficult, treating the algorithm as an undifferentiated whole will often make sense. And if the algorithm's operation does not disproportionately exclude members of protected groups, then no discriminatory harm has occured.

What if the operation of an algorithm produces biased outcomes, but the model's predictions are only one input in the employer's selection process, and, in the end, there is no disparate effect on a protected class? In that case, should the law still hold the employer responsible for relying on a biased data model as part of its process? In the context of workforce analytics, permitting a bottom-line defense makes sense. First, as discussed above, when dealing with algorithms plagued by statistical bias or reproducing structural disadvantage, the harm is systemic rather than individual.[248] Given that the central concern is with workplace systems that disadvantage certain groups, those concerns are alleviated when the operation of the system as a whole does not produce biased outcomes.

More practically, allowing employers a bottom-line defense is more likely to encourage equality-promoting uses of data. If employers are potentially liable for biased effects at each step of their

247. *Id.* at 453-56.

248. *See supra* Part I.C.

hiring or promotion process, they will have little incentive for self-examination or evaluation of the structural impact of their choices. Instead, they are likely either to ignore the risk that algorithms can cause bias or simply to cease using data analytics altogether. In contrast, a legal regime that permits a bottom-line defense will encourage employers to audit the impact of selection tools—including decision-making algorithms—on their workforce composition and to create processes that produce less biased results overall.

* * *

Thus far, this Part has considered how the law should look different from existing doctrine in order to respond to the equality challenges posed by workforce analytics. As explained, the law will have to depart from traditional disparate impact doctrine in significant ways in order to respond effectively. It might do so by recognizing classification bias as a separate type of harm prohibited by Title VII, or, alternatively, by adjusting disparate impact doctrine to be more responsive to the particular risks posed by discriminatory algorithms. Whether framed in terms of a prohibition on classification bias or a revised disparate impact theory, the critical point is that data analytics differ significantly from the employer practices challenged in earlier cases, and thus require a legal response adapted to those particular risks.

D. A Note on Ricci v. DeStefano

The previous Section discussed how Title VII might be applied in ways better suited to meet the challenges to equality posed by workforce analytics. In this Section, I consider whether anything in existing Title VII doctrine would preclude such a development. More specifically, some commentators have interpreted the Supreme Court's decision in *Ricci v. DeStefano* as casting doubt on the viability of disparate impact theory—and by implication, any doctrine that looks at the disparate effects of employer practices.[249] These concerns raise the question: does the Court's holding in *Ricci* bar the development of Title VII doctrine in ways that can meet the

249. *See, e.g.*, Primus, *supra* note 200, at 1344, 1363.

risks of classification bias? For reasons I explain below, I believe the answer is clearly "no." And for the same reasons, Title VII—read as a whole—should pose no barrier to employers' voluntary use of data analytics to try to diagnose and reduce structural forms of bias.

The dispute in *Ricci* arose when the City of New Haven, Connecticut, refused to certify the results of promotional exams.[250] After administering the written portion, the City realized that the exams would have a racially disparate impact if certified: virtually all of the promotions would go to white firefighters, even though a significant proportion of the candidate pool was black or Hispanic.[251] Concerned about a possible disparate impact lawsuit if it made the promotions, the City decided not to certify the results.[252] Some of the firefighters who believed that they would have been promoted sued the City.[253] These firefighters alleged that the City's refusal to use the test results constituted a form of disparate treatment discrimination in violation of Title VII and the Equal Protection Clause because the City had considered the racial impact of the tests in making its decision.

In *Ricci*, the five-justice majority accepted the plaintiffs' argument that the City's decision to discard the test results violated Title VII's disparate treatment prohibition with very little discussion.[254] The majority summarily rejected the district court's reasoning that the City's motivation of avoiding disparate impact liability did not constitute discriminatory intent. Writing for the majority, Justice Kennedy explained, "Our analysis begins with this premise: The City's actions would violate the disparate-treatment prohibition of Title VII absent some valid defense."[255] In the

250. *See* Ricci v. DeStefano, 557 U.S. 557, 562-63 (2009).

251. "Seventy-seven candidates completed the lieutenant examination—43 whites, 19 blacks, and 15 Hispanics. Of those, 34 candidates passed—25 whites, 6 blacks, and 3 Hispanics." *Id.* at 566. The top ten candidates were eligible to fill eight vacant lieutenant positions. *Id.* All ten candidates were white. *Id.* "Forty-one candidates completed the captain examination—25 whites, 8 blacks, and 8 Hispanics. Of those, 22 candidates passed—16 whites, 3 blacks, and 3 Hispanics." *Id.* The top nine candidates were eligible to fill seven vacant captain positions. *Id.* Seven of the candidates were white, and two were Hispanic. *Id.*

252. *See id.* at 562 (describing how the City threw out the examinations after some firefighters threatened to sue the City if it promoted firefighters on the basis of the tests).

253. *Id.* at 562-63.

254. *See id.* at 579-80.

255. *Id.* at 579.

majority's view, the fact that the City accounted for the racially disparate results made its decision a form of intentional discrimination, such that Title VII's disparate treatment and disparate impact prohibitions appeared to be in conflict.[256]

From this starting premise, the majority's analysis turned to whether the City had a lawful justification for taking the action it did. The Court rejected the City's argument that its good faith belief that using the exams would be a disparate impact violation justified discarding the test results.[257] The majority also rejected the plaintiffs' position that an employer may never take race-conscious actions even if the employer knows that it would otherwise violate disparate impact.[258] Instead, the majority concluded that the City must have "a strong basis in evidence to believe it will be subject to disparate-impact liability" to justify its actions.[259] Examining the record evidence, the majority concluded that New Haven lacked the requisite "strong basis in evidence," finding the exams "job related" and "consistent with business necessity."[260] The majority therefore held that discarding the test results violated Title VII.[261]

In the wake of *Ricci*, some commentators have suggested that disparate impact faces an existential threat.[262] If disparate treatment and disparate impact are in conflict, and if the Equal Protection Clause forbids disparate treatment, then is the disparate impact prohibition itself unconstitutional? Justice Scalia clearly

256. *See id.* at 579-80.
257. *Id.* at 581-82.
258. *See id.* at 580.
259. *Id.* at 585.
260. *See id.* at 587.
261. *Id.* at 592.
262. *See, e.g.*, Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After* Inclusive Communities, 101 CORNELL L. REV. 1115, 1126-27 (2016); Kenneth L. Marcus, *The War Between Disparate Impact and Equal Protection*, 2008-2009 CATO SUP. CT. REV. 53, 55; Eang L. Ngov, *When "The Evil Day" Comes, Will Title VII's Disparate Impact Provision Be Narrowly Tailored to Survive an Equal Protection Clause Challenge?*, 60 AM. U. L. REV. 535, 538-39 (2011); Primus, *supra* note 200, at 1343-44; Lawrence Rosenthal, *Saving Disparate Impact*, 34 CARDOZO L. REV. 2157, 2161-62 (2013); *see also* Richard A. Primus, *Of Visible Race-Consciousness and Institutional Role: Equal Protection and Disparate Impact After* Ricci *and* Inclusive Communities, *in* TITLE VII OF THE CIVIL RIGHTS ACT AFTER 50 YEARS: PROCEEDINGS OF THE NEW YORK UNIVERSITY 67TH ANNUAL CONFERENCE ON LABOR 295, 295-96 (Anne Marie Lofaso & Samuel Estreicher eds., 2015) (concluding in light of the Court's decision in *Inclusive Communities* that the statutory disparate impact standard will survive constitutional scrutiny given the current Court composition).

intended to signal a looming constitutional issue in his concurring opinion;[263] however, the rest of the Justices were content to argue the merits in *Ricci* on purely statutory grounds.[264] This approach is sensible because there is a vast difference between a constitutional prohibition on race-based state action and the conclusion that Congress cannot require employers to dismantle practices that operate as "built-in headwinds" for disadvantaged minority groups.[265] Despite the alarms, *Ricci* can easily be read as consistent with the continuing constitutionality of disparate impact liability under Title VII.[266] In *Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc.*, the Supreme Court held that disparate impact claims are cognizable under the Fair Housing Act.[267] This decision suggests that the theory will likely remain viable even if subject to a direct constitutional challenge.[268]

Putting aside the constitutional question—as the Court did in *Ricci*—the question is whether prohibiting classification bias that results from data models would conflict with Title VII's prohibition on intentional discrimination. The Justices in *Ricci* divided five to four over how to frame the question before the Court. While five Justices started from the premise that disparate treatment and disparate impact obligations were in conflict in the case,[269] the four dissenting Justices saw no conflict at all.[270] Justice Ginsburg, who authored the dissent, argued that the best reading of Title VII understands the disparate treatment and disparate impact theories as working in concert to achieve the statute's purposes of "ending workplace discrimination and promoting genuinely equal opportu-

263. *See Ricci*, 557 U.S. at 594 (Scalia, J., concurring) ("[R]esolution of this dispute merely postpones the evil day on which the Court will have to confront the question: Whether, or to what extent, are the disparate-impact provisions of Title VII of the Civil Rights Act of 1964 consistent with the Constitution's guarantee of equal protection?").

264. *See id.* at 576-78, 584 (majority opinion).

265. *See* Griggs v. Duke Power Co., 401 U.S. 424, 432 (1971).

266. *See* Primus, *supra* note 200, at 1374-75 (arguing that disparate impact doctrine will survive constitutional challenge under two of three proposed readings of *Ricci*); *cf. In re* Emp't Discrimination Litig. Against Ala., 198 F.3d 1305, 1324 (11th Cir. 1999) (finding that Title VII's disparate impact provisions are a valid exercise of Congress's Fourteenth Amendment enforcement power).

267. 135 S. Ct. 2507, 2525 (2015).

268. *See, e.g.*, Bagenstos, *supra* note 262, at 1127-28; Primus, *supra* note 262, at 295-96.

269. *See Ricci*, 557 U.S. at 580.

270. *See id.* at 624-25 (Ginsburg, J., dissenting).

nity."[271] In the view of the dissenting Justices, the employer who rejects criteria that systematically disadvantage minorities "due to reasonable doubts about their reliability can hardly be held to have engaged in discrimination 'because of' race."[272]

Thus, the Justices were closely divided on whether discarding New Haven's promotional exams constituted disparate treatment. An even stronger case can be made that abandoning a data model that produces racially biased results is not a form of disparate treatment. Richard Primus argues that one plausible reading of *Ricci* is that the City's actions constituted disparate treatment because they "adversely affected specific and visible innocent parties."[273] Certainly Primus is right that protecting the expectations of the plaintiffs was a significant concern for the Justices in the majority. Justice Kennedy wrote that the City "create[d] legitimate expectations" in the firefighters who took the tests.[274] Some, he noted, "invested substantial time, money, and personal commitment in preparing."[275] The problem arose because once the City established and announced the selection process, invalidating the test results upset legitimate expectations.[276] Justice Alito, in his concurrence, similarly emphasized the personal sacrifices that individual plaintiffs made to qualify for promotion—one firefighter hired someone to read and record the study materials because he was dyslexic, and another gave up a part-time job in order to study.[277]

---

271. *See id.* at 624.

272. *Id.* at 625.

273. Primus, *supra* note 200, at 1362. Some commentators have argued that the challengers were not in fact "victims" at all. *See, e.g.*, *Ricci*, 557 U.S. at 608 (Ginsburg, J., dissenting) ("[The white firefighters] had no vested right to promotion."); *see also* Mark S. Brodin, Ricci v. DeStefano: *The New Haven Firefighters Case & the Triumph of White Privilege*, 20 S. CAL. REV. L. & SOC. JUST. 161, 181, 202-12 (2011). Regardless, Primus is correct that the majority in *Ricci* viewed the challengers as victims because they relied on a process announced in advance. *See* Primus, *supra* note 200, at 1372-73.

274. *Ricci*, 557 U.S. at 583 (majority opinion).

275. *Id.* at 583-84.

276. *See id.* at 583-84, 593 ("The injury arises in part from the high, and justified, expectations of the candidates who had participated in the testing process on the terms the City had established for the promotional process."). *Contra id.* at 630 (Ginsburg, J., dissenting) ("The legitimacy of an employee's expectation depends on the legitimacy of the selection method.").

277. *See id.* at 607 (Alito, J., concurring).

This reading of *Ricci*—that the disparate treatment violation occurred because the City's action created "visible victims"[278]—is not only consistent with the language of the opinions, but it also best fits the statutory language. Title VII does not forbid any employer decision just because it is made with an awareness of race. Instead, it forbids "adverse employment actions" taken "because of an individual's race."[279] Unlike the situation in *Ricci*, prohibiting the use of a biased algorithm does not constitute a disparate treatment violation because there has been no adverse employment action. No employee has been deprived of a job to which he is entitled because no employee has any right or legitimate expectation that an employer will use any particular model. Because data mining models are atheoretical and typically based on past behavioral observations,[280] applicants are unlikely to know exactly which factors weigh into the model, and so they cannot argue that they relied on the process. The applicant who might have been selected if the employer had used a data mining model that it chose to discard is thus in an entirely different position from the white firefighters in *Ricci* who studied in reliance on the announced test. With no reliance interest and no entitlement that the employer use any particular model, employees who might have been hired if a biased model was used have no plausible claim that they have suffered discrimination.

Because disparate treatment violations occur only when employees' legitimate entitlements are disrupted, nothing in *Ricci* precludes interpreting Title VII to prohibit classification bias, nor would the decision prohibit employer attempts to identify and avoid such bias. Barocas and Selbst thus overstate the matter when they suggest that any legislation directed at reducing biased models might "run afoul of *Ricci*."[281] They argue that attempts to regulate data mining are problematic because diagnosing the impact of a model requires taking protected class characteristics into account.[282] As explained above, however, the problem in *Ricci* was not that the

278. *See* Primus, *supra* note 200, at 1345, 1369-75.
279. Civil Right Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(1) (2012).
280. *See supra* Part I.
281. *See* Barocas & Selbst, *supra* note 5, at 725.
282. *See id.* at 725-26.

City took action with an awareness of its racial impact, but that the action entailed adverse employment actions against identifiable persons. Merely being aware of the racial consequences of a selection process does not constitute disparate treatment. Similarly, an employer's efforts to understand the racial consequences of its processes in order to avoid bias does not violate Title VII.

Even the five Justices who disapproved of the City's actions in *Ricci* agreed on this point. As Justice Kennedy wrote, "Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race."[283] And, of course, the only way to ensure that a test is fair regardless of race is to pay attention to race. The clear implication is that mere race-consciousness in developing a selection criterion is not a violation of Title VII. Rather, the Supreme Court has repeatedly emphasized that voluntary compliance by employers is "the preferred means of achieving the objectives of Title VII"[284] and "essential to the statutory scheme."[285] As the majority in *Ricci* recognized, unless employers can act to *avoid* practices that have a disparate impact, the voluntary compliance efforts that Title VII calls for would come "to a near standstill."[286]

Barocas and Selbst also erroneously suggest that *Ricci* poses an obstacle to crafting a remedy for biased classification schemes.[287] They argue that "[a]fter an employer begins to use the model to make hiring decisions, only a 'strong basis in evidence' that the employer will be successfully sued for disparate impact will permit corrective action."[288] However, nothing in *Ricci* prevents a court

283. Ricci v. DeStefano, 557 U.S. 557, 585 (2009); *id.* at 628-29 (Ginsburg, J., dissenting) ("This Court has repeatedly emphasized that the statute 'should not be read to thwart' efforts at voluntary compliance. Such compliance, we have explained, is 'the preferred means of achieving [Title VII's] objectives.'" (alteration in original) (internal citations omitted) (first quoting Johnson v. Transp. Agency, 480 U.S. 616, 630 (1987); and then quoting Local No. 93, Int'l Ass'n of Firefighters v. City of Cleveland, 478 U.S. 501, 515 (1986))).

284. *Id.* at 581 (majority opinion) (quoting *Local No. 93*, 478 U.S. at 515).

285. *Id.* at 583 ("The standard leaves ample room for employers' voluntary compliance efforts, which are essential to the statutory scheme and to Congress' efforts to eradicate workplace discrimination.").

286. *Id.* at 581.

287. *See* Barocas & Selbst, *supra* note 5, at 725-26.

288. *Id.* at 726.

from enjoining the use of a biased model, or an employer from voluntarily ceasing to use the discriminatory algorithm once that bias has been detected. The majority in *Ricci* objected to *undoing* the results of the test once the employer announced and administered it;[289] the Court did not require the City to continue using the test results to make future promotion decisions. To suggest otherwise would lead to the absurd result that an employer, who ordinarily has a great deal of discretion to change its selection processes or criteria, would suddenly be prohibited from changing a practice the moment it learned that it had a disparate effect on a protected group. Such an outcome would produce the exact *opposite* effect that Congress intended Title VII to have—namely, it would freeze into place employer practices that work to systematically disadvantage minority applicants and employees. The way to avoid such an absurd result is to recognize that acting prospectively to prevent classification bias is not a form of intentional discrimination.

A remedy limited to prospective relief is entirely consistent with *Ricci*. Because applicants and employees have no entitlement that an employer will continue to use any particular selection device,[290] the employer harms no one if it discards one practice in favor of a different one. Things would be more complicated if a remedy required the employer to fire current employees who were hired using a biased selection device, but that has not been the type of remedy required in successful disparate impact suits, nor should it be a remedy in cases of classification bias. For similar reasons, employers would not run afoul of Title VII by voluntarily avoiding models that produce biased results. An employer might not be permitted to fire an employee solely because she was selected using a biased data model. However, Title VII should not be read to prohibit the employer from ceasing to use that model once it discovers the bias.

E. The Limits of the Liability Model

Prohibiting classification schemes that disadvantage protected classes is a promising avenue for addressing the equality concerns raised by workforce analytics. Such an approach is grounded in the

289. *See Ricci*, 557 U.S. at 585.
290. *See id.*

text of Title VII and consistent with the statute's purpose. Because the risks posed by workforce analytics stem from different sources than traditional forms of workplace testing,[291] it makes sense to tailor the doctrine to those particular risks rather than to mechanically apply the details of disparate impact doctrine that were developed in a different context. However, relying on the threat of legal liability to prevent classification bias has limitations as well.

In order to enforce its prohibitions on employment discrimination, Title VII relies on both individual and agency enforcement. After exhausting the administrative process, individual workers can file suit under Title VII and seek injunctive and monetary relief.[292] Under the current version of the law, a successful complainant is entitled to lost wages and other forms of equitable relief, compensatory damages, punitive damages (in cases in which the defendant acted with malice or reckless indifference), and attorneys' fees.[293] The law caps the total amount of compensatory and punitive damages based on the size of the employer.[294] This remedial structure is intended in part to incentivize aggrieved individuals to enforce the prohibition against employment discrimination.

In addition to individual suits, the EEOC also has enforcement powers.[295] The EEOC has authority to receive, investigate, and conciliate charges of discrimination under Title VII and other antidiscrimination statutes. In cases in which the EEOC has found cause to believe discrimination occurred but was unable to resolve the dispute through informal conciliation, the EEOC may choose to file suit on behalf of a complaining party.[296]

For several reasons, this scheme may be less effective at enforcing a prohibition on classification bias, as compared with other types of discrimination. First, as previously discussed, the harms that classification bias causes are structural rather than individual in nature.[297] Because the harms are more diffuse, individuals will find it extremely difficult to detect when a biased algorithm has

291. *See supra* Part I.B.
292. *See* Civil Rights Act of 1964 § 706, 42 U.S.C. § 2000e-5(f) (2012).
293. *See* Civil Rights Act of 1991 § 102, 42 U.S.C. § 1981a(a)-(b); 42 U.S.C. § 2000e-5(g).
294. *See* 42 U.S.C. § 1981a(b)(3).
295. *See id.* § 2000e-5.
296. *See id.* § 2000e-5(f).
297. *See supra* Part I.C.

produced an adverse outcome and to understand what caused the model to be biased. Even if these obstacles are overcome, the appropriate remedy would be structural in nature—namely, an injunction to revise or eliminate use of a biased model.[298] The reduced chance of receiving damages makes it less likely that individual employees will step forward to challenge instances of classification bias.

Individual complainants may not be reliable enforcers of a prohibition on classification bias for another reason. Detecting and pursuing claims of classification bias will be highly resource- and time-intensive. Even with a favorable legal regime, plaintiffs will need experts to determine whether data models are producing biased outcomes. Most individual plaintiffs will simply be financially unable to pursue such a case, particularly when the likelihood of a large damage award is slim.

The EEOC might step into the breach, as it often does, by litigating cases that have the potential for significant public impact, but that private litigants are unlikely to pursue.[299] Even if the EEOC makes these cases a priority, however, its limited resources will significantly constrain its efforts. Currently, the EEOC receives nearly 100,000 new charges annually and it faces a persistent backlog of charges.[300] The EEOC's current strategic priorities include cases involving systemic discrimination.[301] That focus would seem to encompass the structural harms threatened by employer reliance on biased data models. If the EEOC decides to prioritize cases involving workforce analytics, it would need to develop methods for detecting when data algorithms are producing discriminatory outcomes. Doing so would require a level of technical expertise and fiscal resources even beyond what is currently needed to tackle large scale systemic cases.[302]

Lowering the standards for establishing liability or increasing the available remedies could resolve the problem of insufficient incentives for private litigants to file suit. If the law swings too sharply

298. *See supra* Parts I.C, III.D.

299. *See* Pauline T. Kim, *Addressing Systemic Discrimination: Public Enforcement and the Role of the EEOC*, 95 B.U. L. REV. 1133, 1141-46 (2015).

300. *See id.* at 1144.

301. *See id.* at 1141-42.

302. *Cf. id.* at 1145-46.

in that direction, however, it may deter employers from attempting to understand whether their data tools have any disparate effects, and they may prefer instead to remain ignorant of any biases those tools may be causing. Alternatively, employers may cease using data models altogether, even though data analytics might help to diagnose and correct existing cognitive or structural biases. Thus, the goal of the law should not be to eliminate the use of all data analytics. Instead, the optimal legal regime would deter the use of biased data models while permitting or encouraging equality-promoting uses of data. The difficulty of balancing these two goals under Title VII suggests that policymakers may need to look beyond a backward-looking, liability-based regime and to consider other regulatory responses.

Fully exploring alternative regimes goes beyond the scope of this Article, but a few examples are illustrative. Technological innovations may make it possible to limit in advance whether a computer will produce an algorithm with a disparate effect on a protected class.[303] Another possibility would be to develop an ex ante regulatory regime to govern algorithms like the one currently used for premarket approval of drugs.[304] An appropriately structured approval process could ensure that data mining models are not statistically biased and that the social costs of using them do not exceed the benefits. Alternatively, a regulatory body might work to develop standards relating to data collection, integrity and preservation, and model validity, such that models that complied with these standards would have a presumption of legality.

None of these alternatives is simple or guaranteed to work, and all are likely to generate resistance. Implementing any of these solutions would require resolving difficult questions about what kinds of bias are unfair and how much should be tolerated. But

303. *See, e.g.*, Sara Hajian & Josep Domingo-Ferrer, *Direct and Indirect Discrimination Prevention Methods*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 56, at 241, 247-51; Faisal Kamiran, Toon Calders & Mykola Pechenizkiy, *Techniques for Discrimination-Free Predictive Models*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 56, at 223, 229-35; Kroll et al., *supra* note 5 (manuscript at 35-45).

304. *Cf.* Andrew Tutt, *An FDA for Algorithms*, 68 ADMIN. L. REV. (forthcoming 2017) (manuscript at 20-25), https://ssrn.com/abstract=2747994 [https://perma.cc/2WCH-WMB9] (arguing for a federal regulatory agency to ensure the safety and efficacy of algorithms before they are introduced in the market).

these types of efforts might offer some kind of safe harbor to employers who, acting in good faith, attempt to leverage data to remove bias from their personnel practices.

CONCLUSION

The data revolution is here to stay. Advances in computing power and the availability of massive amounts of data make it inevitable that employers will harness these tools to manage their workforces. Depending on how employers deploy these tools, data may enhance workplace fairness or exacerbate inequality. When these tools are used—not as guides or aids, but as gatekeepers to critical employment opportunities—they risk reinforcing existing patterns of disadvantage. Because of the nature of data mining techniques, employer reliance on these tools poses novel challenges to workplace equality and thus traditional doctrine will not suffice to address them.

Thinking in terms of classification bias offers a lens through which to better understand these challenges and to consider how to develop an appropriate legal response. Although the term may sound novel, a legal prohibition of classification bias is grounded in the text of Title VII and fully consistent with its purposes. Whether recognized as a distinct type of discrimination under Title VII or a species of disparate impact theory, classification bias offers a way for rethinking how antidiscrimination law should be tailored to respond to the unique challenges raised by data-driven forms of discrimination. Doing so is essential for Title VII's vision of workplace equality to continue to advance in the face of evolving threats.