Spring 4-2020

Is Algorithmic Affirmative Action Legal?

Jason R. Bent

ARTICLES

Is Algorithmic Affirmative Action Legal?

JASON R. BENT*

TABLE OF CONTENTS

* Associate Dean for Academic Affairs and Professor of Law, Stetson University College of Law; University of Michigan Law School, J.D. 2000; Grinnell College, B.A. 1997. This Article won the 2019 Southeastern Association of Law Schools (SEALS) Call for Papers contest and was presented at the SEALS Annual Meeting on July 29, 2019. The project was generously supported by the administration and faculty of the Stetson University College of Law, and was significantly improved by comments from Charles Sullivan, Ramona Paetzold, Steve Willborn, Deborah Hellman, Stephanie Bornstein, Louis Virelli III, Anupam Chander, and Gregory Stein, as well as discussions with Matthew Kusner, Andrew Selbst, Carla Reyes, Marco Jimenez, Kirsten Davis, Roy Gardner, and James Fox. Thank you to the editors of *The Georgetown Law Journal* for insightful comments and suggestions that have substantially improved this Article.

INTRODUCTION

Employers cannot use different employment tests for different racial groups, even if they might have good, nondiscriminatory reasons for doing so. Imagine that, in making promotions to the rank of firefighter captain, the City of Springfield administers an exam with two parts—a written component and a field test component. Historically, the city has weighted the written and field components equally (50% each). If the city obtains evidence that the written test is highly predictive of future success for white candidates but not for nonwhite ones, and that the opposite is true for the field test, then the city may have legitimate reasons for weighting the exam differently for the two groups. Based on its data, the city might want to weight the exam for white candidates at 75% written test and 25% field test but reverse those percentages for nonwhite candidates. The city may also have other reasons to weight components differently, including attaining diversity goals or avoiding disparate impact from equally weighted testing.

Despite the city's legitimate ends, differentially weighting the exam according to race would violate Title VII of the Civil Rights Act[1] and may also violate the

1. *See* Civil Rights Act of 1991, Pub. L. No. 102-166, § 106, 105 Stat. 1071, 1075 (codified at 42 U.S.C. § 2000e-2(l) (2012)) ("It shall be an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests on the basis of race, color, religion, sex, or national origin."); Dean v. City of Shreveport, 438 F.3d 448, 462–63 (5th Cir. 2006) (holding that separating civil service exam scores by race and sex constituted the use of different score cutoffs in violation of Title VII); *see also* CHARLES A. SULLIVAN & MICHAEL J. ZIMMER, CASES AND MATERIALS ON EMPLOYMENT DISCRIMINATION 224 (9th ed. 2017) (explaining that differential test validation by race violates section 703(l) of Title VII).

constitutional guarantee of equal protection.[2] A 1991 amendment to Title VII forbids adjusting the scores of, or using different cutoff scores for, "employment related tests" on the basis of race.[3] The amendment banned the controversial practice of "race-norming" aptitude test scores.[4] Following the 1991 amendment, employers may not use different test standards for different racial groups, even if there might be a legitimate business reason or valid affirmative action justification for doing so.[5]

Now, what if instead of differentially weighting two exam components, the city wanted to make its promotion decisions using a machine-learning algorithm, and that algorithm weighted some observed variables differently for white and nonwhite candidates? Machine learning is an "automated process of discovering correlations" that can be used "to make predictions or estimates of some outcomes."[6] The relevant prediction here is whether the applicant for promotion is likely to be successful in the position (here, of captain). The data made available to the algorithm when searching for correlations could include information about each applicant's race. In other words, the race variable could be used by the algorithm in generating its predictions. If the employer's end goal is legitimate, is using a *race-aware algorithm* legal? What if the employer's legitimate goal for using race in the algorithm is to counteract the statistical disparate impact on racial minorities that the algorithm would otherwise produce if it were blinded to race? Courts will soon be facing these questions, and the answers are not clear. Race-aware algorithms will force courts to confront anew tensions between competing purposes of discrimination law.[7] Law that was developed to govern human

2. *See* U.S. CONST. amend. XIV; *Dean*, 438 F.3d at 454–62 (remanding for a determination of whether city's race-conscious hiring process satisfied strict scrutiny).

3. *See* Civil Rights Act of 1991 § 106 (amending section 703 of the Civil Rights Act of 1964). In the context of physical strength and stamina tests, courts have upheld the use of different cutoffs for men and women—for example, thirty push-ups for men but fourteen for women. *See* Bauer v. Lynch, 812 F.3d 340, 351 (4th Cir. 2016). This "gender norming" of physical tests has been justified on the basis that "[m]en and women simply are not physiologically the same for the purposes of physical fitness programs," and the different cutoffs were "equally burdensome" for men and women. *See id.* at 349–51.

4. *See* Ann C. McGinley, *Cognitive Illiberalism, Summary Judgment, and Title VII: An Examination of* Ricci v. DeStefano, 57 N.Y.L. SCH. L. REV. 865, 891 (2013) ("Race-norming is prohibited [by section 2000e-2(l)] because it is a practice of grading people of different races according to different standards."); Chi. Firefighters Local 2 v. City of Chicago, 249 F.3d 649, 655–56 (7th Cir. 2001) (finding that altering scores to equalize the mean values for each racial group or adding ten points to each black candidate's score would be race-norming in violation of section 2000e-2(l) of Title VII).

5. *See Dean*, 438 F.3d at 462–63; *see also Chi. Firefighters Local 2*, 249 F.3d at 655–56.

6. David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 671 (2017).

7. *See* Samuel R. Bagenstos, *Bottlenecks and Antidiscrimination Theory*, 93 TEX. L. REV. 415, 415 (2014) (book review) ("In American antidiscrimination theory, two positions have competed for primacy. One, anticlassification, sees the proper goal of antidiscrimination law as being essentially individualistic. . . . The other position, antisubordination, sees the proper goal of antidiscrimination law as being more group oriented." (footnote omitted)); *see also* Bradley A. Areheart, *The Anticlassification Turn in Employment Discrimination Law*, 63 ALA. L. REV. 955, 960–61 (2012) (exploring the tension between antisubordination and anticlassification); Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 9–14 (2003) (challenging the "assumption that . . . the anticlassification principle triumphed over the

decisions, including the statutory prohibition on race-norming tests and the Court's equal protection rulings, may not translate readily to the governance of machine-assisted decisions. Legislators and jurists did not have in mind the possibility that powerful computing machines might be capable of formally removing protected characteristics from consideration, yet still inescapably and inadvertently reproducing preexisting human biases by leveraging correlations.

The basic problem of unintentional algorithmic discrimination is by now well-recognized.[8] A recent Reuters report provides an example: Amazon used a secret machine-learning algorithm in recruiting employees, and it systematically favored men over women in scoring potential job candidates.[9] Amazon gave up on that project, but it, and other companies, continue to pursue algorithmic decision processes in personnel matters.[10] Scholars worry that algorithmic decisions will evade regulation under current disparate impact and disparate treatment doctrine.[11] Machine-learning algorithms explore data to identify correlations with

antisubordination principle"); Reva B. Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE L.J. 1278, 1281 (2011) (noting that Supreme Court Justices disagree on whether the Equal Protection Clause dictates antisubordination or anticlassification).

8. *See generally* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 3 (2016) (many mathematical predictive models "encode[] human prejudice, misunderstanding, and bias"); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 674–75 (2016) (explaining that "data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society"); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1039 (2017) (explaining that automated algorithms present problems of "intentional invidious discrimination" and "replicating real world inequalities"); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 13–16 (2014) (noting that credit-scoring algorithms "systematiz[e] [discriminatory practices] in hidden ways" (footnote omitted)); James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 169 (2016) (noting one commentator's argument that under the current constitutional order, data mining is "permitted to exacerbate existing inequalities in difficult-to-counter ways"); Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633, 662–63 (2017) (explaining that artificial intelligence systems are "not immune to inherent racial biases" and that judgments from these algorithms might be suspect); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 190 (2017) (describing the "classification bias" of algorithms); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 680 (2017) ("A significant concern about automated decisionmaking is that it may simultaneously systematize and conceal discrimination."); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 120–23 (2017) (providing examples of unintentional discrimination by algorithms, including Google's AdWords linking "'black-sounding' names to criminal records").

9. *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 9, 2018, 11:12 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/B3JA-QX76].

10. *See id.* (reporting that Amazon has tasked a new group with developing artificial intelligence techniques to improve diversity in hiring); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 860 (2017) ("Employers are increasingly relying on data analytic tools to make personnel decisions, thereby affecting who gets interviewed, hired, or promoted." (footnote omitted)).

11. *See, e.g.*, Barocas & Selbst, *supra* note 8, at 694–714; Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 525–26 (2018) (arguing that stereotyping theory could be effective in regulating discriminatory algorithms); Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395, 428–29

some target characteristic (for example, a good employee). Due to this defining feature, the structure of the algorithm itself may provide a business necessity defense to disparate impact liability.[12] And, because the algorithm is making predictions based on data—without any discriminatory animus or intent—an employer relying on the algorithm may not be engaged in unlawful disparate treatment.[13]

This perceived shortcoming of antidiscrimination law is prompting calls for legal reform, including for modification of the business-necessity defense to disparate impact claims[14] and creation of a regulatory agency to approve algorithms.[15] Others propose technological cures to ensure fairness in algorithm design.[16] But, as we will see, algorithmic fairness cannot be accomplished just by hiding protected characteristics from the machine. Attempting to build colorblind algorithms is an exercise in futility. Machine-learning scholars increasingly agree that the best way to get fair algorithmic results is not by hiding the protected trait, but instead by *using* the protected trait to set a fairness constraint within the algorithm design.[17] Machine-learning scholars are developing an arsenal of mathematical techniques to achieve fairness, some of which look like the algorithmic equivalent of weighting employment tests differently by racial group.[18] Professor

(2018) (noting that machine learning has exposed gaps in the Supreme Court's categories of disparate treatment and disparate impact discrimination).

12. *See* Barocas & Selbst, *supra* note 8, at 709 ("Data mining will likely only be used if it is actually predictive of *something*, so the business necessity defense solely comes down to whether the trait sought is important enough to job performance to justify its use in any context."). *But see* Kim, *supra* note 10, at 909–25 (arguing that a closer reading of Title VII's statutory text reveals a possible theory beyond disparate treatment and disparate impact that would "directly prohibit classification bias when algorithms operate to systematically disadvantage protected groups").

13. *See* Barocas & Selbst, *supra* note 8, at 698 ("Except for masking [that is, the use of algorithms to hide true, intentional discrimination], discriminatory data mining is by stipulation unintentional.").

14. *See* Grimmelmann & Westreich, *supra* note 8, at 173–74 (using a hypothetical judicial opinion to urge modification of the disparate impact business necessity defense to include a burden on the defendant employer to explain *why* an algorithmic model works to predict success if the model's results correlate with protected characteristics).

15. *See generally* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017).

16. "Fairness" is subject to multiple possible definitions, some of which are incompatible with each other. *See infra* Section II.A; *see also* RICHARD BERK ET AL., FAIRNESS IN CRIMINAL JUSTICE RISK ASSESSMENTS: THE STATE OF THE ART 3, 12–15 (2017), https://arxiv.org/pdf/1703.09207.pdf [https://perma.cc/9DDK-4EMJ] ("But, even when attempts are made to clarify what fairness can mean, there are several different kinds that can conflict with one another and with accuracy." (citation omitted)).

17. *See generally* CYNTHIA DWORK ET AL., DECOUPLED CLASSIFIERS FOR FAIR AND EFFICIENT MACHINE LEARNING 2 (2017) [hereinafter DWORK ET AL., DECOUPLED CLASSIFIERS], https://arxiv.org/abs/1707.06613 [https://perma.cc/2AYK-475J]; *see also* CYNTHIA DWORK ET AL., FAIRNESS THROUGH AWARENESS 214 (2011) [hereinafter DWORK ET AL., FAIRNESS THROUGH AWARENESS], https://arxiv.org/pdf/1104.3913.pdf [https://perma.cc/NHZ4-K9HN]; *infra* Section I.B (discussing algorithmic fairness).

18. *See* DWORK ET AL., DECOUPLED CLASSIFIERS, *supra* note 17, at 2 ("Using sensitive attributes [in the algorithm] may increase accuracy for all groups and may avoid biases where a classifier favors members of a minority group that meet criteria optimized for a majority group . . . ."). Dwork suggests training machine-learning classifiers separately for each group. *See id.* at 2; *see also* Dina Bass & Ellen Huet, *Researchers Combat Gender and Racial Bias in Artificial Intelligence*, GLOBE & MAIL (Dec. 4, 2017), https://www.theglobeandmail.com/report-on-business/researchers-combat-gender-and-racial-bias-in-artificial-intelligence/article37177350/ [https://perma.cc/2CMJ-4FV5] ("So for example, female engineering applicants can be evaluated by the criteria best suited to predicting a successful female

Anupam Chander has called for "algorithmic affirmative action,"[19] arguing that achieving fairness requires algorithms to take protected characteristics into account and use them to counteract other unidentified sources of bias.[20]

Is algorithmic affirmative action legal? This question has two components: (1) whether algorithmic affirmative action is consistent with statutory prohibitions on intentional, disparate treatment discrimination; and (2) whether it is consistent with the constitutional guarantee (where applicable) of equal protection. Data scientists have noted the legal issues, but left them for legal scholarship.[21] The few legal scholars to confront such questions have disagreed about whether *Ricci v. DeStefano*[22]—the Supreme Court's 2009 opinion addressing tensions between the disparate impact and disparate treatment theories of discrimination—represents a legitimate threat to algorithmic fairness solutions.[23] In *Ricci*, the Court held that the City of New Haven's refusal to certify firefighter promotion test results due to concerns about the racially disparate impact of the test results constituted disparate treatment because of race.[24] The limited scholarly

engineer and not be excluded because they don't meet criteria that determine success for the larger group."); Kevin Hartnett, *How to Force Our Machines to Play Fair*, QUANTA MAG. (Nov. 23, 2016), https://www.quantamagazine.org/making-algorithms-fair-an-interview-with-cynthia-dwork-20161123/# [https://perma.cc/7WWF-442E]; Gal Yona, *A Gentle Introduction to the Discussion on Algorithmic Fairness*, TOWARDS DATA SCI. (Oct. 5, 2017), https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6 [https://perma.cc/DL29-ZFTV] ("The 'aware' approach . . . [could pick] up on the fact that learning Physics is a stronger signal for future success among males than it is for females.").

19. *See* Chander, *supra* note 8, at 1025 ("My central claim is this: if we believe that the real-world facts, on which algorithms are trained and operate, are deeply suffused with invidious discrimination, then our prescription to the problem of racist or sexist algorithms is *algorithmic affirmative action*." (footnote omitted)); *see also* Kroll et al., *supra* note 8, at 637 (arguing that *ex ante* technological "fairness" solutions—possibly including "fair affirmative action"—embedded within the algorithm structure are preferable to *ex post* solutions).

20. *See* Chander, *supra* note 8, at 1041 ("The counterintuitive result of affirmative action is that the decisionmaker must take race and gender into account in order to ensure the fairness of the result.").

21. *See, e.g.*, BERK ET AL., *supra* note 16, at 3 ("No effort is made here to translate formal definitions of fairness into philosophical or jurisprudential notions in part because the authors of this paper lack the expertise and in part because that multidisciplinary conversation is just beginning." (citations omitted)); DWORK ET AL., DECOUPLED CLASSIFIERS, *supra* note 17, at 2 ("[W]e consider *how* to use a sensitive attribute such as gender or race to maximize fairness and accuracy, assuming that it is legal and ethical."); Zachary C. Lipton et al., Does Mitigating ML's Impact Disparity Require Treatment Disparity 9 (Jan. 11, 2019) (unpublished manuscript), https://arxiv.org/abs/1711.07076 [https://perma.cc/2JWT-9ZA6] (citing legal scholars for the proposition that explicit treatment disparity may be legally tolerated).

22. 557 U.S. 557 (2009). The facts and holding of *Ricci* are outlined *infra* Section II.A.2.

23. *Compare* Kroll et al., *supra* note 8, at 692–95 (observing that *Ricci* may raise "legal difficulties with correcting discriminatory algorithms ex post"), *and* Barocas & Selbst, *supra* note 8, at 725–26 (noting that legislation requiring or enabling employers to audit algorithms for bias "may run afoul of *Ricci*"), *with* Kim, *supra* note 8, at 197–202 (arguing that Kroll et al. misread *Ricci*, and that "nothing in *Ricci* prohibits revising an algorithm after discovering that it has discriminatory effects"), Kim, *supra* note 10, at 930–32 (arguing that Barocas & Selbst overstate the possibility that *Ricci* could be an obstacle), *and* Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67, 131–32 (2017) (agreeing with Kim).

24. *See* 557 U.S. at 563 (holding that such "race-based action" would be justified only where the employer had "a strong basis in evidence that, had it not taken the action, it would have been liable under the disparate-impact statute").

discussion of *Ricci* in the context of machine learning has glossed over the mechanics of race-aware algorithmic fairness solutions, as well as some important distinctions from the use of race by the City of New Haven. And beyond the threat posed by *Ricci*, there remains an equal protection challenge to state actors' use of race-aware fairness techniques to avoid algorithmic disparate impact.[25]

This Article is the first to comprehensively explore whether algorithmic affirmative action is lawful. It concludes that both statutory and constitutional antidiscrimination law leave room for race-aware affirmative action in the design of fair algorithms. Along the way, the Article recommends some clarifications of current doctrine and proposes the pursuit of formally race-neutral methods to achieve the admittedly race-conscious goals of algorithmic affirmative action.

The Article proceeds as follows. Part I introduces algorithmic affirmative action. It begins with a brief review of the bias problem in machine learning and then identifies multiple design options for algorithmic fairness. These designs are presented at a theoretical level, rather than in formal mathematical detail. It also highlights some difficult truths that stakeholders, jurists, and legal scholars must understand about accuracy and fairness trade-offs inherent in fairness solutions. Part II turns to the legality of algorithmic affirmative action, beginning with the statutory challenge under Title VII of the Civil Rights Act. Part II argues that voluntary algorithmic affirmative action ought to survive a disparate treatment challenge under *Ricci* and under the antirace-norming provision of Title VII. Finally, Part III considers the constitutional challenge to algorithmic affirmative action by state actors. It concludes that at least some forms of algorithmic affirmative action, to the extent they are racial classifications at all, ought to survive strict scrutiny as narrowly tailored solutions designed to mitigate the effects of past discrimination.

I. WHAT IS ALGORITHMIC AFFIRMATIVE ACTION?

A. ALGORITHMIC BIAS

Machine-learning algorithms can improve efficiency in decisionmaking and may help minimize some forms of subjective human biases. But they can also inadvertently reproduce existing bias, as Amazon's experience illustrates. Amazon wanted to more quickly and efficiently identify strong candidates for employment, so it began using an "experimental hiring tool" that used a machine-learning algorithm to search the Internet and mechanically identify and rate potential job candidates.[26] Machine-learning algorithms necessarily rely on past examples (the training data) to identify correlations and patterns that are then exploited to

25. *See Ricci*, 557 U.S. at 594–95 (Scalia, J., concurring).

26. *See* Dastin, *supra* note 9; *see also* Jordan Weissmann, *Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women*, SLATE (Oct. 10, 2018, 4:52 PM), https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html [https://perma.cc/PP2N-L8NM].

make predictions or score candidates on target variables of interest.[27] For example, an algorithm might be trained on past examples of good employees (who performed well by some quantifiable metric) and bad employees (who did not). Or the algorithm may be trained on past examples of good credit risks (who missed no more than *x* payments) and bad ones (who missed more than *x* payments).[28] A machine-learning algorithm can then explore everything known about those examples, including all the variables or "features,"[29] that are coded for each individual and identify patterns. It can then use those patterns—and the known variables for a new applicant—to predict whether that person will be a good employee or a good credit risk.[30]

This automated process can remove some of the subjectivity and bias of human decisionmaking, but it may also encode preexisting bias reflected in the information that the algorithm uses as examples. That is what happened in Amazon's experiment. "In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word 'women's,' as in 'women's chess club captain.' And it downgraded graduates of two all-women's colleges . . . ."[31] The algorithm was trained on data from employment applications actually received by Amazon over a ten-year period, and most of those applications were received from men, reflecting a social phenomenon that men tend to dominate employment in technology fields.[32] After observing the disparities generated by its hiring tool, Amazon ultimately "lost hope for the project."[33] According to the report, Amazon still uses a "much-watered down version" of the algorithmic tool, only to perform basic administrative sorting tasks like removing duplicate candidate profiles.[34]

What happened with Amazon is only one illustration of how bias can creep into algorithmic output. In an article widely cited in both the legal and machine-learning literature, Solon Barocas and Andrew Selbst comprehensively describe

27. *See* Kim, *supra* note 10, at 874 ("In order to build a model, its creators must select the data that they will use to build it—the 'training data.'"); Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 89–91 (2014) ("Such algorithms are designed to detect patterns among data. In a typical process, a machine learning algorithm is 'trained' to recognize[, for example,] spam emails by providing the algorithm with known examples of spam for pattern analysis."). This describes a "supervised" machine-learning process, in which human modelers "must actively specify a target variable of interest." Barocas & Selbst, *supra* note 8, at 678 n.24.

28. *See* Barocas & Selbst, *supra* note 8, at 681 (describing the subjective nature of labeling a target variable like creditworthiness).

29. *See id.* at 688 (noting that selecting which features to include in a model involves determining "what attributes [organizations and their data miners] observe and subsequently fold into their analyses").

30. *See id.* at 677 (machine learning "automates the process of discovering useful patterns, revealing regularities upon which subsequent decision making can rely").

31. Dastin, *supra* note 9; *see also* Bornstein, *supra* note 11, at 521.

32. *See* Dastin, *supra* note 9. Note that this social phenomenon may be the product of individual tastes and preferences, the product of potential applicants' knowledge of discriminatory hiring patterns of the particular employer or of the industry, or some combination of these. The true cause of the bias, whether social tastes or past discriminatory practices, cannot be discerned from the data alone.

33. *Id.*

34. *Id.* (internal quotation marks omitted).

the various ways that bias can be unintentionally introduced at different stages of the machine-learning process.[35] A basic understanding of these potential sources of unintended bias is required before examining the proposed fairness solutions. I briefly summarize below four critical vulnerabilities in the machine-learning process, but credit for the careful elucidation of these bias sources belongs to Barocas, Selbst, and others who have added levels of technical detail on the mechanics of machine-learning algorithms.[36]

1. Labeling Target Variables

First, discrimination can be the result of choices made when defining the target variables for the algorithm. There is no single, objective definition of a "good" employee or a "good" employment candidate, so the way the user defines and assigns a specific value to the target variable, if correlated with a protected characteristic, could unintentionally trigger a disparate impact.[37] For example, the employer's coding of a "good" employee can be subjective and track closely to characteristics more heavily represented in the population of male employees than female employees.

For the City of Springfield, constructing a machine-learning algorithm to identify those who would be "good" fire captains would begin by labeling past examples of good and bad captains. But how would the city identify and label which prior captains had been good captains and which had not? Would labels be based on the subjective ratings of each captain by the city's fire chief? Would it be based on some objective data about the performance of the firefighters under each captain's command? Would it be based on some other objective performance data, such as average response times to emergencies by the stations and personnel under the captain's command? Labeling the target variable (the "good" fire captain) will be driven by what the employer values in a successful fire captain and will necessarily be an exercise in human judgment. Because of that, the selection and labeling of the target variable creates a vulnerability in the machine-learning process that can lead to the reproduction of human bias.

---

35. *See* Barocas & Selbst, *supra* note 8, at 677–93. Barocas and Selbst also discuss ways that machine learning can be used to "mask" old-fashioned intentional discrimination by an organization or decisionmaker. *See id.* at 692–93. If proven, such actions would trigger disparate treatment liability. For purposes of this Article, I will set aside the potential for masking and focus only on unintended discrimination. However, Barocas and Selbst highlight an important and interesting question about whether an actor's intentional use of an algorithm, having only knowledge that it leads to discriminatory results, is sufficient to establish disparate treatment liability without a further showing that the actor actually had an intent or motive to discriminate. *See id.* at 700–01. For the influence of Barocas and Selbst in both the law and machine-learning literature, see Zachary C. Lipton et al., Does Mitigating ML's Impact Disparity Require Treatment Disparity? 14 (Mar. 2, 2018) (unpublished manuscript), http://zacklipton.com/media/papers/impact-disparity-treatment-disparity-lipton2018.pdf [https://perma.cc/6NNE-CTR5] (noting the article's cross-disciplinary audience).

36. *See, e.g.*, Lehr & Ohm, *supra* note 6, at 665–66 (crediting Barocas and Selbst for "being one of the earliest to provide a detailed legal analysis of data mining," and building upon that work with a more detailed exploration of the stages of the machine-learning process).

37. *See* Barocas & Selbst, *supra* note 8, at 679–80; *see also* Lehr & Ohm, *supra* note 6, at 703 ("[O]utcome variables can be disadvantageously defined . . . .").

2. Training Data: Encoding Preexisting Bias and Selection Bias

Second, the characteristics of the training data can lead to biased results in two ways. If the training data include examples that are themselves tainted by discrimination or bias, then the algorithm will simply encode and reproduce that bias.[38] For example, if one feature in the algorithm is prior performance review scores, and those performance reviews are subjectively scored by a human being who discriminated against women (whether intentionally or not), then the algorithm's output will likewise reproduce that discrimination against women.[39]

Barocas and Selbst illustrate the problem of biased training data with an oft-cited real world example: the discriminatory admission of medical school applicants to St. George's Hospital in the United Kingdom. Historically, the hospital had systematically discriminated against minorities and women in considering applicants. When prior admissions decisions were used as training data, the algorithm encoded and reproduced that discrimination when considering new applicants, thus perpetuating the preexisting human bias.[40]

Training data can also lead to discriminatory output if that training data are drawn from a biased sample of the relevant population because certain types of individuals are overrepresented or underrepresented in that training data.[41] If, for example, data are more readily available for men than women, or for younger applicants than older applicants, the algorithm may unintentionally disfavor the underrepresented group due to the data's inaccurate reflection of the relevant population. The Amazon case illustrates one form of biased selection of training data. Amazon trained its algorithm with data drawn from job applications that it had actually received. Because women were significantly underrepresented in the training data, the algorithm learned to disfavor women.

38. *See* Barocas & Selbst, *supra* note 8, at 681–82; *see also* Lehr & Ohm, *supra* note 6, at 703 ("[D]ata can have baked into them preexisting human biases . . . .").

39. *See* Barocas & Selbst, *supra* note 8, at 681–82; *see also* Dino Pedreschi et al., *Discrimination-aware Data Mining*, *in* YING LI ET AL., KNOWLEDGE DISCOVERY & DATA MINING, KDD '08: PROCEEDINGS OF THE 14TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 560, 560 (2008), http://pages.di.unipi.it/ruggieri/Papers/kdd2008.pdf [https://perma.cc/5JRA-4KAL] ("[L]earning from historical data may mean to discover traditional prejudices that are endemic in reality, and to assign to such practices the status of general rules, maybe unconsciously, as these rules can be deeply hidden within a classifier."); Richard Zemel et al., *Learning Fair Representations*, 28 PROC. MACHINE LEARNING RES. 325, 325 (2013), https://www.cs.toronto.edu/~toni/Papers/icml-final.pdf [https://perma.cc/W4YV-CMB3] ("Systems trained to make decisions based on historical data will naturally inherit the past biases."); Sullivan, *supra* note 11, at 397 (noting how data used by artificial intelligence may incorporate prior conscious or unconscious human discrimination in past performance reviews).

40. *See* Barocas & Selbst, *supra* note 8, at 682.

41. *See id.* at 684–87 ("If a sample includes a disproportionate representation of a particular class (more or less than its actual incidence in the overall population), the results of an analysis of that sample may skew in favor of or against the over- or underrepresented class."); Lehr & Ohm, *supra* note 6, at 703 ("[D]ata can be collected in a nonrepresentative manner . . . .").

3. Feature Selection

Third, machine-learning algorithms can unintentionally discriminate due to feature selection.[42] The features of an algorithm are the observed variables that the algorithm has access to when seeking out correlations and patterns. The data available to the algorithm can never fully capture all the unquantifiable complexities of the real world. The data set only captures what can be quantified and used in mathematical calculations. As an example, Barocas and Selbst point to the use of academic credentials in hiring.[43] An algorithm may assign a particular numerical weight to the academic reputation of the school attended by a candidate, but that choice of feature may systematically discriminate against certain groups. Students from poorer backgrounds, for example, may attend prestigious schools at lower rates for economic reasons, despite having similar academic competence.

4. Proxies

Finally, algorithms unintentionally discriminate by identifying proxies. Even where race, sex, or another protected characteristic is removed as a feature, making the algorithm theoretically unaware of that variable, the algorithm will often track that variable through other closely correlated proxies. For example, street address or neighborhood can often serve as a fairly reliable proxy for race.[44] Job tenure can serve as a proxy for gender because women who leave jobs for family care reasons will decrease average job tenure for all women.[45] Data scientists have consistently identified such "redundant encoding" as one reason that fairness through blinding is ultimately ineffective.[46] Removing race or sex as a variable for the machine does not prevent the machine from discovering and using correlations based on proxies for race or sex, resulting in unintentionally discriminatory output.

---

42. *See id.* at 688–90; Kroll et al., *supra* note 8, at 681 ("Of particular concern are choices about which data models should consider, a problem computer scientists call *feature selection*.")

43. *See* Barocas & Selbst, *supra* note 8, at 689.

44. *See, e.g.*, *id.* at 712 ("Due to housing segregation, neighborhood is a good proxy for race and can be used to redline candidates without reference to race."); DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 17, at 23 (noting the redlining problem, or "the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor"); JOSHUA R. LOFTUS ET AL., CAUSAL REASONING FOR ALGORITHMIC FAIRNESS 1–2 (2018), https://arxiv.org/abs/1805.05859 [https://perma.cc/HU6E-GP6N] (noting that "knowledge of someone's address makes it easy to predict their race with relatively high accuracy").

45. *See* Kroll et al., *supra* note 8, at 681 ("[W]omen who leave a job to have children lower the average job tenure for all women, causing this metric to be a known proxy for gender in hiring applications[].").

46. *See* DWORK ET AL., DECOUPLED CLASSIFIERS, *supra* note 17, at 1–2 ("Still speaking informally, it is known that 'ignoring' these attributes [such as gender and race] does not ensure fairness, both because they may be closely correlated with other features in the data and because they provide context for understanding the rest of the data, permitting a classifier to incorporate information about cultural differences between groups." (citation omitted)).

Professor Chander offers an illustration in the context of higher education admissions.[47] If a college admissions office wanted to choose applicants who were likely to succeed in the job market after graduation, it could use a machine-learning algorithm designed to predict job market success. If whites fare better than nonwhites in the job market, then the training data will reflect that fact and the algorithm will reproduce the racial discrepancy by giving higher scores to white applicants. Blinding the algorithm to race will not solve this problem. The algorithm will use proxies—such as the applicants' zip code or high school attended—to score applicants in a way that favors whites. The race variable becomes redundantly encoded in these proxies, still permitting the algorithm to discriminate on the basis of race in scoring applicants.

B. ALGORITHMIC FAIRNESS

Having described the problem of algorithmic bias, we turn to the technological cure currently under development: algorithmic fairness. The Reuters report on Amazon's experimental hiring algorithm concludes by noting that a new team has been assembled at Amazon to work on automated employment screening, "this time with a focus on diversity."[48] This work may include attempts to incorporate a fairness constraint into the algorithmic process. The machine-learning literature on fairness is quickly proliferating, with many creative ideas for adjusting or constraining machine-learning algorithms to counteract unintentional bias. This section introduces the leading approaches.[49]

1. Fairness Through Blindness

The most obvious, though disfavored, approach to machine-learning fairness is blinding the algorithm to any sensitive characteristics that are protected by anti-discrimination law, including race, color, religion, sex, or national origin;[50] age;[51] disability;[52] or genetic information.[53] To attempt this, data handlers simply omit these variables from the data features provided to the machine-learning algorithm.

At some level, this approach is intuitively appealing. It fits an idealized version of pure anticlassification decisionmaking. It fits Chief Justice Roberts's anticlassification rhetoric that "[t]he way to stop discrimination on the basis of race is to

47. *See* Chander, *supra* note 8, at 1038–39.

48. Dastin, *supra* note 9.

49. Professor Stephanie Bornstein refers to these technological approaches as attempts to "improve the algorithms," as opposed to attempts to "improve the law." Bornstein, *supra* note 11, at 533, 537.

50. *See* 42 U.S.C. § 2000e-2(a)(1)–(2) (2012).

51. *See* Age Discrimination in Employment Act of 1967 (ADEA), Pub. L. No. 90-202, § 4(a)(1)–(2), 81 Stat. 602, 603, 29 U.S.C. § 623(a)(1)–(2) (2012).

52. *See* Americans with Disabilities Act of 1990 (ADA), Pub. L. No. 101-336, § 102(a), 104 Stat. 327, 331–32, 42 U.S.C. § 12112(a) (2012).

53. *See* Genetic Information Nondiscrimination Act of 2008 (GINA), Pub. L. No. 110-233, § 202(a), 122 Stat. 881, 907, 42 U.S.C. § 2000ff-1(a) (2012). Notably, GINA does not provide a cause of action for disparate impact. *See* 42 U.S.C. § 2000ff-7(a).

stop discriminating on the basis of race."[54] And it fits with some salient success stories in human decisionmaking. A famous example, cited often by Justice Ginsburg in discussions about human biases, involves symphony orchestras dropping a curtain between the players auditioning for positions and the decisionmakers.[55] The curtain blinded decisionmakers to gender, resulting in more women successfully earning positions in symphony orchestras. A study of the impact of blind orchestra auditions concluded that they substantially increased the likelihood of orchestras selecting female candidates, explaining approximately one-third of the increase in female positions in major orchestras.[56]

Blinding algorithms to sensitive characteristics by removing variables like gender and race seems analogous to dropping a curtain between a symphony orchestra candidate and a human decisionmaker. Unfortunately, it doesn't work the same way. In the symphony context, humans need to be careful that no clues or potential proxies are allowed to reach the decisionmaker. In addition to a curtain, some symphonies use soft carpet flooring or ask candidates to remove their shoes to hide the sound of footsteps that might reveal the candidate's probable gender.[57] That works for symphony auditions. But, as Amazon's experiment illustrates, machines are much better at detecting patterns, and at identifying and leveraging on nonobvious proxies for omitted sensitive variables.[58]

---

54. Parents Involved in Cmty. Schools v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 748 (2007) (plurality opinion).

55. *See, e.g.*, Ruth Bader Ginsburg, Gillian Metzger & Abbe Gluck, *A Conversation with Justice Ruth Bader Ginsburg*, 25 COLUM. J. GENDER & L. 6, 18 (2013); *The Robert L. Levine Distinguished Lecture: A Conversation with Justice Ruth Bader Ginsburg and Professor Aaron Saiger*, 85 FORDHAM L. REV. 1497, 1508 (2017).

56. *See* Claudia Goldin & Cecilia Rouse, *Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians*, 90 AM. ECON. REV. 715, 716 (2000); Christine Jolls, *Is There a Glass Ceiling?*, 25 HARV. WOMEN'S L.J. 1, 3–5 (2002) (describing Goldin and Rouse's study).

57. *See* Ginsburg, Metzger & Gluck, *supra* note 55, at 18 ("When I mentioned the curtain last summer . . . a star violinist, Jennifer Koh, told me, 'You left out one thing. Not only is there a curtain but we audition shoeless so they don't hear a woman's heels coming on stage.'"); Goldin & Rouse, *supra* note 56, at 721; Jolls, *supra* note 56, at 3. Of course, there are many situations where human decisionmaking processes cannot be completely blinded to race or sex. When reviewing human decisionmaking that is not blinded, courts rely on a variety of burden-shifting proof frameworks to try to discern from the evidence available (often indirect, circumstantial evidence) whether an adverse employment decision was made "because of" the protected characteristic. *See, e.g.*, Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 336 (1977) (systemic disparate treatment); McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802–05 (1973) (individual disparate treatment); *see also* 42 U.S. C. § 2000e-2(a) (2012) (primary operative provisions); *id.* § 2000e-2(m) (mixed motives). Nonetheless, employers will often attempt to blind themselves to protected characteristics where possible to establish a defense to any potential claim of discrimination. *See generally Interviewing Candidates for Employment*, SOC'Y FOR HUM. RESOURCE MGMT, https://www.shrm.org/resourcesandtools/tools-and-samples/toolkits/pages/interviewingcandidatesforemployment.aspx ("Questions relating either directly or indirectly to age, sex, race, color, national origin, religion, genetics or disabilities should be avoided entirely.").

58. *See, e.g.*, Yona, *supra* note 18 ("[R]emoving the protected attribute alone will often not suffice. In most practical cases, the protected attribute is actually redundantly encoded in the rest of the observed features; most [machine-learning] algorithms are complicated computational machine[s] that will be able to pick up on and leverage this fact."). Removing any variable that might be a potential proxy is also

Machine-learning scholars appear to have reached a general consensus on this point: attempting to blind an algorithm to a sensitive characteristic is usually ineffective.[59] Sensitive features will indirectly affect the algorithm's result because of their relationship with other variables available to the algorithm. Computer scientists are trying to tell lawyers and judges that the way to stop machines from discriminating on the basis of race is *not* by trying to blind the algorithms to race, but by recognizing the influence of unspecifiable sources of bias and implementing race-aware fairness solutions.

Unlike "fairness through blindness," the other leading approaches to algorithmic fairness all take protected characteristics into account in an effort to achieve algorithmic fairness. The specific use of the sensitive characteristic varies, as discussed below, but each of the following fairness approaches is aware of (that is, the algorithm has access to information about) the sensitive variable and leverages that information to define and attain some form of fairness. In that sense, each of the approaches detailed in the remainder of this section might be described as a form of algorithmic affirmative action.[60]

---

problematic. *See* Kroll et al., *supra* note 8, at 681 ("Eliminating proxies can be difficult, because proxy variables often contain other useful information that an analyst wishes the model to consider . . . .").

59. *See* DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 44, at 22–23 (discussing the redundant encoding problem facing fairness through blindness techniques); DWORK ET AL., DECOUPLED CLASSIFIERS, *supra* note 17, at 1–2 ("[I]t is known that 'ignoring' these attributes does not ensure fairness . . . ."); MORITZ HARDT ET AL., EQUALITY OF OPPORTUNITY IN SUPERVISED LEARNING 1 (2016), https://arxiv.org/pdf/1610.02413 [https://perma.cc/63UY-WJ3K] ("A naïve approach might require that the algorithm should ignore all protected attributes . . . . However, this idea of 'fairness through unawareness' is ineffective due to the existence of *redundant encodings*, ways of predicting protected attributes from other features."); Toshihiro Kamishima et al., *Fairness-aware Learning Through Regularization Approach*, *in* IIEE COMPUT. SOC'Y, ICDMW '11: PROCEEDINGS OF THE 2011 IEEE 11TH INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS 643, 644 (2011), http://www.kamishima.net/archive/2011-ws-icdm_padm.pdf [https://perma.cc/Y5YG-RJRG] ("[S]imply removing sensitive features is insufficient, and affirmative actions have to be adopted to correct the unfairness in machine learning."); MATT KUSNER ET AL., COUNTERFACTUAL FAIRNESS 2 (2018), https://arxiv.org/pdf/1703.06856 [https://perma.cc/8SW8-A7EW] ("Despite its compelling simplicity, [fairness through unawareness] has a clear shortcoming as elements of [the observed attributes] can contain discriminatory information analogous to [the omitted protected characteristic] that may not be obvious at first."); LOFTUS ET AL., *supra* note 44, at 2 ("[I]t is . . . easy to accidentally create algorithms that make decisions without knowledge of a persons [sic] race or gender, but still exhibit a racial or gender bias."); Aditya Krishna Menon & Robert C. Williamson, *The Cost of Fairness in Binary Classification*, 81 PROC. MACHINE LEARNING RES. 107, 107 (2018), http://proceedings.mlr.press/v81/menon18a/menon18a.pdf [https://perma.cc/M3UK-4NZ6] ("[M]erely ignoring the sensitive feature is inadmissible, owing to it potentially being predictable by other features"); Zemel et al., *supra* note 39, at 325 ("These [biases] may be ameliorated by attempting to make the automated decision-maker blind to some [potential proxy] attributes. This however, is difficult, as many attributes may be correlated with the protected one."); Kroll et al., *supra* note 8, at 685 ("From a technical perspective, however, this approach is naive. Blindness to a sensitive attribute has long been recognized as an insufficient approach to making a process fair."); *see also* Barocas & Selbst, *supra* note 8, at 691 (noting the problem of redundant encodings even when "proscribed criteria have been removed from the dataset"); Chander, *supra* note 8, at 1041 ("The obvious remedy to the problem . . . would seem to be to mandate race or gender neutrality. In reality, however, . . . racial or sex neutrality would in fact perpetuate the problem of algorithmic replication of existing racism.").

60. Professor Chander may have meant to include even steps short of race-aware fairness constraints in his use of the term "algorithmic affirmative action." He writes: "Here, I mean 'affirmative action' in

2. Group Fairness

Group fairness is a category of fairness definitions that attempts to measure fairness by comparing the target variable outcomes of a machine-learning process between two groups sorted along the sensitive variable. For example, a group-fairness approach might be interested in the percentage of white candidates predicted by the algorithm to be good employees versus the percentage of black candidates predicted to be good employees.

The simplest example of a group-fairness approach is a "demographic parity" or "statistical parity" approach.[61] At its most restrictive, this would require that the predicted target variable success rates (good employee) be equal for both groups. The most significant problem with this approach is that it leads to "highly undesirable," inaccurate predictions of future success.[62] It forces the algorithm to predict success for white and black candidates (or male and female candidates, etc.) at exactly the same rate, regardless of whether that leads to accurate predictions of future success.[63]

A more relaxed version of demographic parity would allow some differences in predicted success rates between the groups, but within some defined boundaries. This resembles the Equal Employment Opportunity Commission's (EEOC) four-fifths rule of thumb for measuring potential disparate impact by comparing differences in the selection rates produced by a particular employment practice.[64] Under this version of parity, an algorithm might be required to produce a positive prediction (good employee) at a rate for minority candidates that is no less than four-fifths of the positive prediction rate for majority candidates. This approach

---

its broadest sense, as a set of proactive practices that recognize deficiencies in the equality of opportunity and act in a multiplicity of ways to seek to correct for those deficiencies." Chander, *supra* note 8, at 1040–41. As examples, Chander includes the removal of variables suspected of being proxies or of being tainted with encoded human bias. *See id*. at 1042–43; *see also* Yona, *supra* note 18 ("A quick fix which is often used in practice is to also remove all the other attribute[s] that are highly correlated (e. g [sic] above some threshold) with the protected attribute."). My use of the term "algorithmic affirmative action" is focused more narrowly on race-aware fairness constraints that arguably make classifications, or at least adjust calculations, on the basis of race.

61. *See* BERK ET AL., *supra* note 16, at 13; HARDT ET AL., *supra* note 59, at 1–2 (insisting that demographic parity means that "membership in a protected class should have no correlation with the decision"); Zemel et al., *supra* note 39, at 325 ("Group fairness, also known as statistical parity, ensures that the overall proportion of members in a protected group receiving positive (negative) classifications are identical to the proportion of the population as a whole.").

62. *See* BERK ET AL., *supra* note 16, at 13 (citing DWORK ET AL., FAIRNESS THROUGH AWARENESS *supra* note 44).

63. *See* DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 44, at 215 (arguing that statistical parity can lead to "blatantly unfair" results on an individual basis); HARDT ET AL., *supra* note 59, at 1–2 (positing that demographic parity does not ensure fairness and cripples utility of the algorithm); Zemel et al., *supra* note 39, at 325 ("While statistical parity is an important property, it may still lead to undesirable outcomes that are blatantly unfair to individuals, such as discriminating in employment while maintaining statistical parity among candidates interviewed by deliberately choosing unqualified members of the protected group to be interviewed in the expectation that they will fail.").

64. *See* 29 C.F.R. § 1607.4(D) (2018); *see also* MICHAEL FELDMAN ET AL., CERTIFYING AND REMOVING DISPARATE IMPACT 2–3 (2015), https://arxiv.org/pdf/1412.3756 [https://perma.cc/PG6V-JVLK] (proposing a measure of algorithm accuracy linked to the EEOC's disparate impact four-fifths rule, and a method for "repairing" data to remove disparate impact results).

could, for example, build something like the EEOC's four-fifths rule directly into the algorithm's optimization process.

Comparing selection rates for different groups is not the only possible way of measuring fairness at a group level. Fairness can also be measured by comparing false-positive rates (candidates who were predicted to be good employees but ultimately turned out not to be) or false-negative rates (candidates who were predicted to be "bad employees" but ultimately turned out not to be).[65] Another slightly more complicated alternative involves the following inquiry: looking only at those individuals for whom the algorithm predicted success (good employees), what is the probability that those individuals will actually achieve success? Are those probabilities the same or similar for each racial group?[66] Or, one could compare the *ratio* of false negatives to false positives for each group. Are these ratios the same or similar?[67] Still other possibilities for measuring group fairness exist, though they are either advocated less often[68] or satisfied only under unrealistic, idealized conditions.[69]

The point is not to advocate for any particular measure of group fairness, but to recognize that there are a host of ways to measure it, each of which requires awareness of the protected characteristic and using it to impose some constraint on the machine-learning algorithm. A related and important point is that these measures of fairness cannot all be optimized at once.[70] Setting demographic parity as a goal will mean that, in practice, false-positive or false-negative rates will not be equal. Requiring equality of false-negative rates will mean that, in practice, overall selection rates cannot be equal. There are inherent trade-offs in different measures of group fairness.[71]

Optimizing for any particular measure of group fairness also involves a trade-off with the algorithm's accuracy.[72] That is, applying a group fairness constraint

---

65. *See* BERK ET AL., *supra* note 16, at 13–14 (using the more formal terminology "conditional procedure accuracy equality" to describe this measure, and noting that Hardt uses the terms "equalized odds" and "equality of opportunity" to describe similar, specific measures).

66. *See id.* at 14 (referring to this measure as "conditional use accuracy equality," and noting that some machine-learning scholars refer to this measure as "calibration"); *see also* LOFTUS ET AL., *supra* note 44, at 2–3 (comparing the formal definitions of the "equalized odds," "calibration," and "demographic parity/disparate impact" measures of group fairness).

67. *See* BERK ET AL., *supra* note 16, at 14–15 (referring to this measure as "treatment equality").

68. *See id.* at 13 (referring to "overall accuracy equality" that measures the proportion of all predictions that were correct—true positive and negative predictions, compared to the total number of predictions—for each group, but noting that it is not often used because it treats false positives and false negatives as equally undesirable).

69. *See id.* at 15 (referring to "total fairness," which is satisfied only when all the foregoing measures of accuracy are simultaneously met, a condition that "cannot be achieved" in practice).

70. *See id.*; Yona, *supra* note 18 (noting the zero-sum trade-offs between various measures of fairness, and concluding that "it is ultimately up to the stakeholders to determine the tradeoffs").

71. *See* BERK ET AL., *supra* note 16, at 33–34 ("Except in stylized examples, there will be tradeoffs. These are mathematical facts subject to formal proofs. Denying that these tradeoffs exist is not a solution." (citations omitted)).

72. *See id.* at 3 ("[E]ven when attempts are made to clarify what fairness can mean, there are several different kinds that can conflict with one another and with accuracy." (citation omitted)); Menon & Williamson, *supra* note 59, at 107 ("There is typically an unavoidable tradeoff between *how accurate*

necessarily requires some reduction in the overall accuracy of the algorithm's predictions, sometimes called a "price of fairness."[73] The reason is that any fairness constraint imposed by a modeler will prevent the algorithm from simply maximizing accuracy based on all the features that would otherwise be available to the algorithm.

3. Individual Fairness

Recognizing the shortcomings of group-fairness measures alone, some machine-learning scholars advocate a fundamentally different approach called individual fairness. The critical difference is that, instead of focusing on comparisons at the group level, individual fairness approaches measure disparities in treatment at the individual level for individuals with similar features. The basic principle of individual fairness is: "any two individuals who are similar with respect to a particular task should be classified [on the target variable] similarly."[74]

Cynthia Dwork, a leading proponent of individual fairness metrics, is exploring ways to attain individual fairness, while also attempting to satisfy measures of group fairness in some circumstances.[75] Dwork suggests a metric to measure the "distance" between any two individuals calculated based on their observed nonsensitive features.[76] Where this distance is small enough (that is, the two individuals are similar enough across the nonsensitive variables), then the algorithm should predict the same target variable for the two individuals.[77]

As with the various definitions of group fairness, some challenges exist. First, individual fairness approaches may not solve the redundant-encoding (proxy)

---

our classifier is with respect to the target feature, and *how fair* it is with respect to the sensitive feature.").

73. *See, e.g.*, Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, *in* ASS'N FOR COMPUTING MACH., KDD '17: THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 797, 802 (2017), https://arxiv.org/pdf/1701.08230 [https://perma.cc/5FUZ-7TYS] (considering the public safety price of criminal recidivism models constrained for fairness).

74. DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 44, at 1 (emphasis omitted); *see also* KUSNER ET AL., *supra* note 59, at 2 (defining individual fairness as "[a]n algorithm is fair if it gives similar predictions to similar individuals").

75. *See* Zemel et al., *supra* note 39, at 325 (describing Dwork's line of work as "an ambitious framework which attempts to achieve both group and individual fairness"); Hartnett, *supra* note 18; Claire Cain Miller, *Algorithms and Bias: Q. and A. With Cynthia Dwork*, N.Y. TIMES (Aug. 10, 2015), https://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html.

76. *See* DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 44, at 1 ("[A] task-specific *similarity metric* describing the extent to which pairs of individuals should be regarded as similar for the classification task at hand."); FELDMAN ET AL., *supra* note 64, at 4; LOFTUS ET AL., *supra* note 44, at 3–4. Some observed, nonsensitive features that might be considered in the distance metric include years of education, years of relevant experience, college major, prior job performance metrics, or other similar variables.

77. *See* DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 44, at 1 (suggesting that the similarity metric imposed could be set by a "regulatory body" or proposed "by a civil rights organization"); LOFTUS ET AL., *supra* note 44, at 4 (quoting same). Coming up with an appropriate distance metric is a significant challenge for this individual fairness approach. *See* KUSNER ET AL., *supra* note 59, at 4070 (the distance metric "must be carefully chosen, requiring an understanding of the domain at hand beyond black-box statistical modeling"); Zemel et al., *supra* note 39, at 326.

problem. If the similarity distance metric includes consideration of nonsensitive features that correlate with the sensitive feature, then an algorithm could satisfy individual fairness despite producing group bias by proxies.[78] Second, although ensuring individual fairness may tend to produce group fairness in some situations, under other conditions, individual fairness will not produce group fairness, meaning that an individual fairness constraint can still produce disparate impacts along the sensitive variable.[79] Finally, as with any fairness constraint, there is necessarily some trade-off between overall prediction accuracy and a completely unconstrained algorithm. However, proponents of individual fairness argue that it actually improves accuracy *within all groups*.[80] That is, an algorithm optimized to achieve individual fairness and that has access to the gender variable should be better at predicting both whether a female candidate will be successful and whether a male candidate will be successful, as compared to an algorithm that is blinded to gender. The idea of improving intragroup accuracy by allowing the algorithm to use sensitive variables ("fairness through awareness") closely tracks the City of Springfield's hypothetical use of differential test score weights, discussed in the Article's Introduction.

4. Causal Modeling and Counterfactual Fairness

Recently, some machine-learning scholars have argued that no measure of group or individual fairness based solely on observational data will be satisfactory. These scholars propose fairness measures that are founded on causal modeling, which requires the analyst to have a theory about how the sensitive variable interacts with the target variable and with other observed variables.[81] Instead of solely relying on data, this approach incorporates a theory about how the world works.[82] For an example of causal modeling, see Figure 1.[83]

78. *See* LOFTUS ET AL., *supra* note 44, at 4 ("[M]any variables vary along with protected attributes such as race or gender, making it challenging to find a distance measure that will not allow some implicit discrimination.").

79. *See* DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 44, at 2 (describing the conditions under which individual fairness will produce group fairness); Hartnett, *supra* note 18 (asking Dwork, "What's a situation where individual fairness wouldn't be enough to ensure group fairness?," to which she replies, "If you have two groups that have very different characteristics. . . . If you have two groups that have very different performance on standardized tests, then you won't get group fairness if you have one threshold for the standardized-test score"). Dwork discusses a possibility for implementing "fair affirmative action," by which statistical parity could be forced while still preserving as much individual fairness as possible. *See* DWORK ET AL., FAIRNESS THROUGH AWARENESS, *supra* note 44, at 2, 11–15.

80. *See* DWORK ET AL., DECOUPLED CLASSIFIERS, *supra* note 17, at 2.

81. *See* LOFTUS ET AL., *supra* note 44, at 4 ("These [causal modeling] works depart from the previous approaches in that they are not wholly data-driven but require additional knowledge of the structure of the world, in the form of a causal model.").

82. These approaches are predicated on the work of computer scientist Judea Pearl in formally modeling the structure of causal relationships. *See generally* JUDEA PEARL, CAUSALITY: MODELS, REASONING, AND INFERENCE (2d ed. 2009).

83. Credit for the specific model presented in Figure 1 is due to Matt Kusner and coauthors Joshua Loftus, Chris Russell, and Ricardo Silva. This model is reproduced with their permission.

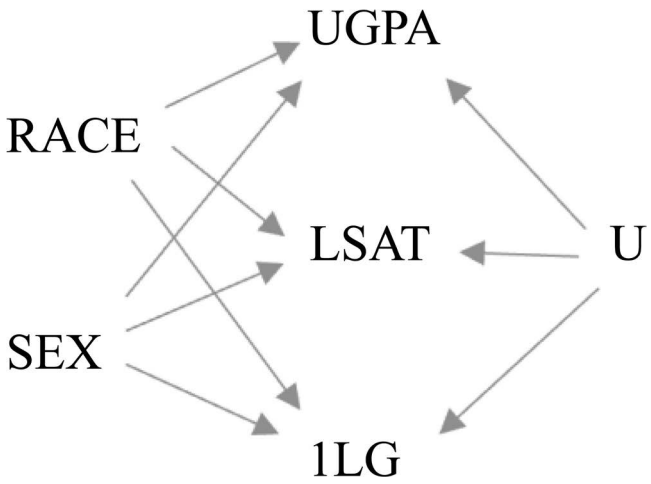

Figure 1

In Figure 1, directed arrows are used to show causal relationships between several observed variables (race, sex, undergraduate GPA (UGPA), LSAT score) and a target variable (first-year law school grades (1LG)). In addition, a variable, *U*, represents all other latent, unobserved characteristics of an individual (which could include, for example, general knowledge of legal topics) that are assumed not to be caused by any of the observed variables in the model. A directed arrow indicates that a variable has some causal influence on another, with causation flowing in the direction of the arrow. In this model, for example, the directed arrows indicate that race has some causal influence on the values for the variables undergraduate GPA and LSAT Score, and also on the target variable, first-year grades.

Of course, there is no guarantee that this model is correct; it is just one possible model of the causal relationships among the relevant variables. But having some model of how the observed variables interact provides additional information that can then be leveraged by analysts when seeking to measure algorithmic fairness. In fact, machine-learning scholars advancing causal approaches to algorithmic fairness argue that other measures of fairness include covariate adjustments that "are often based on implicit causal reasoning"[84]—in other words, that other fairness techniques are driven by unstated, unexamined models of causal interactions between the relevant variables.[85] Advocates of causal approaches contend that it is better to explicitly state the causal model assumptions and then design fairness solutions accordingly.[86]

84. LOFTUS ET AL., *supra* note 44, at 8.
85. *See id.*
86. *See id.*

A prominent example of a causal approach to algorithmic fairness is "counterfactual fairness," which posits that a predictor is fair toward an individual if it gives the same prediction on the target variable in both the actual world and in the counterfactual world where the individual's sensitive group membership (for example, race) is switched.[87] For the model shown in Figure 1, an algorithmic classifier might predict an individual's first-year grades (and then admit or deny admission accordingly) based only on the two features: undergraduate GPA and LSAT. To determine whether this simple algorithmic classifier is counterfactually fair, the analyst would use all the known, observed variables, including the sensitive variables (race, sex, UGPA, and LSAT score), combine all of that data with the causal model depicted in Figure 1, and use that data to compute the implied value of the unobserved variable *U*. Recall that in the causal model, *U* simply represents all the latent, unobserved characteristics of an individual that influence an individual's undergraduate GPA, LSAT score, and first-year grades. In the model, an individual's *U* should remain the same regardless of race or sex. Next, the analyst would switch the race variable (for example, from white to black, its counterfactual value). Finally, the analyst would use the causal model, the now-computed value for the individual's *U*, and the new counterfactual race variable to recompute all the remaining observed variable values (UGPA and LSAT). If the classifier would reach the same prediction for first-year grades based on the new, recomputed values for undergraduate GPA and LSAT score, then it is counterfactually fair by this definition.[88] This particular use of causal modeling operates on an individual level,[89] although there is a somewhat similar approach that operates on a group level.[90]

87. *See* KUSNER ET AL., *supra* note 59, at 2 ("[W]e provide the formal definition of counterfactual fairness, which enforces that a distribution over possible predictions for an individual should remain unchanged in a world where an individual's protected attributes had been different in a causal sense.").

88. *See id.* at 8–9 (applying this definition of counterfactual fairness to a law school success problem using the model shown in Figure 1). In implementation, the authors propose allowing slight variation in the counterfactual predictions, up to a defined amount. *See* Matt J. Kusner, *Counterfactual Fairness: Matt Kusner, The Alan Turing Institute*, YOUTUBE (Feb. 28, 2018), https://www.youtube.com/watch?v=ZfuOw02U7hs (from 24:56–26:07 of the video). A further variation on this notion of counterfactual fairness is to try to model and isolate "path specific" causal routes from the sensitive variable (for example, race) to the target variable (here, predicted 1L GPA) and apply corrective measures only along identified causal pathways that are deemed by the modeler be "unfair," while not applying corrective measures to any "fair" causal pathways from the sensitive variable to the target. *See* SILVIA CHIAPPA & THOMAS P.S. GILLAM, PATH-SPECIFIC COUNTERFACTUAL FAIRNESS 1, 2 (2018), https://arxiv.org/pdf/1802.08139 [https://perma.cc/MQM5-NM4V]; LOFTUS ET AL., *supra* note 44, at 15–18 (discussing path-specific variations).

89. *See* KUSNER ET AL., *supra* note 59, at 3 ("We also emphasize that counterfactual fairness is an individual-level definition."). Indeed, individual fairness through use of a distance metric as proposed by Dwork is similar to matching up pairs of individuals who are quite similar except that they are differ on the protected trait. Creating matched pairs is one way to get a sense of what to expect in the counterfactual world (what one might expect from switching one person's protected trait to its opposite).

90. *See* NIKI KILBERTUS ET AL., AVOIDING DISCRIMINATION THROUGH CAUSAL REASONING 656, 657 (2018), https://arxiv.org/pdf/1706.02744 [https://perma.cc/G2PU-8EW3] ("[O]ur work proposes a shift from trying to find a single statistical fairness criterion to arguing about properties of the data and which assumptions about the generating process are justified. Causality provides a flexible framework for

The primary drawback to the counterfactual fairness approach is the need to posit a causal model depicting a theory about how all the variables are interrelated.[91] Advocates of causal reasoning approaches acknowledge this, but maintain that causal modeling is critical for algorithm fairness.[92] For our purposes, it is not necessary to determine whether causal fairness approaches are necessarily superior to purely observational approaches. We need only note that, like the group-fairness and individual-fairness approaches, causal approaches require awareness of the sensitive characteristic to operate. Causal approaches will also impose an accuracy price because requiring the algorithm to satisfy the counterfactual fairness constraint necessarily makes its predictions less accurate than an unconstrained algorithm.[93]

5. Implementing Fairness Solutions

As the foregoing shows, machine-learning scholars are working on interesting, complex, and creative ways to approach the fundamental problem of fairness in machine-learning algorithms. The leading proposals all require some sort of *protected-characteristic awareness* adjustment at some point in the machine-learning stages to constrain or alter the output the algorithm otherwise would have generated had it been completely unconstrained.

The means by which these fairness solutions are actually implemented in the process varies. Possibilities include: preprocessing the training data (that is, changing some of the data, removing some variables that correlate with the sensitive variable, or changing some predicted target values so that a fairness constraint is satisfied);[94] imposing some kind of penalty for unfairness as the machine learns patterns and optimizes its classification algorithm;[95] or allowing

organizing such assumptions."). Kusner describes the difference as follows: Kusner's approach compares the same individual with a different, counterfactual version of him/herself, while Kilbertus's approach compares different individuals with those who have the same or similar observed features. *See* Kusner, *supra* note 88, at 17:28.

91. *See* KUSNER ET AL., *supra* note 59, at 7 ("[C]ausal models always require strong assumptions, even more so when making counterfactual claims.").

92. *See* LOFTUS ET AL., *supra* note 44, at 7–8 ("[W]hen choosing policies and designing systems that will impact people, we should minimise or eliminate the causal dependence on factors outside an individual's control, such as their perceived race or where they were born. Since such factors have influences on other aspects of peoples' lives that may also be considered relevant for determining what is fair, applying this intuitive notion of fairness requires careful causal modelling as we describe here.").

93. *See* Kusner, *supra* note 88, at 22:48.

94. *See generally* BERK ET AL., *supra* note 16, at 25–29 (discussing preprocessing, inprocessing, and postprocessing as potential solutions for group fairness issues); Zemel et al., *supra* note 39, at 325 ("Several papers . . . aim to achieve the first goal, group fairness, by adapting standard learning approaches in novel ways, primarily through a form of fairness regularizer, or by re-labeling the training data to achieve statistical parity."); *id*. at 328–29 (giving an example of "massaging" the training data labels by changing the values of some labels to remove the discrimination with the least possible changes to the data).

95. *See, e.g*., Zemel et al., *supra* note 39, at 328–29 (discussing use of a regularizer during optimization); Kusner, *supra* note 88, at 30:07.

the machine to access the sensitive attributes while it is learns from the training data, but later hiding them from the machine when it is makes new predictions.[96]

For purposes of antidiscrimination law, the method of implementation may prove important. The time at which the fairness solution and the sensitive variable are injected into the process could be relevant under current affirmative action doctrine. And if courts are concerned about the optics of the process and potential balkanization effects, then the transparency of the fairness solution and the ability to identify the individuals who are advantaged or disadvantaged by the fairness constraint could be important.

II. THE STATUTORY CHALLENGE TO ALGORITHMIC AFFIRMATIVE ACTION

This Part turns from the design of algorithmic affirmative action to its legality. Are race-aware algorithmic fairness solutions permissible under U.S. antidiscrimination law? In this Part, I will rely on the following hypothetical situation: an employer covered by Title VII has delegated its hiring decisions to a machine-learning algorithm that is optimized to select good employees, but does so subject to one of the race-aware fairness constraints described in Part I (either group fairness, individual fairness, or counterfactual fairness). The employer, during the algorithm design process, has chosen to add the race-aware fairness constraint out of concern that an unconstrained algorithm would produce racial disparities disfavoring minorities. The employer's concern is rooted in the literature on machine learning, as well as the employer's own initial testing of its algorithm design on sample data. After the employer makes hiring decisions using the fairness-constrained algorithm, an unsuccessful candidate sues the employer for disparate treatment discrimination on the basis of race.

The starting point for our statutory analysis is the primary operative text for disparate treatment discrimination, section 703(a):

> (a) Employer practices
>
> 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

96. For criticism of this approach, sometimes called "disparate learning processes" (DLPs), see Lipton et al., *supra* note 21 (manuscript at 2). Lipton argues that DLP approaches cannot solve the redundant encoding problem, and that they can generate discrimination within groups that is based on irrelevant features because the DLP attempts to implement treatment disparity using proxies. *See id.* (manuscript at 2–3). Lipton concludes that a better trade-off between accuracy and fairness (parity) is accomplished by transparently using the sensitive variable, rather than attempting to hide it from the machine via a DLP. *See id.* (manuscript at 3).

> an employee, because of such individual's race, color, religion, sex, or national origin.[97]

The following sections will consider whether our employer's use of a race-aware, fairness constraint in a hiring algorithm violates this statutory prohibition on disparate treatment.

A. THE PRIMA FACIE CASE: IS ALGORITHMIC AFFIRMATIVE ACTION AN ADVERSE ACTION "BECAUSE OF" A PROTECTED CHARACTERISTIC?

First, our employer's use of a race-aware, algorithmic fairness constraint likely constitutes at least a prima facie instance of disparate treatment discrimination in violation of section 703(a) of Title VII. Our employer uses a machine-learning algorithm as part of its hiring selection process, and that algorithm includes a race-aware fairness constraint. An unsuccessful candidate will allege that by using the race-aware fairness constraint the employer has "fail[ed] or refuse[d] to hire" her "because of" her race, or that the employer "classif[ied]" her in a way that deprived her of employment opportunity, in violation of section 703(a).[98]

1. Discriminatory Motive or Intent

In cases involving human decisions the Supreme Court has sharply divided Title VII violations into two categories, disparate treatment and disparate impact.[99] The Court emphasizes that, unlike disparate impact, a disparate treatment violation requires "'intent' or 'motive' (the terms are often used interchangeably)."[100] This emphasis on "intent" or "motive" presents some interpretive challenges when moving from human decisions to machine decisions. Professor Charles Sullivan imagines a machine that runs amok by making its outcome predictions about who would be a good employee solely on the basis of race or gender because (in his illustration) those features happen to correlate with the target variable.[101] Even in this scenario, it is not clear that an employer's reliance on the machine's output would necessarily trigger disparate treatment liability under the Court's Title VII precedents.[102] How can a computer act with

97. 42 U.S.C. § 2000e-2(a) (2012). In its mixed motive provisions, Title VII further provides that an unlawful employment practice is established if plaintiff demonstrates that a protected characteristic "was a motivating factor for any employment practice, even though other factors also motivated the practice." *Id.* § 2000e-2(m). A separate provision limits the plaintiff's remedies in a mixed motives case if the employer can demonstrate that it "would have taken the same action in the absence of the impermissible motivating factor." *Id.* § 2000e-5(g)(2)(B). This lower mixed motive standard is unavailable in age discrimination cases under the ADEA, where plaintiffs must show that age was a "but-for" cause of the adverse action. *See* Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 177–78 (2009).

98. 42 U.S.C. § 2000e-2(a).

99. *See* Sullivan, *supra* note 11, at 405.

100. *Id.* at 404–05 (citation omitted).

101. *Id.* at 402–10 (considering artificial intelligence that is "going rogue" by classifying based on protected characteristics).

102. *See id.* at 405–06. Sullivan suggests that a more careful reading of the statute, apart from the Supreme Court's "bifurcated structure" (disparate treatment versus disparate impact) in the human decision context, should indicate liability under these circumstances for "failing or refusing to hire" or for "classifying them in a way that would deprive them of employment opportunities." *Id.* at 408.

discriminatory intent or motive when it only identifies correlations with a target variable and acts upon them? How can an employer have discriminatory intent or motive when it relies entirely on the computer's data-driven predictions?

With algorithmic affirmative action, however, there is an identifiable human component to the employer's practice that is lacking in Professor Sullivan's hypothetical. Professor Sullivan proposed turning the hiring decisions over to an algorithm with the sole task of identifying good employees.[103] The algorithmic affirmative action hypothetical I set forth above involves a different, more complex instruction. By definition, incorporation of any race-aware, fairness solution means instructing the computer to pick good employees, subject to this defined fairness constraint. If an employer or its algorithm designer—a human, not a machine—instructs the computer to follow a race-aware fairness constraint, then it is presumably acting with sufficient intent to trigger disparate treatment protections, even under the Court's human decision precedents. The human instruction to follow a race-aware fairness constraint intentionally injects a protected characteristic into the computer's programming in a way that simply asking the computer to find good employees—even with access to sensitive variables—does not.

2. *Ricci v. DeStefano* and Discriminatory Intent

The Supreme Court's decision in *Ricci v. DeStefano*[104] bolsters the conclusion that courts would find sufficient discriminatory intent to support a disparate treatment claim where an employer directed the use of a race-aware fairness constraint. *Ricci*'s analysis of the tension between disparate treatment and disparate impact in the context of human decisions makes it a critical case for scholars who are considering how to respond to algorithmic disparate impacts.[105]

In *Ricci*, the City of New Haven, Connecticut's fire department administered objective examinations (part written, part oral) to identify the best candidates for promotion to lieutenant and captain.[106] The tests would determine who would be eligible for those promotions for two years, and many candidates "studied for months, at considerable personal and financial cost."[107] The tests were developed by a third-party contractor that took steps to ensure that the test results would not inadvertently discriminate against minority candidates. For example, in conducting interviews to prepare a job analysis, the contractor oversampled minority firefighters.[108] The contractor also ensured that, for the oral portion of the

Barocas and Selbst explore some possible arguments for application of disparate treatment doctrine to discriminatory output of data mining but conclude that "disparate treatment doctrine does not appear to do much to regulate discriminatory data mining." *See* Barocas & Selbst, *supra* note 8, at 699–701.

103. *See* Sullivan, *supra* note 11, at 395 ("It gives the computer only one instruction: 'Pick good employees.'").

104. 557 U.S. 557 (2009).

105. *See supra* note 23.

106. *Ricci*, 557 U.S. at 562, 564.

107. *Id*. at 562.

108. *See id*. at 565.

examination, each three-member panel of assessors included two minority members.[109]

After the city obtained the test results, it grew "concern[ed] that the tests had discriminated against minority candidates."[110] Most of the firefighters eligible for immediate promotion based on the test results were white.[111] After a protracted public debate about whether to certify the test results and proceed with the promotion procedure, the city's civil service board ultimately refused to certify the results of the tests.[112]

Several of the firefighters—who passed the examinations and might have received promotions had the tests been certified—sued the city alleged disparate treatment in violation of section 703(a) of Title VII and in violation of the Equal Protection Clause of the Fourteenth Amendment.[113] The city argued that its attempt to comply with Title VII's disparate impact provisions could not be the basis for finding a disparate treatment violation. The District Court agreed, granting summary judgment in favor of the city, in part because the city's "motivation to avoid making promotions based on a test with a racially disparate impact . . . does not, as a matter of law, constitute discriminatory intent" necessary to support a disparate treatment violation of Title VII.[114] The Court of Appeals agreed.[115] The Supreme Court not only reversed the summary judgment in favor of the city, but it also went on to direct summary judgment in favor of plaintiffs on their Title VII claim.[116] It held that the city's refusal to certify the tests constituted disparate treatment in violation of Title VII because the city lacked a "strong basis in evidence to believe it would face disparate impact liability if it had certified the examination results."[117]

In reaching its conclusion, the Supreme Court considered whether the city had sufficient discriminatory intent to support a section 703(a) disparate treatment violation when the city's motivation was not to discriminate against minorities, but instead to avoid disparate impact liability under Title VII.[118] The Court had no problem finding that the city's refusal to certify the test results constituted sufficient intent to establish a section 703(a) disparate treatment violation given its knowledge of the racial disparities that the test had produced.[119] It wrote:

---

109. *See id*. at 565–66.

110. *Id*. at 566.

111. The top ten scores on the lieutenant test were eligible for immediate promotion to one of eight open lieutenant positions. All ten were white. The top nine scores on the captain test were eligible for immediate promotion to one of seven open captain positions. Of those nine, seven were white and two were Hispanic. *See id*. at 566.

112. The civil service board vote was two to two, with one member recused. As a result, the board did not certify the tests. *See id*. at 574.

113. *See id*. at 574–75.

114. *See id*. at 576 (quoting Ricci v. DeStefano, 554 F. Supp. 2d 142, 160 (D. Conn. 2006)).

115. *See id*. at 576 (citing Ricci v. DeStefano, 264 F. App'x. 106 (2d Cir. 2008) (per curiam), *opinion withdrawn and superseded by* 530 F.3d 87 (2d Cir. 2008)).

116. *See id*. at 593. The Court declined to resolve the equal protection question. *See id*.

117. *Id*. at 592.

118. *See id*. at 579–80.

119. *See id*.

> But both of those statements [about intent by the District Court and by the United States government by amicus] turn upon the City's objective—avoiding disparate-impact liability—while ignoring the City's conduct in the name of reaching that objective. Whatever the City's ultimate aim—however well intentioned or benevolent it might have seemed—the City made its employment decision because of race. The City rejected the test results solely because the higher scoring candidates were white. The question is not whether that conduct was discriminatory but whether the City had a lawful justification for its race-based action.[120]

The Court's Title VII analysis therefore started with the premise that the city's action's violated section 703(a)'s prohibition on disparate treatment because, absent some legal justification or defense, "this express, race-based decisionmaking violates Title VII's command that employers cannot take adverse employment actions because of an individual's race."[121]

Applying *Ricci*'s logic to our algorithmic affirmative action hypothetical, the necessary intent for a section 703(a) disparate treatment claim exists. Our employer's decision to use a race-aware fairness constraint has the benevolent objective of avoiding algorithmic results that produce a disparate impact on racial minorities. Yet, the employer's conduct in deploying a race-aware fairness constraint would likely be viewed as "race-based action," like the city's refusal to certify test results, because it injects consideration of race into the hiring decision.

One might try to distinguish *Ricci* by arguing that the employer's action there was based solely on race, while the decision to hire or reject a candidate based on the output of a race-aware algorithm would not be *solely* because of race.[122] Race would be only one of many variables considered by our employer's algorithm. But, Title VII's causation standard does not require that race be the *sole* reason for an action; it requires only either "but-for" causation or "motivating factor" causation.[123] As discussed in the following section, race could be shown to be a but-for cause of some candidate rejections, which is enough to sustain a disparate treatment violation under any of the major antidiscrimination statutes.[124]

3. "But-For" Causation

If a human instruction to introduce a race-aware fairness constraint does constitute a sufficient motive or intent for disparate treatment purposes, then the next step of the analysis is to determine whether there is a sufficient causal link between the protected characteristic and the adverse employment action. For

120. *Id.*

121. *Id.* at 579.

122. *See supra* text accompanying note 118.

123. *See supra* note 97; *see also* Andrew Verstein, *The Jurisprudence of Mixed Motives*, 127 YALE L. J. 1106, 1137–41 (2018) (distinguishing a but-for motive standard from a "sole" motive standard).

124. For Title VII violations, a plaintiff need only show that race was one motivating factor among several possible motivating factors for the adverse employment action to trigger liability. *See supra* note 97. In contrast, the ADEA requires a showing of but-for causation. *See supra* note 97.

Title VII, the prohibited characteristic need only be a motivating factor to trigger liability, but, under the ADEA, age must be a but-for cause of the adverse action.[125] Even the more demanding but-for causation requirement will not thwart a statutory challenge to algorithmic affirmative action.

To show but-for causation, a plaintiff alleging discrimination needs to demonstrate that she would not have suffered the adverse employment action but for the employer's consideration of the protected characteristic.[126] In our hypothetical, the adverse action is a failure to hire, based on the results of a machine-learning algorithm that includes a race-aware fairness constraint. A plaintiff could establish but-for causation by showing that she would have been hired had it not been for consideration of her race.

This required showing of but-for causation might prove challenging for real-world litigants, but in theory it should not be difficult at all. If a disappointed candidate in our example can get access (presumably through civil litigation discovery) to the employer's algorithm program, the underlying training data, and the data upon which the algorithm based its predictions, the candidate would have all the tools necessary to determine whether the race-aware fairness constraint was decisive. The race-aware fairness constraint could be temporarily removed from the computer's optimization program. Programmers can delete the fairness constraint instructions and leave the program with only one optimization instruction: "pick good employees."[127] Then the results for any individual candidate could be directly compared, with and without the fairness constraint. A plaintiff might have been classified as "bad employee—don't hire" using an algorithm with the race-aware fairness constraint, but classified as "good employee—hire" using the same algorithm without the fairness constraint. If so, the plaintiff would have an unusually strong case that the race-aware fairness constraint was a but-for cause of the adverse employment action.

In practice, information asymmetries typical for failure-to-hire or failure-to-promote cases are at work.[128] Disappointed candidates will not have much detailed information about the reasons they were not selected. They will not have access to information about the hiring algorithm source code, the training data, or the applicant pool data—all of which are in the possession of the employer. They likely will not be able to obtain this information prior to civil litigation discovery.

---

125. *See supra* note 97.

126. *See* Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 346–47 (2013) (citing multiple authorities on tort law for the proposition that "[i]n the usual course, this standard [causation in fact] requires the plaintiff to show 'that the harm would not have occurred' in the absence of—that is, but for—the defendant's conduct"); Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 176–77 (2009).

127. This could be done by "commenting out" the part of the computer code that implements the fairness constraint, which would convert that portion of code from an instruction to a comment that the computer ignores. Skype Interview with Matt Kusner, Associate Professor, Univ. Coll. London (Oct. 16, 2018) (on file with author).

128. *See* Suja A. Thomas, *Oddball* Iqbal *and* Twombly *and Employment Discrimination*, 2011 U. ILL. L. REV. 215, 222 nn.37–38 (discussing information asymmetries in employment discrimination cases, where plaintiff typically "lacks information in the defendant's possession that may be vital to the proof of his case").

Nonetheless, if a plaintiff can survive a motion to dismiss and obtain discovery about the hiring process, a strong showing of but-for causation appears quite possible.

The but-for causation standard, much less the more lenient motivating factor standard, is unlikely to stand in the way of a prima facie showing of disparate treatment in our hypothetical. A court's analysis of algorithmic affirmative action, therefore, will likely begin with the same starting premise as the *Ricci* Court's analysis: the employer's race-based conduct, however benevolent its objective, will violate section 703(a) absent some legal justification or defense for it.[129]

B. THE "STRONG-BASIS-IN-EVIDENCE" JUSTIFICATION

One possible legal justification for the employer's conduct is the one announced in *Ricci*: the employer's strong basis in evidence for believing that it would be subject to disparate impact liability in the absence of the race-based conduct.[130] The Court adopted this standard in the context of Title VII to resolve any tensions between the statutory disparate impact framework and the disparate treatment framework.[131]

An employer using a race-aware fairness constraint in a machine-learning algorithm would likely argue that the fairness constraint is necessary to prevent the algorithm from producing a disparate impact violation of Title VII. At least one proposed variant of a race-aware group fairness constraint directly incorporates the mathematics of the EEOC's four-fifths rule to determine whether a disparity exists in the data accessible to an algorithm, and then recommends "repairing" the "biased" data in a way that removes that disparity.[132]

One difficulty with this justification is that *Ricci* requires employers to consider not only the prima facie showing of statistical disparities, but also the employer's potential defenses to a disparate impact claim. In *Ricci*, the statistical disparities were significant, and the city was therefore "compelled to take a hard look at the examinations to determine whether certifying the results would have had an impermissible disparate impact."[133] But the city was obligated to consider its defenses, too. The city would be liable only if the tests "were not job related and consistent with business necessity, or if there existed an equally valid, less-discriminatory alternative that served the City's needs but that the City refused to

129. *See supra* note 119 and accompanying text.

130. Ricci v. DeStefano, 557 U.S. 557, 584 (2009). The Court borrowed this standard from its equal protection jurisprudence in cases involving race-based governmental action to remedy past discrimination. *See id.* at 582–84.

131. *See id.* at 584.

132. *See* FELDMAN ET AL., *supra* note 64. The "repair" procedure is a variant of data preprocessing and looks somewhat like the type of norming that would be a violation of section 703(l) if performed on employment-related tests scores. The modeler would change or remove any attributes in the data set that could be used to predict whether a person was in the minority group. *See id.* at 6, 11, 13 (describing and illustrating the proposed "repair" procedures).

133. *Ricci*, 557 U.S. at 587.

adopt."[134] The Court found no genuine issue of material fact on these points; the city lacked a strong basis in evidence for believing that the tests were not job-related or that there was a less discriminatory alternative.[135]

In our hypothetical, an employer has adopted a race-aware fairness constraint in order to prevent expected disparate impact from an unconstrained algorithm. But, even if we assume that statistically significant disparities would be produced by the unconstrained algorithm, *Ricci* requires an employer to consider its business necessity defense: is the unconstrained algorithm job-related and consistent with business necessity? We are now back to the question, addressed in some of the legal literature on algorithms, whether existing disparate impact doctrine can effectively police machine-learning algorithms that produce discriminatory results. Many scholars think not, believing that a machine-learning algorithm designed to identify correlations with a target variable like "good employee" is, maybe by definition, job-related and consistent with business necessity.[136] This possibility prompted Professors Grimmelmann and Westreich to propose a modification to (or perhaps a clarification of) the business necessity defense to require that the employer provide an understandable explanation for the machine's decision.[137] Although this issue is far from resolved in the courts, there is at least a viable argument that an algorithm designed to uncover nonobvious variables correlating with being a good employee is job-related and consistent with business necessity. If so, then our hypothetical employer, like the City of New Haven in *Ricci*, has a viable defense to disparate impact liability, thereby precluding it from engaging in race-based conduct to avoid statistical disparities.

The strong-basis-in-evidence test might also be met, thus permitting the employer to implement the race-aware fairness constraint, if a disappointed candidate under an unconstrained algorithm could point to a less discriminatory alternative practice that would still meet the employer's needs, but that the employer

134. *Id*. This proof framework for disparate impact was set out in the 1991 amendments to Title VII and is codified at 42 U.S.C. § 2000e-2(k).

135. *See id*. at 587–92.

136. *See supra* note 12.

137. *See* Grimmelmann & Westreich, *supra* note 8, at 171–74; *see also* Barocas & Selbst, *supra* note 8, at 708–09 ("Data mining will likely only be used if it is actually predictive of *something*, so the business necessity defense solely comes down to whether the trait sought is important enough to job performance to justify its use in any context."); Sullivan, *supra* note 11, at 411 ("There is yet another reason why disparate impact will not resolve the problem of Rogue Arti: as hypothesized, the gender exclusion seems to be justified by business necessity and therefore is not illegal."). Our employer may be lacking a strong basis in evidence for believing there is a disparate impact violation for a second reason—the algorithm may be classifying on the basis of a protected characteristic to make its predictions, and therefore may be "facially discriminatory"—and subject only to disparate treatment theory—rather than "facially neutral"—the domain of disparate impact theory. *See id*. at 410–11. *But see* Allan G. King & Marko J. Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555, 571 (2016) (arguing that a simple correlation with a target variable is not sufficient to satisfy disparate impact's job-relatedness requirement, and if it were, then the defense would reduce to a tautology).

refused to adopt.[138] The simplest alternative employment practice to an unconstrained algorithm—attempting to blind the algorithm to the sensitive variable—will often produce an outcome that is just as discriminatory as the unconstrained algorithm, as explained above.[139] And, one problem with race-aware fairness alternatives to an unconstrained algorithm is that they will necessarily involve some trade-off with accuracy. Measured by overall accuracy of predictions, an algorithm instructed only to identify good employees will beat an algorithm instructed to identify good employees subject to a fairness constraint, because of the necessary trade-off with accuracy.[140] Arguably, a less accurate predictor is not "equally valid" and may not serve the employer's need to identify good employees as well as the unconstrained predictor.[141]

Because of the likely business necessity defense to a disparate impact claim, our hypothetical employer is probably in the same boat as the City of New Haven in *Ricci*. It likely lacks a strong basis in evidence for believing that it would face disparate impact liability if it refused to implement its race-aware fairness constraint. On this reading of the business necessity defense, our search for a lawful justification will need to continue.

C. *RICCI*'S "AMPLE ROOM" FOR VOLUNTARY COMPLIANCE WITH TITLE VII

Even if algorithmic affirmative action might not be justified under the strong-basis-in-evidence test, it might be justified under other language found in *Ricci*'s dicta that protects certain acts of voluntary Title VII compliance. The *Ricci* Court emphasized the importance of voluntary compliance as integral to Title VII's statutory scheme and clarified that its ruling left "ample room" for employers' voluntary compliance efforts.[142] In an enigmatic portion of the *Ricci* opinion potentially relevant to algorithmic fairness, the Court wrote:

> Nor do we question an employer's affirmative efforts to ensure that all groups have a fair opportunity to apply for promotions and to participate in the process by which promotions will be made. But once that process has been established and employers have made clear their selection criteria, they may not then invalidate the test results, thus upsetting an employee's legitimate expectation not to be judged on the basis of race. . . .

138. *See* 42 U.S.C. § 2000e-2(k)(1)(A), (C) (2012) (returning the law on "alternative employment practice[s]" in disparate impact cases to its state before *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989)); *Ricci*, 557 U.S. at 589–92 (finding no evidence of an available "equally valid, less-discriminatory" alternative employment practice).

139. *See supra* note 59 and accompanying text.

140. *See supra* note 72 and accompanying text.

141. *Cf. Ricci*, 557 U.S. at 587, 589 (providing no evidence that changing the test weighting of written and oral components from 60/40 to 30/70 would be "an equally valid way to determine whether candidates possess the proper mix of job knowledge and situational skills").

142. *Id.* at 583 ("The [strong-basis-in-evidence] standard leaves ample room for employers' voluntary compliance efforts, which are essential to the statutory scheme and to Congress' efforts to eradicate workplace discrimination.").

> 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. And when, during the test-design stage, an employer invites comments to ensure the test is fair, that process can provide a common ground for open discussions toward that end.[143]

Professor Pauline Kim rightly notes that this language would allow an employer who has detected bias in its algorithm to voluntarily cease using it.[144] The City of New Haven was required to certify the test results once firefighters had relied on the process, but there was nothing requiring the city to continue using that test format in making promotions in the future.[145]

This observation is true as far as it goes, but it doesn't go far enough to legally justify race-aware fairness constraints. There is a significant difference between *discarding* a biased algorithm and *fixing* a biased algorithm by introducing a race-aware fairness constraint. That difference is the employer's affirmative use of race to calculate mathematical fairness constraints within an algorithm used to make hiring decisions, which will necessarily result in adverse employment actions.

Critiquing Barocas and Selbst's observation that *Ricci* might present a roadblock to employer responses to biased algorithms, Professor Kim writes:

> 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." Unlike . . . *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.[146]

She goes on to observe that "[m]erely being aware of the racial consequences of a selection process" is not a disparate treatment violation.[147] Therefore, an employer's attempt to identify, understand, and avoid any racial disparate impact produced by a selection process is not a violation. In *Ricci*, for example, if during

143. *Id*. at 585.

144. Kim, *supra* note 10, at 931–32.

145. *Id*. at 932. There is an interesting ambiguity lurking in this dichotomy. Of course, the City of New Haven was not required to certify the test results and then to continue using those scores for promotions in perpetuity. On the other hand, it seems fair to infer that the city might not be able to discard the results after filling only the immediately open positions. The record indicates that the test results would be used in making promotion decisions for the next two years, and the candidates no doubt relied on that understanding when studying for the high-stakes test. *See Ricci*, 557 U.S. at 562.

146. Kim, *supra* note 10, at 930 (footnote omitted) (quoting Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(l) (2012)). Kim notes that an applicant who might have been hired under a biased algorithm but was not because the employer discarded the model is in a meaningfully different position from a firefighter who studied in reliance on an announced testing process. *See id*.

147. *Id*. at 931.

the test design phase and prior to announcing the test process, the city had discovered that using a 30/70 weighting (written to oral components) rather than its planned 40/60 weighting would have produced equally valid results with less likelihood for disparate impact, then switching to 30/70 weighting would not constitute disparate treatment.[148] Importantly, though, that sort of action by the city would be a race-aware but *race-neutral* voluntary fix to the test design.[149]

Beyond simply discarding a biased algorithm, Professor Kim does not consider ways that an employer might try to *fix* an algorithm to remove the bias.[150] If the employer's fix is a race-neutral attempt to blind the algorithm to sensitive variables, it will likely be ineffective.[151] If the employer's fix is to implement a race-aware fairness constraint, then individual disappointed candidates will likely have a prima facie case that they suffered an adverse employment action because of the individual's race.[152]

The voluntary compliance efforts cited by the *Ricci* Court and Professor Kim are distinguishable from the situation facing our hypothetical employer. Although *Ricci*'s dicta described "ample room" for employers to design and revise selection procedures *in race-neutral ways* with a race-aware goal of enhancing diversity or avoiding disparate impact, that "ample room" may not encompass voluntary efforts by an employer that include *race-based* methods. This poses an interesting question with importance for the constitutional equal protection issue: whether a race-aware fairness constraint is race based, as the Court described the City of New Haven's conduct, or is instead race

148. *See Ricci*, 557 U.S. at 585; *see also supra* note 141 (considering 30/70 weighting as a possible alternative employment practice in *Ricci*). A race-based differential weighting (for example, 30/70 only for minorities, but 60/40 for whites) would presumably have been disapproved by the *Ricci* Court, given section 703(l).

149. Whether a race-aware fairness constraint is a racial "classification," for constitutional purposes, or is instead a formally race-neutral measure is explored in section III.A. *See* Reva B. Siegel, *Race-Conscious but Race-Neutral: The Constitutionality of Disparate Impact in the Roberts Court*, 66 ALA. L. REV. 653, 655–56 (2015) (noting the Court's apparent approval of race-conscious attempts to improve diversity so long as they are implemented in a formally race-neutral way, such as the State of Texas's "top ten percent" plan that provides automatic admission for any student finishing in the top ten percent of his or her class, regardless of race).

150. *See* Kim, *supra* note 10, at 928–32.

151. *See supra* note 59 and accompanying text.

152. *See* 42 U.S.C. § 2000e-2(a) (operative disparate treatment text); *supra* Section II.A (prima facie case); *cf. supra* text accompanying note 144. Group fairness approaches might fare better than individual fairness approaches under the "ample room" language from *Ricci*. An employer could argue that an algorithm using a group fairness constraint, such as the four-fifths selection rate rule, is analogous to: (1) applying a race-neutral selection technique; (2) checking the results against the four-fifths rule; (3) rejecting the technique if it fails that rule; (4) selecting the next-best race-neutral selection technique; and then using computational power to continually repeat that four-step process rapidly until it finds a selection technique that satisfies the fairness constraint. This analogy, however, could easily open the door to comparisons to quota systems. *See generally infra* note 236 and accompanying text (noting the Supreme Court's distinction between the "flexible" and "nonmechanical" use of race in higher education admissions programs and the impermissible use of quota systems). Individual fairness approaches, including counterfactual fairness, are not as readily analogized to the four-step process described above because they incorporate sex or race directly into the computations, at the individual level, as part of the fairness constraint.

neutral.[153] Certainly, the *Ricci* Court did not cite an example of conduct akin to race-aware algorithmic fairness constraints that would fit within the "ample room" for voluntary compliance. *Ricci*'s dicta, by itself, does not seem to provide statutory justification for race-aware fairness constraints.

D. THE VOLUNTARY AFFIRMATIVE ACTION JUSTIFICATION UNDER TITLE VII

The final, and most promising, legal justification for our hypothetical employer's use of a race-aware fairness constraint is permissible voluntary affirmative action. This section contends that current Title VII affirmative action doctrine already permits some uses of race-aware algorithmic fairness constraints, and that a clarification or modification to update the doctrine could justify algorithmic affirmative action more broadly.

Title VII permits employers to engage in voluntary race-based affirmative action pursuant to a valid affirmative action plan.[154] Where an employer acts pursuant to a valid affirmative action plan, it serves as the employer's legitimate "nondiscriminatory rationale" for a challenged employment action.[155] The Supreme Court in *Johnson v. Transportation Agency* applied the *McDonnell Douglas* framework as follows:

> Once a plaintiff establishes a prima facie case that race or sex has been taken into account in an employer's employment decision, the burden shifts to the employer to articulate a nondiscriminatory rationale for its decision. The existence of an affirmative action plan provides such a rationale. If such a plan is articulated as the basis for the employer's decision, the burden shifts to the plaintiff to prove that the employer's justification is pretextual and the plan is invalid.[156]

A plaintiff challenging our hypothetical employer's use of a race-aware fairness constraint would have no difficulty establishing a prima facie case that race has been taken into account in the employer's employment decision.[157] The inclusion of a race-aware fairness constraint in a hiring algorithm would, by definition, establish that.

At this point, our employer could point to an "affirmative action plan" to justify its use of race in the algorithmic hiring process, which would place the burden on the plaintiff to show that the plan was invalid. An initial question, though, is whether use of a race-aware fairness constraint in a machine-learning algorithm is really an affirmative action plan at all. The Supreme Court's decisions in

153. *See infra* Section III.A.

154. *See, e.g.*, Johnson v. Transp. Agency, 480 U.S. 616, 642 (1987) (Title VII not violated where employer took sex into account as part of a valid affirmative action plan featuring "a moderate, flexible, case-by-case approach" to improving minority and female representation in certain jobs); United Steelworkers v. Weber, 443 U.S. 193, 208 (1979) (Title VII "does not condemn all private, voluntary, race-conscious affirmative action plans.").

155. *See Johnson*, 480 U.S. at 626.

156. *Id.*

157. *See supra* Section II.A.

Johnson and *Weber* establishing this defense do not define the term "affirmative action plan."[158] A related question is whether there is any daylight between a voluntary affirmative action plan governed by *Weber* and *Johnson* and the sort of voluntary employer conduct governed by *Ricci*'s strong-basis-in-evidence standard. The *Ricci* Court did not evaluate the city of New Haven's refusal to certify test results under the *Weber* and *Johnson* affirmative action precedents, but rather borrowed its strong-basis-in-evidence test from the Court's equal protection jurisprudence, which includes *Croson* and *Wygant*.[159] The Supreme Court should not be presumed to have overruled the *Weber* and *Johnson* cases *sub silentio*, meaning that there must be some cases still governed by the *Weber–Johnson* doctrine.[160]

The Second Circuit attempted to clarify the line between situations governed by *Ricci* and affirmative action plans governed by *Weber* and *Johnson*. According to the Second Circuit, that distinction turns on whether the employer's conduct involves a forward-looking procedure benefitting an entire class or backward-looking individualized relief upsetting previously established procedures. The court explains:

> [W]hen an employer, acting *ex ante*, although in the light of past discrimination, establishes hiring or promotion procedures designed to promote equal opportunity and eradicate future discrimination, that may constitute an affirmative action plan. But where an employer, already having established its procedures in a certain way—such as through a seniority system—throws out the results of those procedures *ex post* because of the racial or gender composition of those results, that constitutes an individualized grant of employment benefits which must be individually justified [under *Ricci*], and not affirmative action.[161]

---

158. *See* United States v. Brennan, 650 F.3d 65, 99 (2d Cir. 2011) ("The Supreme Court has never said what, for purposes of the *Weber*/*Johnson* defense to § 703(a), is an affirmative action plan.").

159. *See* SULLIVAN & ZIMMER, *supra* note 1, at 162 ("While *Ricci* did not expressly discuss the validity of affirmative action under the *Weber*/*Johnson* analysis, there is at least considerable tension between its analysis and that of the earlier cases."); *supra* note 129 and accompanying text; *see also* Ricci v. DeStefano, 557 U.S. 557, 626 (2009) (Ginsburg, J., dissenting) ("This litigation does not involve affirmative action. But if the voluntary affirmative action at issue in *Johnson* does not discriminate within the meaning of Title VII, neither does an employer's reasonable effort to comply with Title VII's disparate-impact provision by refraining from action of doubtful consistency with business necessity."); Shea v. Kerry, 796 F.3d 42, 54–55 (D.C. Cir. 2015) (holding that *Ricci* did not implicitly overrule *Weber* and *Johnson*); Roberto L. Corrada, Ricci*'s Dicta: Signaling a New Standard for Affirmative Action Under Title VII?*, 46 WAKE FOREST L. REV. 241, 255–56 (2011) (arguing that *Ricci*'s holding must be confined to tests already taken, but that its dicta may suggest a new standard for voluntary affirmative action); Sachin S. Pandya, *Detecting the Stealth Erosion of Precedent: Affirmative Action After* Ricci, 31 BERKELEY J. EMP. & LAB. L. 285, 287 (2010) (arguing that *Ricci* involved a "stealth erosion" of *Weber* and *Johnson*).

160. *See Shea*, 796 F.3d at 54–55.

161. *Brennan*, 650 F.3d at 102; *see also* SULLIVAN & ZIMMER, *supra* note 1, at 162–63. The Second Circuit concluded that awarding retroactive competitive seniority relief fell on the *Ricci* side of the line, as ex post individualized restorative remedies rather than affirmative action plans. *See Brennan*, 650 F.3d at 104–05, 109.

If the Second Circuit has drawn the line between *Ricci* and *Weber–Johnson* correctly, then our hypothetical employer should be governed by *Weber–Johnson*.[162] The employer is not discarding, altering, or modifying the results of an announced test or an established procedure to provide backward-looking individualized relief. Rather, our hypothetical employer recognized that past discrimination can produce unintended bias in algorithmic output, saw indications of that possibility when testing the algorithm design, and introduced the fairness constraint acting ex ante to benefit an entire class of racial minorities and provide equal opportunities in the hiring process. Indeed, race-aware fairness constraints in many ways resemble traditional affirmative action plans that set hiring goals for minorities.[163] Group fairness constraints that require metrics like selection rates or false negative rates to be within some defined interval are just more mathematically complex and conceptually nuanced measures of the types of targets adopted in *Weber* and *Johnson*.[164] Individual fairness and counterfactual fairness concepts carry these ideas a step further, but they too are roughly analogous to traditional affirmative action plans.

Is our hypothetical employer's affirmative action plan valid? Under *Weber* and *Johnson*, an affirmative action plan is valid if: (1) there exists "manifest imbalance" in "traditionally segregated job categories";[165] (2) the plan does not "unnecessarily trammel[]" the rights of other employees or present an "absolute bar to their advancement";[166] and (3) the plan is temporary if it "sets aside positions according to specific numbers."[167]

The first requirement presents the biggest hurdle and will be met if the employer can point to a racial or gender imbalance in the relevant job category. For example, if Amazon introduced a gender-aware fairness constraint in an algorithm used to hire for technology jobs, it should satisfy the first requirement

---

162. The Second Circuit's opinion in *Brennan* was cited with approval by the D.C. Circuit in *Shea*, 796 F.3d at 55. In *Shea*, the court found that *Weber* and *Johnson* controlled (rather than *Ricci*) because the employer in *Shea* "did not modify the outcomes of personnel processes for the asserted purpose of avoiding disparate-impact liability under Title VII," but instead "acted to 'expand[] job opportunities for minorities and women' and to 'eliminate traditional patterns of racial segregation.'" *Id*. at 55 (alteration in original) (citation omitted) (quoting Johnson v. Transp. Agency, 480 U.S. 616, 622 (1987), and United Steelworkers v. Weber, 443 U.S. 193, 201 (1979)).

163. *See Johnson*, 480 U.S. at 621–22, 636 (involving an agency's long-term goal of a work force "whose composition reflected the proportion of minorities and women in the area labor force" and a short-term goal of "3 women for the 55 expected openings in [the Skilled Craft Workers] job category—a modest goal of about 6% for that category"); *Weber*, 443 U.S. at 199 (upholding a collectively bargained affirmative action plan providing that "at least 50% of the new trainees were to be black until the percentage of black skilled craftworkers in the . . . plant approximated the percentage of blacks in the local labor force").

164. *See supra* Section I.B.2 for group fairness metrics.

165. *Johnson*, 480 U.S. at 631 (internal quotation marks omitted) (quoting *Weber*, 443 U.S. at 197); *see* Corrada, *supra* note 159, at 243.

166. *Johnson*, 480 U.S. at 637–38; *see* Corrada, *supra* note 159, at 243–44 (arguing that an employer cannot require discharge of white employees or pose absolute bar to their advancement).

167. *Johnson*, 480 U.S. at 639–40; *see* Corrada, *supra* note 159, at 244 ("[The plan] can only be used to attain, and not to maintain, racial balance.").

because of historical male dominance in technology jobs.[168] Indeed, similar historical imbalances can be an important factor motivating employers to switch from subjective human decisions to machine-assisted decisions.

But, unlike Amazon, some employers considering the use of machine-learning algorithms may not be able to point to an existing imbalance in workforce demographics in traditionally segregated job categories. Those employers should still be able to use a race- or gender-aware fairness constraint if, during the design of an algorithmic hiring or promotion process, they learn that an unconstrained algorithm would likely produce actionable statistical disparities. On this point, *Weber* and *Johnson*, decided in the 1970s and 1980s, may require clarification or updating to deal with the new realities of machine-learning processes that tend to reproduce the effects of past bias in new employment decisions.

A sensible modification to (or clarification of) the first *Weber–Johnson* requirement would be to permit race-based "affirmative action" design adjustments to a procedure that has not yet been established, but only if either: (a) there exists a manifest imbalance in traditionally segregated job categories (*Weber–Johnson*), or (b) in the absence of the race-based adjustment, the procedure is likely to produce a statistical disparity that would satisfy a prima facie case of either disparate impact (*Griggs*)[169] or systemic disparate treatment (*Teamsters*, *Hazelwood*).[170]

Professor Roberto Corrada has argued that the voluntary compliance dicta in *Ricci* should be read as signaling that voluntary affirmative action is still permissible after *Ricci* if an employer can point to "a statistical showing disciplined by a technical analysis . . . in affirmative action or voluntary remediation cases in which a test has not yet been given."[171] This idea comports with the Second Circuit's distinction between backward-looking, *Ricci*-like conduct and forward-looking, voluntary affirmative action plans. *After* a test is given or a selection procedure is established, the employer can take race-based action to remedy statistical disparities only if it shows a strong basis in evidence of disparate impact liability, including a consideration of the employer's defenses. But *before* a test is

168. *See supra* notes 31–32 and accompanying text.

169. *See generally* Griggs v. Duke Power Co., 401 U.S. 424, 426 (1971) (first recognizing a disparate impact claim under Title VII where a minimum educational or intelligence test requirement had the effect of disqualifying minorities at "a substantially higher rate than white applicants").

170. *See generally* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 307 (1977) (approving the use of statistical analysis of disparities in applying the systemic disparate treatment theory to hiring practices); Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 336 (1977) (first recognizing a systemic disparate treatment theory under Title VII where a plaintiff can demonstrate that "racial discrimination was the company's standard operating procedure—the regular rather than the unusual practice.").

171. Corrada, *supra* note 159, at 259 (emphasis omitted) (asserting that standard deviation or regression analysis would satisfy statistical showing). Corrada bases his argument on the *Ricci* Court's approving citation of *Croson* and, in particular, on Justice O'Connor's concurring opinion in *Wygant*. *See id*. For the application of standard deviation analysis, see *Hazelwood*, 433 U.S. at 309 n.14. For multivariate regression analysis, see *Bazemore v. Friday*, 478 U.S. 385, 398–404 (1986). Although Professor Corrada's analysis focused on a prima facie showing of systemic disparate treatment, similar logic would support forward-looking affirmative action steps in response to a prima facie showing that disparate impact would result from a contemplated, but not yet administered, employment practice.

given or procedure is established, the employer can take forward-looking, race-based action if it has a "firm basis for determining that affirmative action is warranted," which could be satisfied by evidence sufficient to support only a prima facie showing of disparity, without consideration of the employer's possible defenses.[172] Applying this modification to *Weber–Johnson*'s first requirement, an employer would be able to implement a race-aware fairness constraint in a machine-learning algorithm if it had evidence to show that an unconstrained algorithm would likely produce disparities sufficient to trigger a prima facie case of disparate impact[173] or systemic disparate treatment.[174]

This view of the first requirement is further supported by existing EEOC regulations, which indicate that employers "may take affirmative action based on an analysis which reveals facts constituting actual or *potential* adverse impact, if such adverse impact is likely to result from existing or *contemplated* practices."[175] The EEOC regulations recognize that race-based affirmative action may be justified by potential adverse impact that is likely to result from contemplated practices, apart from whether the employer can point to an existing manifest imbalance.[176] The EEOC regulations further clarify that, where an employer's self-analysis shows that a practice will "tend to have an adverse effect," it has a reasonable basis to adopt affirmative action measures "without regard to whether there exists arguable defenses to a title VII action."[177] This is consistent with Professor Corrada's view that forward-looking, race-based affirmative action is justified by a prima facie case of statistical disparity, without regard to the business necessity defense.

The second *Weber–Johnson* requirement, that the action not "unnecessarily trammel" others' rights or pose an "absolute bar" to their advancement, should be readily met in algorithmic affirmative action cases. The race-aware fairness constraint in our hypothetical would introduce race and fairness metrics as only one

172. Wygant v. Jackson Bd. of Educ., 476 U.S. 267, 292 (1986) (O'Connor, J., concurring in part and concurring in the judgment) (arguing that such a "firm basis" standard would be met by "demonstrable evidence of a disparity . . . sufficient to support a prima facie Title VII pattern or practice claim"); *see* Corrada, *supra* note 159, at 259–60 (stating that a prima facie case of systemic discrimination should be "strong enough to ground voluntary remedial efforts").

173. Here, the EEOC's four-fifths rule of thumb for selection procedures could guide the employer. *See* 29 C.F.R. § 1607.4(D) (2018).

174. Here, *Hazelwood*'s "greater than two or three standard deviations" test, 433 U.S. at 309 n.14, and *Bazemore*'s discussion of the evidentiary value of multivariate regressions, 478 U.S. at 398–404, could guide the employer.

175. 29 C.F.R. § 1608.3(a) (2018) (emphasis added). In her *Ricci* dissent, Justice Ginsburg argued that the city should have been given the opportunity, on remand, to invoke reliance on this regulation as "a shield to liability." Ricci v. DeStefano, 557 U.S. 557, 631 n.9 (2008) (Ginsburg, J., dissenting). She notes that 29 C.F.R. § 1608.3(a) authorizes affirmative action "based on an analysis which reveals facts constituting actual or potential adverse impact," and that "[i]f 'affirmative action' is in order, so is the lesser step of discarding a dubious selection device." *Id.* (quoting 29 C.F.R. § 1608.3(a) (internal quotation marks omitted)).

176. In a separate subsection, the Guidelines authorize affirmative action "to correct the effects of prior discriminatory practices," which "can be initially identified by a comparison between the employer's work force . . . and an appropriate segment of the labor force." 29 C.F.R. § 1608.3(b).

177. 29 C.F.R. § 1608.4(b)(1), (3).

component of the algorithm. It would, of course, not pose an absolute bar to the selection of majority candidates for hiring or promotion. And race-aware fairness constraints are not likely to "unnecessarily trammel" the rights of those in the majority in failure-to-hire or failure-to-promote applications. Short of setting a tightly defined demographic-parity fairness rule and using that algorithmic fairness rule to make determinations about which current employees to *terminate* (as opposed to hire or promote), algorithmic affirmative action should not run afoul of the second requirement.[178]

The third *Weber–Johnson* requirement is designed to ensure that affirmative action measures are used to attain, but not maintain, balance in workforce demographics.[179] Where a plan sets aside a specific number or percentage of positions for underrepresented groups, the plan may need to include express assurance that it is temporary.[180] This requirement arguably would apply to use of a restrictive demographic parity fairness constraint, which demands that an algorithm produce a positive prediction (good employee) at equal rates for majority and minority groups.[181] It likely would not apply to other race-aware measures of algorithmic fairness.[182] In any event, the race-aware affirmative measures should be viewed as necessary to "attain" fair results from the use of a particular hiring or promotion process, rather than a means of "maintaining" a certain quota, ratio, or demographic balance in the employer's workforce in perpetuity. The EEOC regulations require an ongoing evaluation of any race-based measures: "The race, sex, and national origin conscious provisions of the [affirmative action] plan or program should be maintained only so long as is necessary" to achieve the objectives identified by the employer in the self-analysis and to "ensure that employment systems operate fairly in the future, while avoiding unnecessary restrictions on opportunities for the workforce as a whole."[183]

Employers can and should continually reassess their algorithm model design, including a reassessment of whether a race-aware fairness constraint remains necessary to avoid statistical disparities that would trigger a prima facie case of disparate impact or systemic disparate treatment.[184] Combined with the modification

178. *See* Johnson v. Transp. Agency, 480 U.S. 616, 638 (1987) ("Thus, denial of the promotion unsettled no legitimate, firmly rooted expectation on the part of petitioner. Furthermore, while petitioner in this case was denied a promotion, he retained his employment with the Agency . . . and remained eligible for other promotions."); United Steelworkers v. Weber, 443 U.S. 193, 208 (1979) ("[T]he plan does not unnecessarily trammel the interests of the white employees. The plan does not require the discharge of white workers and their replacement with new black hirees.").

179. *See Johnson*, 480 U.S. at 630, 639.

180. *See id.* at 639–40.

181. *See supra* notes 61–63 and accompanying text.

182. *See supra* Sections I.B.2–4.

183. 29 C.F.R. § 1608.4(c)(2)(i) (1979).

184. *Cf. id.* ("The plan should be tailored to solve the problems which were identified in the self analysis and to ensure that employment systems operate fairly in the future, while avoiding unnecessary restrictions on opportunities for the workforce as a whole. The race, sex, and national origin conscious provisions of the plan or program should be maintained only so long as is necessary to achieve these objectives." (citation omitted)).

to *Weber–Johnson*'s manifest imbalance requirement, this understanding of the third *Weber–Johnson* requirement would make for a sensible approach to voluntary, race-aware fairness techniques in designing algorithmic hiring models and would be consistent with the EEOC's Guidelines on affirmative action.

E. A FINAL STATUTORY HURDLE: ARE MACHINE-LEARNING ALGORITHMS "EMPLOYMENT RELATED TESTS" SUBJECT TO SECTION 703(L)?

If race-aware fairness constraints are justified as valid affirmative action plans under *Weber–Johnson*, one final statutory hurdle remains: section 703(*l*)'s prohibition on race-norming. As the Introduction illustrated, using different cutoffs for, or adjusting the scores of, "employment related tests" according to race, gender, or another protected characteristic (that is, race-norming) is prohibited by section 703(*l*).[185] This is true even for affirmative action plans that are otherwise valid under *Weber–Johnson*.[186] The hypothetical City of Springfield therefore could not differentially weight its firefighter-promotion exam by race even if it tried to do so as part of a valid affirmative action plan.

Some algorithmic fairness techniques use protected traits to identify variables that predict success for one particular group (for example, females), even if different values of the same variable predict success for other groups (for example, males). As Professor Dwork articulates, *using* the race or gender variable can help provide important context, like cultural differences, that can actually increase the accuracy of target variable predictions within groups.[187] Dwork gives the following example:

> Suppose you had a minority group in which the smart students were steered toward math and science, and a dominant group in which the smart students were steered toward finance. Now if someone wanted to write a quick-and-dirty classifier to find smart students, maybe they should just look for students who study finance because, after all, the majority is much bigger than the minority, and so the classifier will be pretty accurate overall. The problem is that not only is this unfair to the minority, but it also has reduced utility compared to a classifier that understands that if you're a member of the minority and you study math, you should be viewed as similar to a member of the majority who studies finance. That gave rise to the title of the paper, "Fairness Through Awareness," meaning cross-cultural awareness.[188]

Is this kind of algorithmic affirmative action a variant of race-norming of employment-related tests that is prohibited by section 703(*l*)? Are other fairness constraints, like group fairness and counterfactual fairness, prohibited under

185. *See* 42 U.S.C. § 2000e-2(l) (2012).

186. *See* Dean v. City of Shreveport, 438 F.3d 448, 462–63 (5th Cir. 2006) (arguing that even if the process was a valid affirmative action plan, it must not violate the specific prohibitions of section 703(l)).

187. *See supra* notes 18 and 46.

188. Hartnett, *supra* note 18.

section 703(*l*)? This section contends that the answer should be no, based on the text, history, and purpose of section 703(*l*).

First, the text of section 703(*l*) covers only employment-related tests. It does not, by its terms, cover machine-learning algorithms, observational data review techniques or formulas, or hiring or promotion procedures or practices other than employment-related tests.[189] Congress did not define employment-related tests in Title VII.[190] The use of this narrow term stands in contrast to Title VII's disparate impact provisions (also added by Congress in 1991), which use the broader term "employment practice" to describe employer conduct that may trigger disparate impact liability.[191] Likewise, the primary operative provisions of Title VII use the broader term employment practice.[192] The narrow use of employment-related tests in section 703(*l*) also differs from broader language used elsewhere in federal antidiscrimination statutes. For example, the ADA refers to "*qualification standards*, employment tests *or other selection criteria* that screen out or tend to screen out an individual with a disability."[193] These textual differences suggest that section 703(*l*)'s more narrowly drawn prohibition does not apply to employment standards, criteria, or selection procedures other than employment-related tests.[194]

Second, the legislative history of section 703(*l*) shows that its drafters never contemplated its application to non-testing selection procedures. Nor did its drafters envision it governing complex systems, like fairness-constrained machine-learning algorithms, that are now capable of evaluating the within-group predictive accuracy of some subset of observed variables (that is, individual fairness). Rather, section 703(*l*) was intended to prohibit the relatively unsophisticated techniques for race- or gender-norming the overall scores of so-called "neutral" or "nondiscriminatory" ability tests taken by applicants. This purpose is shown in the debates over race-norming scores on ability tests like the Department of Labor's General Aptitude Test Battery (GATB) that occurred in the lead-up to passage of the Civil Rights Act of 1991.[195]

---

189. *See* 42 U.S.C. § 2000e-(2)(l).

190. *See generally* 42 U.S.C. § 2000e (lacking any explanation of what constitutes an employment-related test).

191. *See* 42 U.S.C. § 2000e-2(k)(1)(A)(i) (stating that using "a particular *employment practice* that causes a disparate impact" can establish an unlawful employment practice (emphasis added)).

192. *See* 42 U.S.C. § 2000e-(2)(a).

193. 42 U.S.C. § 12112(b)(6) (emphasis added); *see also* 42 U.S.C. § 12113(a), (c) (using the same language in describing defenses).

194. The EEOC's regulations interpreting Title VII also describe a category of employment selection procedures beyond tests. For example, the Uniform Guidelines on Employee Selection Procedures, adopted in 1978, apply to "tests and other selection procedures which are used as a basis for any employment decision[s]." *See* 29 C.F.R. § 1607.2(B) (2018).

195. *See* Paul S. Greenlaw & Sanne S. Jensen, *Race-Norming and The Civil Rights Act of 1991*, 25 PUB. PERSONNEL MGMT. 13, 13 (1996); Kimberly West-Faulcon, *Fairness Feuds: Competing Conceptions of Title VII Discriminatory Testing*, 46 WAKE FOREST L. REV. 1035, 1051–52 (2011) ("In the politically turbulent years during which Congress debated the failed Civil Rights Act of 1990 and the 1991 Act, the race norming of employment test scores became a hot button political issue and the subject of intense media scrutiny." (footnotes omitted)).

By the late 1980s, the Department of Labor's GATB was the most widely used employment test, used to screen and refer applicants for many public and private sector jobs.[196] Based on a statistical analysis suggesting that the GATB could serve as a valid predictive test "for selecting employees for all 12,000 jobs in the U.S. economy," the Department of Labor initiated a job-referral system using GATB results.[197] Realizing that the use of GATB raw score rankings would lead to racially disparate results and likely run afoul of the disparate impact doctrine announced in *Griggs*, the Department of Labor race-normed the scores by converting raw GATB scores into a "racial-group-based percentile ranking."[198] A raw score of 300 on the GATB might place a white candidate at the forty-fifth percentile for white test-takers, but that same raw score might place a black candidate at the eighty-third percentile of black test-takers.[199] Referrals were then made using the percentile score. This type of simplistic race-norming was a highly divisive issue during the 1991 congressional debates over civil rights reform.[200]

Congress's 1990 attempt to amend the Civil Rights Act was vetoed by President George H.W. Bush out of concern that it would lead employers to hire by quota.[201] The 1990 bill did not include express prohibition on race-norming, and some believed it actually *required* race-norming to avoid disparate impact liability from testing.[202] Following President Bush's veto, the 102nd Congress began a second effort at reforming Title VII. The first bill to directly address race-norming was Senate Bill 478, introduced by Senator Alan Simpson.[203] That bill included a section entitled "Prohibition on Discrimination in Employment Ability Tests."[204] The operative language read:

> It shall be an unlawful employment practice for an employer, when selecting applicants for employment, or when selecting current employees for promotion, to interpret or adjust the results of *ability tests* in a manner that would

196. The Department of Labor's GATB included twelve subtests designed to evaluate various skills and qualities, including "cognitive, perceptual and manual dexterity skills." *See* Greenlaw & Jensen, *supra* note 195, at 13.

197. West-Faulcon, *supra* note 195, at 1053–54; *see also* Greenlaw & Jensen, *supra* note 195, at 13–14.

198. West-Faulcon, *supra* note 195, at 1055; *see also* Greenlaw & Jensen, *supra* note 195, at 14 (stating that race-norming was invented "to promote federal equal employment opportunity and affirmative action goals").

199. *See* Greenlaw & Jensen, *supra* note 195, at 14.

200. *See* West-Faulcon, *supra* note 195, at 1051–52. Other forms of simplistic race-norming added points to the scores of minority candidates. *See* Greenlaw & Jensen, *supra* note 195, at 14.

201. *See* Reginald C. Govan, *Honorable Compromises and the Moral High Ground: The Conflict Between the Rhetoric and the Content of the Civil Rights Act of 1991*, 46 RUTGERS L. REV. 1, 148–51 (1993); Greenlaw & Jensen, *supra* note 195, at 18; Steven A. Holmes, *President Vetoes Bill on Job Rights; Showdown Is Set*, N.Y. TIMES, Oct. 22, 1990, at A1.

202. *See* Greenlaw & Jensen, *supra* note 195, at 18.

203. *See* S. 478, 102d Cong. (1991); *see also* Govan, *supra* note 201, at 179.

204. S. 478, 102d Cong. § 5 (1991).

> discriminate against any individual because of the race, color, religion, sex, or national origin of the individual.[205]

A second subsection provided that, in disparate impact cases where the employer had demonstrated a business necessity for an "ability test," plaintiffs could not prevail by showing that the results could have been "adjusted on the basis of" a protected characteristic in order to "reduce the disparate impact of the test."[206]

Senator Simpson explained that his bill addressed the "new issue" of "whether neutral, nondiscriminatory employment tests should be race normed" to benefit minority groups.[207] Senator Simpson cited support from R. Gaull Silberman, the Vice Chairman of the EEOC, and offered a transcript of her remarks opposed to race-norming employment tests. Vice Chair Silberman specifically cited the "within-group" race-norming of the GATB by State Employment Service offices.[208] She also cited differential standards under the Foreign Service Officer competitive written exam and an employer's competitive test that was "a sample of the actual job."[209] The issue of race-norming remained one of the key issues during the 1991 debates and negotiations over compromise, bipartisan bills.[210] Despite early indications to the contrary, Democrats in the House of Representatives realized that any successful bill would need to include language prohibiting race-norming of written tests.[211]

Senator Simpson's bill was referred to committee but did not advance further. A bipartisan compromise bill (Senate Bill 1745) eventually emerged in September 1991 that included similar (though not identical) language prohibiting within-group norming.[212] That language would eventually become section 703(*l*) of Title VII. Statements in support of Senate Bill 1745 show that the concern motivating the relevant language was the practice of adjusting scores for ability tests like the GATB.[213]

Nothing in the legislative history suggests that section 703(*l*) was directed at any employment-selection procedures other than neutral and nondiscriminatory employment related tests, like the GATB and similar ability tests were claimed to

---

205. *Id*. § 5(m)(1) (emphasis added).

206. *Id*. § 5(m)(2).

207. 137 CONG. REC. 9077 (1991) (statement of Sen. Alan Simpson).

208. *Id*. at 9,077–78 (excerpts of comments of R. Gaull Silberman before the Equal Employment Advisory Council, Feb. 28, 1991).

209. *Id*. at 9,078.

210. *See* Govan, *supra* note 201, at 6, 183–84, 188, 190 (describing race-norming of tests as one of four major policy issues in the debates and noting that "in the political context it was dynamite").

211. *See id*. at 184, 188, 190.

212. *See* Civil Rights Act of 1991, Pub. L. No. 102-166, § 106, 105 Stat. 1071, 1075 (1991).

213. *See* 137 CONG. REC. H9547 (daily ed. Nov. 7, 1991) (statement of Sen. John Danforth) ("This means, for instance, that discriminatory use of the Generalized Aptitude Test Battery (GATB) by the Department of Labor's and state employment agencies' is illegal."); 137 CONG. REC. S15,476 (daily ed. Oct. 30, 1991); 137 CONG. REC. H3951 (daily ed. June 5, 1991) (citing the "controversial score adjustments" for the GATB "which has sparked the debate over race norming").

be. There is no evidence in the legislative history even hinting that an employer operating within a valid affirmative action plan is prohibited from "adjusting" or setting separate "cutoffs" for any other items in a candidate's portfolio when making comparisons between candidates. Imagine, for example, an employer operating within a valid affirmative action plan who considers for interviews only those white applicants having at least a 3.0 high school GPA but considers minority applicants with at least a 2.75 GPA. There is nothing in the text or the legislative history suggesting that section 703(*l*) would prohibit such a practice. It may be the use of a different cutoff for a hiring criterion, but it is not a different cutoff for an employment-related test. The employer uses GPA in this scenario as a predictor of future success. Algorithms using observational variables to make predictions about future success are likewise not employment-related tests. And, of course, there is nothing in the 1991 legislative history suggesting the possible application of section 703(*l*) to complex machine-learning processes (which had not yet been developed) performing calculations on a set of observed variables.

Section 703(*l*) was included in the 1991 Civil Rights Act as part of a larger compromise. The widespread race-norming of general aptitude tests was controversial, and section 703(*l*) definitively ended that practice. But section 703(*l*) was limited in scope to employment-related tests. It did not prevent the use of different cutoffs or adjustments for other selection processes or criteria. It did not prohibit the use of preferences pursuant to a valid affirmative action plan.[214] And it does not prohibit an employer's use of race-aware fairness constraints in machine-learning algorithms as part of a valid affirmative action plan.

III. THE CONSTITUTIONAL CHALLENGE TO ALGORITHMIC AFFIRMATIVE ACTION

For public sector employers, race-aware fairness constraints can also be challenged as a constitutional equal protection violation.[215] Although the constitutional question is framed differently, many of the central considerations are the same as for the statutory analysis. This Part examines how algorithmic affirmative action may fare under the Court's recent Equal Protection Clause jurisprudence. The discussion focuses on the public employer hiring context, though a similar analysis would be applicable to other governmental uses of algorithmic fairness constraints, such as public university admissions processes.

---

214. *See* Civil Rights Act § 116 ("Nothing in the amendments made by this title shall be construed to affect court-ordered remedies, affirmative action, or conciliation agreements, that are in accordance with the law."); *see also* Michael J. Zimmer, Taxman*: Affirmative Action Dodges Five Bullets*, 1 U. PA. J. LAB. & EMP. L. 229, 235 (1998) ("Whether or not section 116 amounts to a full reenactment of *Weber*/*Johnson*, it is a statement recognizing the law in those cases. Thus, under any view, section 116 bolsters *Weber*/*Johnson* and the stare decisis effect that the courts should give to that law.").

215. *See* U.S. CONST. amends. V, XIV; *see also* Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 227 (1995) (holding that "all racial classifications, imposed by whatever federal, state, or local governmental actor, must be analyzed by a reviewing court under strict scrutiny"); City of Richmond v. J. A. Croson Co., 488 U.S. 469, 493–94 (1989) (applying strict scrutiny in an equal protection challenge to a municipal contract system that favored minority-owned businesses); Wygant v. Jackson Bd. of Educ., 476 U.S. 267, 273–74 (1986) (applying strict scrutiny to race-based preferences in school board layoffs).

Would the City of Springfield's use of an algorithm with a race-aware fairness constraint to make firefighter promotions survive constitutional challenge? The starting point for the constitutional analysis is the text of the Equal Protection Clause of the Fourteenth Amendment, which provides: "No State shall . . . deny to any person within its jurisdiction the equal protection of the laws."[216]

A. CLASSIFICATION: IS ALGORITHMIC AFFIRMATIVE ACTION RACE-NEUTRAL?

The Supreme Court in *Adarand Constructors, Inc. v. Pena* held that "all racial classifications" by governmental actors must be justified under the strict scrutiny test; they must be narrowly tailored to further a compelling governmental interest.[217] In *Adarand*, the Court held that the federal Department of Transportation's affirmative action program providing race-based preferences in awarding contracts could be justified "only for the most compelling reasons."[218]

An initial question is whether the state employer's use of a race-aware fairness constraint is a racial "classification" at all. Professor Sam Bagenstos observes that the Court's most recent equal protection cases have placed tremendous importance on the meaning of the term classification[219]—a term the Court has never defined.[220] Bagenstos predicts that the importance placed on the meaning of classification may lead the Court to "interpret[] the concept elastically, by balancing the equal protection interests it finds salient, but doing so offstage."[221] Whether a race-aware algorithmic fairness constraint constitutes a classification is a complex question, especially in light of the Court's ambiguity. The answer may depend on the specific type of fairness constraint (group, individual, or causal/counterfactual) and on the method of implementation (preprocessing, optimization penalties, or other methods).

---

216. U.S. CONST. amend. XIV, § 1. This provision of the Fourteenth Amendment directly protects employees or applicants for employment by state or local governments. The Fifth Amendment's Due Process Clause protects employees or applicants for federal employment, and its protections have been held to be coextensive with the Fourteenth Amendment. *See* U.S. CONST. amend. V; *Adarand*, 515 U.S. at 227 ("[W]e hold today that all racial classifications, imposed by whatever federal, state, or local governmental actor, must be analyzed by a reviewing court under strict scrutiny. In other words, such classifications are constitutional only if they are narrowly tailored measures that further compelling governmental interests.").

217. *See Adarand*, 515 U.S. at 227.

218. *See id.*

219. Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities*, 101 CORNELL L. REV. 1115, 1166 (2016) ("By formalistically relying on the existence of a classification as the trigger for strict scrutiny, the Court's approach puts great pressure on the definition of 'classification.'").

220. *See id.* at 1119 n.10 (citing Reva B. Siegel, *The Supreme Court, 2012 Term—Forward: Equality Divided*, 127 HARV. L. REV. 1, 48–49 (2013)).

221. *See id.* at 1166. Bagenstos posits that the Court's jurisprudence has evolved to mean that all racial classifications must meet strict scrutiny but that race-neutral state actions that are not classifications will be subjected to strict scrutiny only if adopted to "harm [a] racial group[]," rather than for the purpose of "promot[ing] integration or clos[ing] racial gaps." *See id.* at 1158–59.

Some guideposts do exist.[222] Equal protection cases where racial classifications must have occurred include the affirmative action cases *Gratz v. Bollinger*,[223] *Grutter v. Bollinger*,[224] *Parents Involved in Community Schools v. Seattle School District No. 1*,[225] and *Adarand*,[226] each of which required the application of strict scrutiny.[227] At the other end of the spectrum, the Court has approvingly considered examples of formally race-neutral techniques that do not appear to count as racial classifications, even though motivated by race-conscious goals. One example is Texas's Top Ten Percent Plan, in which high school students graduating in the top ten percent of their class are automatically admitted to the state university. The top ten percent aspect of the University's admission plan was not subject to strict scrutiny in *Fisher v. University of Texas at Austin*,[228] even though it was unquestionably adopted with the specific purpose of improving racial balance.[229] Justice Kennedy gave other examples in his concurrence in *Parents Involved*: "strategic site selection of new schools; drawing attendance zones with general recognition of the demographics of neighborhoods; allocating resources for special programs; recruiting students and faculty in a targeted fashion; and tracking enrollments, performance, and other statistics by race."[230] According to Justice Kennedy, these techniques may be "race conscious" but they "do not lead to different treatment based on a classification that tells each student he or she is to be defined by race, so it is unlikely any of them would demand strict scrutiny to be found permissible."[231]

Ricci is a tougher case to characterize. It may have involved a racial classification in the Court's view. The majority described the city's conduct in refusing to certify the test results as "express, race-based decisionmaking" even though the

222. *See id.* at 1157 (describing cases where the state "concededly act[ed] on the basis of the race of the particular individuals who sought a benefit from them").

223. 539 U.S. 244, 255 (2003) (involving the use of race to award a predetermined amount of points in undergraduate admissions to the University of Michigan).

224. 539 U.S. 306, 343 (2003) (involving the use of race as just one factor in a more nebulous and flexible manner in law school admissions in service of the compelling state interest of promoting diversity in higher education).

225. 551 U.S. 701, 709–10, 717 (2007) (involving the use of race in determining school assignments and ruling on transfer requests in public school districts, where districts maintained predetermined range of target racial balance in schools).

226. 515 U.S. 200, 227 (1995) (involving the use of race in making rebuttable presumptions of social and economic disadvantage in connection with awarding federal government contract work).

227. *See* Bagenstos, *supra* note 219, at 1157.

228. 136 S.Ct. 2198, 2209 (2016).

229. *See* Bagenstos, *supra* note 219, at 1145 ("Notably, none of the justices who joined the majority opinion . . . expressed any doubt that the 'Top Ten Percent plan,' which was framed in race-neutral terms but plainly aimed at increasing racial diversity, was constitutional" (footnote omitted)); Siegel, *supra* note 149, at 674 ("No Justice raised questions about the constitutionality of the percent plan. The Court's acceptance of the percent plan illustrates that government may act in race-conscious but facially neutral ways to promote equal opportunity, even where government seeks to alter racial outcomes." (emphasis omitted)).

230. *Parents Involved*, 551 U.S. at 789 (Kennedy, J., concurring in part and concurring in the judgment).

231. *See id.*

city had the benign purpose of avoiding impact liability from the racially skewed outcome of the testing.[232] On the other hand, the city's conduct in refusing to certify test results might be thought of as formally race-neutral.[233] As others have observed, one plausible reading of *Ricci* is that the city's conduct was a racial classification, rather than race-neutral conduct, only because the city acted *after* administering the tests and learning which candidates would be eligible for promotions.[234] If that reading is correct, then perhaps the forward-looking use of a race-aware fairness constraint is formally race-neutral, in the same sense that disparate impact legislation itself can be thought of as formally race-neutral.[235] The reading would require the algorithm's output to meet some overall group target range, which necessarily requires consideration of race in the aggregate, but it would not make *individual* determinations on the basis of race in the same way as *Gratz* (automatically awarding points to individual applicants according to race) or *Parents Involved* (after a certain threshold was reached, denying school transfer requests if the individual requesting student would contribute to racial imbalance). Of course, this question is by no means settled. The definition of classification remains murky. Courts could readily conclude that some or all of the race-aware fairness constraints described in Part I constitute classifications for equal protection purposes and therefore demand strict scrutiny.

Of the three basic types of race-aware fairness constraints described in Part I, a government's use of group fairness constraints seems most vulnerable to constitutional attack. The starkest example would be governmental use of a demographic-parity constraint, where positive (good employee) predictions must be made at the same (or nearly the same) rate for both the majority and minority groups. Without any carve-out for business relatedness, the government's use of a race-aware, demographic-parity constraint in an algorithm could easily be equated to a form of quota and deemed to impose a racial classification.[236] In contrast, an individual fairness constraint may not constitute racial classifications. Because it operates by identifying individuals who are similar as measured by relevant variables and then comparing the predictions for those individuals, an individual

232. Ricci v. DeStefano, 557 U.S. 557, 579 (2009); *see supra* text accompanying note 121; *see also* Bagenstos, *supra* note 219, at 1151 ("One way of reading the *Ricci* Court's holding . . . is that even formally race-neutral efforts to avoid racially disparate impacts are the equivalent of racial classifications.").

233. *See* Bagenstos, *supra* note 219, at 1150 ("New Haven's refusal to certify the promotion tests could be understood as formally race-neutral.").

234. *See id.* at 1151; Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1369–74 (2010) (describing a "visible victims" reading of *Ricci*); Siegel, *supra* note 149, at 682–83.

235. *See* Bagenstos, *supra* note 219, at 1130 ("Because prohibitions on disparate impact do not individually classify people based on their race, in this view, the prohibitions are not themselves constitutionally suspect simply because they seek to achieve the 'race-conscious' goals of promoting integration and closing racial gaps.").

236. *Cf.* Grutter v. Bollinger, 539 U.S. 306, 334 (2003) (holding that a race-conscious higher education admission plan cannot be narrowly tailored if it uses a "quota system," but can be narrowly tailored if it uses race in a "flexible, nonmechanical way" as part of an admissions program that involves "truly individualized consideration" of applicants).

fairness approach might be described as race-aware but race-neutral. Counterfactual fairness might also be understood as race-neutral rather than as a racial classification. Although it requires knowledge and manipulation of each individual's race variable, the counterfactual fairness model does not care whether any individual's race is switched from black to white or vice-versa. It demands only that the machine's predictions in the racially counterfactual world closely resemble the machine's predictions under the actual facts.

Perhaps the most interesting implication of the Court's recent equal protection jurisprudence is an idea that both legal and machine-learning scholars have not yet seized upon: algorithmic affirmative action might be on better legal footing if it were designed to operate in formally race-neutral ways, despite its admitted and obvious race-conscious purpose. Imagine, for example, an algorithm that is formally blinded to each individual's race, but is optimized to achieve counterfactual fairness when the person's *neighborhood* is flipped from a poor neighborhood known to be predominantly minority to a wealthy one known to be predominantly white. Think of it as algorithmic affirmative action by proxy. It would work on the same principle as the Texas ten percent plan and Justice Kennedy's example of drawing school district lines with knowledge of the general neighborhood demographics. Of course, the machine, if asked to, could accurately guess each individual's probable race based on the individual's neighborhood, but the same is true of human decisionmakers drawing school district lines. Under the Court's recent equal protection jurisprudence, known proxies for protected characteristics can be leveraged in algorithm design to accomplish the race-conscious goal of reducing racial disparities in race-neutral ways.

B. APPLYING SCRUTINY

Assuming that governmental use of a race- or gender-aware fairness constraint is a classification for equal protection purposes, should it survive heightened scrutiny? Under current equal protection jurisprudence, a race-aware constraint would need to be narrowly tailored to compelling governmental interest (strict scrutiny).[237] A gender-aware constraint would need to be substantially related to an important governmental interest (intermediate scrutiny).[238]

The governmental interest at stake in either case would be avoiding unintended discrimination on the basis of race or sex in hiring that would otherwise be

237. *See* Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 227 (1995).

238. *See* United States v. Virginia, 518 U.S. 515, 533 (1996); Miss. Univ. for Women v. Hogan, 458 U.S. 718, 724 (1982); Craig v. Boren, 429 U.S. 190, 197 (1976); *see also Adarand*, 515 U.S. at 247 (Stevens, J., dissenting) (noting the "anomalous result that the Government can more easily enact affirmative-action programs to remedy discrimination against women than it can enact affirmative-action programs to remedy discrimination against African-Americans—even though the primary purpose of the Equal Protection Clause was to end discrimination against the former slaves"). The Court has not equated sex-based classifications with racial classifications, which receive the highest level of scrutiny. Nonetheless, the Court "has carefully inspected official action that closes a door or denies opportunity to women (or to men)." *Virginia*, 518 U.S. at 532.

produced by an unconstrained algorithm, due to one or more sources of bias identified by Barocos and Selbst.[239] Whether this is a compelling or substantial governmental interest is unclear. The only two interests that the Supreme Court has definitively held compelling enough to satisfy strict scrutiny are remedying the effects of prior intentional discrimination by the state actor in question[240] and diversity in higher education.[241] Remedying the effects of general societal discrimination was held not sufficiently compelling in *Croson*.[242]

Some governmental employers may fit within the first recognized compelling interest, if they can point to their own prior intentional discrimination. But many other governmental actors will need to look elsewhere for a compelling interest. It is an open question whether diversity in public employment (or in certain spheres of public employment) is a compelling interest.[243] Short of a diversity interest, perhaps the governmental actor has a sufficient interest in *preventing* its own contemplated employment practices from producing statistical disparate impact by reproducing the effects of prior discrimination.[244] There is some authority, if attenuated, for this view. Although it was resolved on purely statutory grounds,[245] *Johnson* approved gender-based affirmative action by a government employer even in the absence of a showing that the manifest imbalance in traditionally segregated job categories was traceable to the defendant's own prior, intentional discrimination.[246] Second, the EEOC regulations interpret Title VII to permit race-conscious affirmative action when a self-analysis shows that a contemplated practice is likely to lead to actual or potential disparate impacts.[247] We should not lightly presume the EEOC adopted an interpretation of Title VII that condones racial classifications under circumstances that would never establish a compelling state interest. Third, Justice Ginsburg observed that affirmative action might have been warranted under the facts in *Ricci*, which did not involve the

239. *See supra* Section I.A.

240. *See* Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720–21 (2007); Wygant v. Jackson Bd. of Educ., 476 U.S. 267, 277 (1986) (plurality opinion).

241. *See Parents Involved*, 551 U.S. at 720, 722; *Grutter*, 539 U.S. at 328.

242. *See* City of Richmond v. J. A. Croson Co., 488 U.S. 469, 505 (1989).

243. *Compare* Petit v. City of Chicago, 352 F.3d 1111, 1114 (7th Cir. 2003) (holding that "the City of Chicago has set out a compelling operational need for a diverse police department"), *with* Rothe Dev. Corp. v. Dep't of Defense, 545 F.3d 1023, 1027, 1040 (Fed. Cir. 2008) (holding that there was no "substantially probative and broad-based statistical foundation" justifying a "nationwide, race-conscious action" implemented by the Department of Defense to award defense contracts to entities controlled by "socially and economically disadvantaged individuals"), *and* Alexander v. City of Milwaukee, 474 F.3d 437, 445–46 (7th Cir. 2007) (holding that "[a] race-conscious promotion system" favoring women and minorities "cannot pass constitutional scrutiny").

244. *See* Deborah Hellman, *Indirect Discrimination and the Duty to Avoid Compounding Injustice*, *in* FOUNDATIONS OF INDIRECT DISCRIMINATION LAW 105, 120 (Hugh Collins & Tarunabh Khaitan eds., 2018).

245. *See* Johnson v. Transp. Agency, 480 U.S. 616, 620 n.2 (1987).

246. *See id*. at 659 (Scalia, J., dissenting) ("Most importantly, the plan's purpose was assuredly not to remedy prior sex discrimination by the Agency. It could not have been, because there was no prior sex discrimination to remedy.").

247. *See supra* note 175 and accompanying text.

remediation of the City of New Haven's own prior discrimination.[248] Finally, there is academic support for the proposition that affirmative action ought to be justified where deployed in response to statistical evidence of disparities (short of findings of prior intentional discrimination)[249] or to prevent discrimination within an employer's own workplace.[250]

If there is a compelling interest, are race-aware fairness constraints narrowly tailored to serve that interest? The answer is necessarily dependent on the algorithmic design selected, other race-neutral or less restrictive options considered, and how the algorithm's predictions are being used. But a few generalizations can be made. First, algorithms used in making hiring or promotion decisions are much more likely to be upheld than algorithms used to make termination or layoff decisions.[251] Second, race-aware algorithmic fairness constraints by their nature do not define individuals solely by their race, but include race as just one factor among all the observed variables available in the data. In this way, they more closely resemble the narrowly-tailored law school admissions practices upheld in *Grutter* (race as one factor in holistic review)[252] than they do the undergraduate admissions practices struck down in *Gratz* (awarding individuals predefined points based on race).[253] This cuts in favor of finding that race-aware fairness constraints are narrowly tailored to serve the government's interests.

Finally, before voluntarily instituting any race-aware fairness constraint, government employers should consider other less restrictive options, which may include formally race-neutral means that operate on proxies to accomplish race-conscious goals. Simply blinding the algorithm to the sensitive variable will not work in most cases. But, blinding the algorithm and then leveraging known proxies to calculate a fairness constraint might work to avoid or reduce racially disparate impacts from the algorithmic output. If a governmental employer does choose a race-aware fairness constraint, it should be as flexible as possible while still capable of serving its purposes. For example, a group fairness metric should not demand strict equality between, say, false-positive rates if a wider permissible range of difference in the false-positive rates would work nearly as well in eliminating disparate impacts. If using Kusner's counterfactual fairness technique, the metric specified to determine whether counterfactual predictions are close enough to predictions in the actual world should permit leeway for some differences

248. *See supra* note 175.

249. *See* Corrada, *supra* note 159, at 251–52, 258–59.

250. *See* Michael J. Yelnosky, *The Prevention Justification for Affirmative Action*, 64 OHIO ST. L.J. 1385, 1386–87 (2003) (arguing that prevention of discrimination ought to justify affirmative action under Title VII and that this interpretation is consistent with existing Title VII doctrine).

251. *See* Wygant v. Jackson Bd. of Educ., 476 U.S. 267, 282 (1986) (plurality opinion) ("Though hiring goals may burden some innocent individuals, they simply do not impose the same kind of injury that layoffs impose.").

252. *See* Grutter v. Bollinger, 539 U.S. 306, 334 (2003). However, a strict group fairness approach, and particularly a strict demographic parity requirement, might be equated to a quota system that was disapproved in *Grutter* because it "insulat[es] the individual from comparison with all other candidates for the available seats." *Id.* (quoting Regents of Univ. of Cal. v. Bakke, 438 U.S. 265, 317 (1978)).

253. *See* Gratz v. Bollinger, 539 U.S. 244, 270 (2003).

while still serving its purpose.[254] By carefully cabining race-aware fairness constraints and using them only where necessary to serve the governmental interest of preventing algorithmic disparate impacts, algorithmic affirmative action should be able to survive strict scrutiny.

CONCLUSION

I have argued that race-aware fairness constraints should be able to survive both statutory disparate treatment and constitutional equal protection challenges, at least under certain conditions. My arguments have been grounded primarily in the applicable texts, Title VII's legislative history, and the Supreme Court's Title VII and equal protection decisions. These conclusions are certainly not textually or doctrinally preordained, and there remains substantial uncertainty in fitting machine decisions into a body of law that was designed to govern human decisions. At this point, then, it is worth stepping back to look at the larger picture. How does algorithmic affirmative action square with the fundamental principles of antidiscrimination law? The answer to this question may help us predict how courts will deal with the new complexities introduced by machine decisions.

The prevailing account is that U.S. antidiscrimination law has taken a turn toward an anticlassification (formal equality) principle and away from an antisubordination (substantive equality) principle.[255] If so, this spells trouble for race-aware fairness constraints, at least to the extent that they are understood to be racial classifications or "race-based" conduct. The implications for the use of machine-learning algorithms in personnel decisions would be severe. Achieving the anticlassificatory ideal of colorblindness is impossible in the machine-learning context, because we cannot create truly colorblind algorithms. Machines are too effective in identifying proxies. So disparate outcomes might be the inevitable result of machine decisions, and they may well be immune from disparate impact liability under the business-necessity defense.

But Professor Reva Siegel has advanced a third plausible explanation for the Court's more recent antidiscrimination decisions: an antibalkanization principle.[256] If the avoidance of balkanization is the concern driving the close cases, it could bode well for algorithmic affirmative action. Race-aware fairness constraints do not create easily identifiable, visible victims who are likely to feel as though they have been governmentally defined and classified by their race. Compare that to the students requesting school transfers in *Parents Involved*, whose applications were denied because their race would contribute to a school's

254. *See* KUSNER ET AL., *supra* note 59, at 2.

255. *See* Areheart, *supra* note 7 at 993–95 ("[I]t appears the Court has, in *Ricci*, turned hard toward anticlassification values and away from antisubordination rationales that once animated disparate impact analyses."); Siegel, *supra* note 7, at 1287 (describing "the conventional account" that the "anticlassification understanding of equal protection ultimately prevailed" but arguing that in reality, neither has fully displaced the other).

256. *See* Siegel, *supra* note 7 (presenting the antibalkanization principle as a plausible explanation for the Court's recent antidiscrimination decisions).

racial imbalance, or to undergraduate applicants in *Gratz* who learned that individuals of certain preferred races automatically received a predetermined number of points simply because of their race. Race-aware fairness constraints operating within a machine-learning process are not as likely to make race such a visible, salient factor in decisions, and are therefore not as likely to risk escalating social tension. Under this view, carefully circumscribed uses of race or gender variables in calculating and applying algorithmic fairness constraints ought to be a permissible form of voluntary compliance with the disparate impact provisions of Title VII.